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

Kamei, Kenju and Tabero, Katy

Durham University, Durham University

15 November 2021

The Individual-Team Discontinuity Effect on Institutional Choices:

Experimental Evidence in Voluntary Public Goods Provision

November 2021

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

[#1] Department of Economics and Finance, Durham University, Mill Hill Lane, DH1 3LB.
Email: kenju.kamei@gmail.com

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

Abstract: Teams are known to be more cognitively able, and accordingly behave more efficiently, than individuals. This paper provides the first experimental evidence of the so-called "individual-team discontinuity effect" in an institutional setting. In a finitely repeated public goods game where sanctioning institutions are available, teams sustain cooperation surprisingly better than individuals. The superiority of teams is driven by their effective use of punishment. Given an opportunity to construct a formal sanction scheme in their groups, teams enact deterrent schemes by voting much more frequently than individuals. When peer-to-peer punishment is possible, teams inflict costly punishment more frequently on low contributors than individuals, thereby reducing the relative frequency of "misdirected" punishment among teams. These results underscore the effectiveness of having teams as a decision-making unit in organizations in mitigating collective action dilemmas.

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. Additional funding was provided by NINE DTP. The authors thank John Hey for his hospitality when they conducted the experiment at the University of York, and for Louis Putterman and Stefan Penczynski 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, Mohrman, and Ledford, 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 choices and their behaviors under 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. While the members observe the aggregate level of contributions in their group, they are not aware of individual members' contributions, parallel to the feature of unobservable individual effort provision in organizations or the workplace. 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 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 be more cognitively sophisticated and thus 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.*, 2013), coordination games (e.g., Feri *et al.*, 2010), and monetary policy decisions (e.g., Blinder and Morgan, 2005). It is also possible that teams construct institutions differently from individuals by voting, and hence, the former governs groups more efficiently than the latter in the voluntary provision of public goods.

This paper provides the first experimental study by utilizing one of the most commonly-used frameworks – 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 in their group. The institutional formation, their behaviors under constructed institutions and the resulting efficiency are compared against the case where the decision-making units are individuals. The results of the experiment show surprisingly that 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, teams enact deterrent schemes, i.e., schemes that make free riding materially unprofitable, much more frequently than individuals. When informal (peer-to-peer) punishment is enacted, teams punish low contributors more frequently than individuals, which helps reduce the relative frequency of "misdirected" punishment among teams. Moral hazard in groups is a central issue in organizations hurting productivity (e.g., Holmstrom, 1982). While recent experimental research suggests that it can endogenously be resolved by allowing agents to construct institutions (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 effective shape of organizations for efficiency.

The present paper is related to the large literature in 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 features 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 theoretical model using teams of individuals that have homogenous preferences (that align with the common goal), but heterogeneous information. The model focuses on ways in which team members eliminate the information gap among team members in order to face the same decision that an individual decision-maker would. However, teams usually face difficulties in doing so, due to the costs of gathering information and mixed incentives of sharing such 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 have bounded rationality, due to the teams'

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

increased ability to store and process information, for example 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 process even with symmetric information.

Further, this paper also contributes to empirical literature on organizational economics, management, 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 the data from self-assessed/reported questionnaires. Second, team production (the bottom of a hierarchy in firms, such as in production 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 briefly summarizes the related literature in experiments, and then Section 3 describes the experimental design. Section 4 reports experiment results. As a detailed analysis, Sections 5 and 6 report results from finite mixture modeling and communication dialogues, respectively. Section 7 concludes.

2. Related Literature and Research Questions

This study contributes to two large branches of literature in economics and the related social sciences: (a) social dilemmas and endogenous choices of institutions, and (b) group or team decision-making.

First, there is extensive literature on social dilemmas contributed by not only economists but also scholars in neighboring fields, such as political science and psychology. One of the most frequently-used set-ups in this area is a public goods game (PGG). In a public goods game, individuals are randomly allocated to a group of N ($N > 2$), given a fixed endowment, and then simultaneously decide how much to contribute to their group. Parameters are set such that members have private incentives to free ride, while contributing certain amounts to the group is Pareto efficient. For years, such experimental public goods games have been demonstrating that while individuals do not behave as predicted by the assumption of self-interest and the common knowledge of rationality, it is quite challenging to sustain cooperation without any institutions. On average subjects contribute around 40% to 60% of the endowment, even in a one-shot linear PGG without institutions or in earlier rounds of a repeated PGG (e.g., Ledyard, 1995; Chaudhuri, 2011). Despite some individuals' inclinations to cooperate, non-cooperation remains rife and features the expected downward trend of cooperation norms.

Two kinds of institutions can counter the free riding problem. First, groups can sustain cooperation through monitoring and *informal* punishment, provided that punishment acts are not too costly to the punisher, although the standard theoretical prediction still suggests full free riding of members (e.g., Fehr and Gächter, 2000, 2002). This has been replicated by much subsequent research (e.g., Denant-Boemont *et al.*, 2007; Kamei and Putterman, 2015; Nikiforakis and Normann, 2008), underlining the role of human other-regarding preferences in stabilizing cooperation (e.g., Fehr and Schmidt, 2006; Sobel, 2005). The second approach is to introduce *centralized* mechanisms (emulating formal governance) aiming to make cooperation the rational decision through incentive changes. Many of these mechanisms have also seen success. For example, Falkinger *et al.* (2000) studied the behavioral relevance of a tax-subsidy scheme (in which redistribution is exerted from low to high contributors so that cooperation constitutes a Nash Equilibrium outcome), demonstrating in an experiment that contribution rates were sustained close to full efficiency. For the last 15 years, strong development has been made through research conducted by a number of scholars, e.g., Gürerk *et al.* (2006), Kosfeld *et al.* (2009), Sutter *et al.* (2010), Ertan *et al.* (2009), Kamei *et al.* (2015), and Fehr and Williams (2018), allowing individuals to endogenously construct sanctioning mechanisms by voting. These suggest the possibility of self-governance. The main findings are that: (a) without any guidance, individuals are able to construct an efficient formal mechanism by voting, consistent

with theory; and, nevertheless, intriguingly (b) groups prefer and sustain cooperation with *decentralized* mechanisms, such as informal punishment, instead of relying on centralized mechanisms, under certain conditions, if doing so leads to a more efficient outcome. For example, Kamei *et al.* (2015) let individuals choose between formal and informal sanction schemes, the formal similar to Falkinger *et al.* (2000) with the addition that individuals could set their own punishment parameters by voting. They found that both formal and informal mechanisms were effective in incentivizing contribution to a public good. However, informal mechanisms were popular if the formal mechanism entailed a modest administrative cost, despite the standard theory prediction, the benefits of consistency, and reduced risk the formal mechanism offers (see also Fehr and Williams (2018) for strong performance of informal mechanisms through the creation of cooperative norms). To the authors' knowledge, all of the previous studies used *individuals* as the decision-making units. The present paper is the first to study how *teams*, as a decision-making unit, behave differently from individuals in institutional environments if a group consists of multiple teams, and teams make joint institutional decisions by voting.

The second, closely-related area is a substantial literature in economics, management, and psychology on group or team decision-making, which informs the basis of the hypotheses in the present study. Prior experimental studies have demonstrated what is termed the "discontinuity effect" (Schopler *et al.*, 1991) by which individuals behave differently from teams (e.g., Charness and Sutter, 2012; Kugler *et al.*, 2012; Kerr *et al.*, 2004) in a number of contexts; the present study is the first to study the discontinuity effect in an institutional setting. One persistent finding is that teams display greater cognitive ability than individuals in logic or problem-solving activities. For instance, Kocher and Sutter (2005) found teams playing a repeated beauty contest game reached the game-theoretic prediction of 0 much quicker than individuals did, showing a stronger ability to work out the logic of the game. Similarly, in a replica of monetary policy decision making, Blinder and Morgan (2005) found that teams consistently out-performed individuals, and were not slower at reaching a decision. These kinds of sophisticated team behaviors have also been seen in various games, such as centipede games, signaling games, ultimatum games, and trust games (see the survey articles listed above). This tendency is expected to be relevant to the designing of a mechanism in the present study as teams may be better able to set efficient parameters, for example setting punishment rates high enough that it is

rational to contribute under centralized mechanisms. It may also prevent incidences of "perverse punishment" by which agents choose to punish high contributions under peer-to-peer punishment (Cinyabuguma *et al.*, 2006) or choose to punish the public instead of private account through selecting inefficient policies under formal punishment (Kamei *et al.*, 2015).

Another persistent, important finding in the literature is that teams may be less myopic loss averse than individuals (e.g., Bougheas *et al.*, 2013; Sutter, 2007 and 2009). This behavioral tendency also predicts teams' better institutional formation than individuals since they may be more willing than individuals to incur punishment costs to enforce social norms, as units can enjoy strong benefits from building long-term cooperative relationships in their group. Hence, the literature suggests affirmative answers to the following two research questions of the paper.

***Question 1*:** Do teams utilize sanctioning institutions more efficiently than individuals to enforce cooperation norms?

Question 2: As a result, do teams sustain cooperation more easily than individuals in an institutional setting?

Teams' more rational choices could result from fear and greed (e.g., Wildschut *et al.*, 2003; see also Ahn *et al.* [2001]). Greed is made more acceptable for teams as the team dynamics provide anonymity and support for otherwise socially unpopular views, for example, the decision to contribute less is not marked against any single individual in the team. This anonymity also makes it easier for others to agree to strategic decisions that exploit peers, as they are equally protected. Fear, comparatively, encourages distrust and low cooperation as it is expected by teams that the opponent will be aggressive or deceitful, referred to as the 'out-group schema' (Schopler and Insko, 1992, pp. 128). This view is supported in a paper by Schopler *et al.* (1993), in which a team were allowed to withdraw from a game and accept a medium payoff instead of trying to compete or cooperate with another team. They found that teams made much higher use of the withdrawal option than individuals, and that when they did not withdraw they tended to compete more than individuals, also demonstrating a reluctance to cooperate. Hence, in the context of the present study, without sanctioning institutions, teams may behave more in line with standard game-theoretic predictions than individuals. This distrust would, nevertheless, be expected to vary greatly by mechanism design as certain treatments require communication which is crucial for cooperation (for example, see Brosig *et al.* [2003] and Kamei [2019b]).

Having said this, some studies have also found evidence of teams behaving more

cooperatively than individuals in a *repeated* environment, perhaps relating to their increased understanding of the game. Hence, no clear predictions are possible for discontinuity effects when sanctioning institutions are absent. For example, in Wildschut *et al.* (2003) the individual-team discontinuity effect was minimized most when reciprocal strategies were practiced. Specifically, when the team responds in the same way as their opponent responded in the previous round, intergroup interactions were more cooperative. Kreps *et al.* (1982) showed theoretically that this may be a rational strategy where there is uncertainty over whether the unit's opponent will play a "tit-for-tat" or non-cooperative strategy, where a tit-for-tat strategy yields a higher payoff. This is empirically supported with experimental evidence given by Kagel and McGee (2016) who found that when teams were able to play multiple matches against different opponents, while mostly non-cooperating in the initial game for safety concerns, they shifted to a more reciprocal strategy in later matches. Gillet *et al.* (2009), Feri *et al.* (2010), Kamei (2019b), and Müller and Tan (2013) report such teams' more cooperative behaviors in a repeated common-pool resource problem, weakest-link/average-opinion game, PGG, and Stackelberg market game, respectively.

3. Experiment Design

The experiment is built on a linear public goods game. A between-subjects design is used where subjects play the games under one treatment condition.[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 decision-making units can use sanction schemes. Two different strengths of punishment are considered because the efficacy of sanctioning mechanism is known to depend on its strength. The six 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 has the ability to vote on whether to execute a formal or informal sanction scheme, and also the parameters of such a scheme. A novel part of the design is that unlike all prior studies on

[2] A between-subjects design is more appropriate than a within-subjects design, to avoid possible behavioral spill-over across different environments (e.g., Bednar *et al.* 2012; Cason *et al.*, 2012). Kamei (2016) also found that subjects' experiences in one institutional environment may affect their behaviors in another environment.

endogenous choices of institutions (e.g., Kamei *et al.*, 2015; Kosfeld *et al.*, 2009; Traulsen *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 as the decision-making units will act as a control treatment.

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: Each team consists of three subjects. [#1] The ratio, 1: x, means that for each point a punisher spends in reducing the payoff of a player, x points are deducted from the payoff of the punished. The punishment cost ratio of 1:3 (1:5.5) means $x = 3$ and $y = 5$ ($x = 5.5$, and $y = 10$) – see Subsection 3.2 for the detail.

3.1. 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. 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 to their public account. The remaining points will be automatically allocated to their own private account. 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:

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

It is clear from Equation (1) that contributing nothing to the public account is the strictly dominant strategy for every decision-making unit. Hence, according to standard theory based on agents' self-interest and common knowledge of rationality, mutual free riding characterizes the unique Nash Equilibrium of this game. Notice that repetition does not alter this prediction with the logic of backward induction.

In the treatments with teams (T-No, T-Voting and T-Voting-ST), each member in a team *i* receives the team's payoff to make the payoff consequence of team members in the team treatments the same as that of individuals in the Individual treatments.[3] At the end of a given period, each decision-making 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 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 free riding problem. Phases 2 to 6 (collectively "Part 2," hereafter) differ by whether they can use sanction schemes, as summarized in Sections 3.2 and 3.3.

3.2. The Individual Treatments

In the sanction-free I-No treatment, subjects play the PGG under the NS scheme for all five phases after the first phase (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). As explained below, subjects in the treatments with voting have a voting stage at the onset of each phase.

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 simultaneous, mandatory, and does not cost subjects. As discussed below, the standard theory predictions are different between FS and IS. At the beginning of each phase, the decision-making units decide 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. Section 3.2.1 (3.2.2) summarizes the details of the IS (FS) scheme.

[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).

3.2.1. 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 $p_{i\to j} \in \{0, 1, 2, \ldots, 10\}$. This is a costly punishment activity. While each punishment point costs the recipient *x* points ($x > 1$), it costs the punisher one point. Following prior experimental frameworks of punishment (e.g., Fehr and Gächter 2000, 2002; Kamei *et al.*, 2015), 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. In short, the payoff for decision-making unit *i* in period *t* playing IS can be expressed as follows:

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

Standard theory predicts that having IS does not alter equilibrium play from that in the NS scheme because punishment activities are costly. $\frac{\partial \pi_{i,t}}{\partial p_{i\to j}} = -1 < 0$ for all *i*. Note that each unit may choose not to 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)). 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.

3.2.2. Formal Sanction Scheme

When FS is in place, the allocation stage is followed by a punishment stage in each period, similar to the IS scheme. However, unlike the IS scheme, groups in the FS scheme determine by voting in advance at what rate allocations to the *private* account are penalized. Voting is simultaneous, 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. 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 decision-making units in the group, meaning that each unit pays (1/3)(3/*y*) = 1/*y* of the sanctions.[5] There is also a fixed cost (administrative cost) imposed when using the FS of 4 points, which is applied to each unit's payoff for that period.[6]

To parallel the IS scheme, the deductions resulting from the sanction rate 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 *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 additional 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 treatments, while strong punishment intensity, $x = 5.5$ and $y = 10$, is used for the I-Voting-ST and T-Voting-ST treatments.

Standard theory prediction in the FS scheme is different from that in the NS or IS scheme

[5] To mirror the cost of punishment in the IS Scheme the FS scheme also features a proportional cost. However, unlike the IS mechanism this cost will be borne by the whole group. Depending on the punishment strength of the treatment, the cost of punishing will be the summed punishment inflicted on all decision-making units in that group, multiplied by the relevant sanction rate, and then shared equally between each unit in that group (including those punished).

[6] See the 6-C treatment of Kamei *et al.* (2015).

(as in Falkinger *et al.* [2000] and Kamei *et al.* [2015]). $\frac{\partial \pi_{i,t}}{\partial p_{i \to j}} = -0.4 + SR_t \left(1 + \frac{1}{y}\right)$. It therefore predicts that units contribute nothing when the enacted *SR* is less than 0.4, but they contribute the full endowment amount when it is greater than 0.4. When *SR* = 0.4, each unit still has a material incentive to contribute everything because of the per capita share of imposing the fine ($\frac{\partial \pi_{i,t}}{\partial p_{i \to j}} = -0.4 + 0.4 + \frac{0.4}{y} = \frac{0.4}{y} > 0$). Each unit obtains a payoff of 32 points (= 0.6×60 – 4) when a deterrent sanction rate is enacted, while they obtain a payoff of 16 points (= 20 + 0.6×0 – 4) when a non-deterrent sanction rate is enacted instead. This means that the theory predicts groups would choose FS rather than IS, and then vote for a deterrent sanction rate (see Kamei *et al.* [2015]).

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

3.3. The Team Treatments

The T-No, T-Voting and T-Voting-ST treatments are identical to the I-No, I-Voting and I-Voting-ST treatments, respectively, except that the decision-making units are *three-person teams*, not individuals (Figure 1). 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 procedure is similar to Kamei (2019b, 2021) and is as follows: three members in a team communicate with each other for 60 seconds using a computer chat screen before making each team decision.[7] This enables us to perform transcript analysis

[7] The use of electronic chat windows is one of the most common procedures used in the literature on team decision-making (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 decisions within 60 seconds in the experiment. This suggests that the 60-seconds duration was sufficient in the communication stage. A detailed analysis of communication logs will be executed with two independent coders (see Section 6).

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. In the communication stage, the members are only able to communicate with other members of their own team. Anonymity is also kept preserved in the communication stage. In the chat screen, the subjects are identified by Player IDs which are allocated randomly at the start of the experiment; they are instructed that disclosing any information that may identify themselves or using offensive language in communication is prohibited.[8]

A team's three kinds of joint decisions are determined using the median voting rule (Figure 1.C). This includes the allocation decisions between the private and the public account in all three team treatments, and punishment decisions under the IS scheme and sanction rate votes under the FS scheme in the 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 teammates. After the communication stage, each team member privately and simultaneously submits their preferred decision (e.g., an amount they wish to contribute as the team's joint contribution decision).[9] The median of three members' submissions becomes the team's joint 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 decisions for which scheme to vote for, 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.[10]

3.4. Implementation

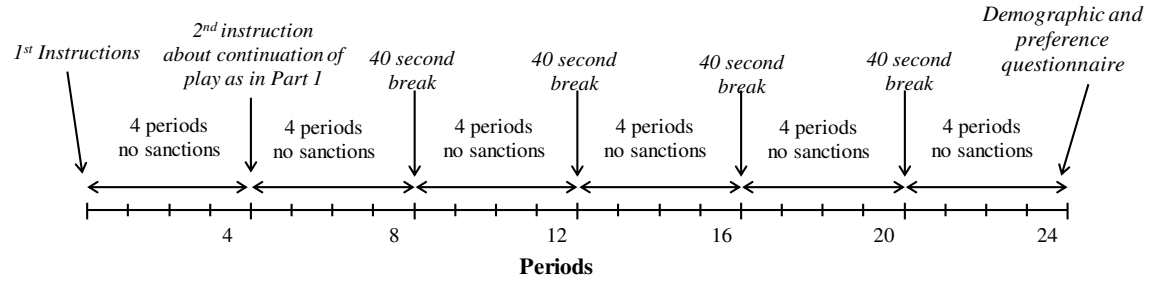
The experiment was conducted at the EXEC laboratory in the University of York from July 2019 through January 2020. Observations of 11 or 12 groups were collected for each treatment condition by conducting six or seven (two) sessions in each Team (Individual)

[8] A subject receives a fine of 10 pounds with an apparent violation of this rule. No one disclosed any idenfiable information, and only 7 out of 306 subjects (2.28%) in the three team treatments had to pay the fine with the rule of offensive language in team discussions.

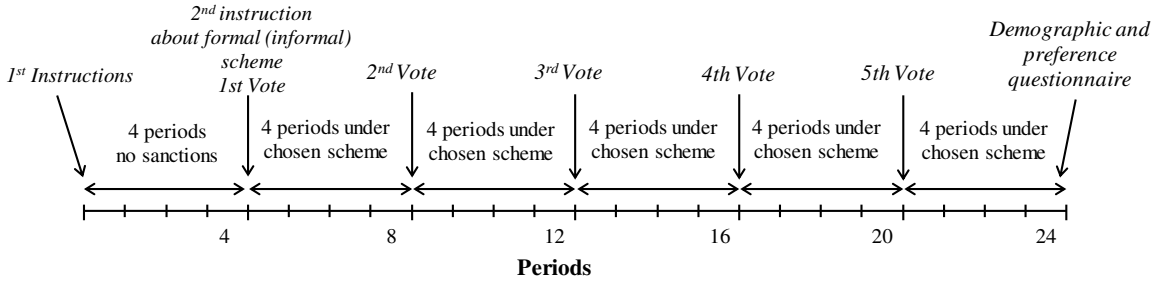
[9] Where the team members agree on a decision, they can submit that decision. If they do not agree on a decision as a team, however, they can submit whatever decision they prefer. Three team members submitted the same decisions in almost all cases in the team treatments (2,049 out of 2,448 team allocation decisions, 581 out of 672 team sanction rate votes, and 1,176 out of 1,296 team informal punishment decisions).

[10] 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).

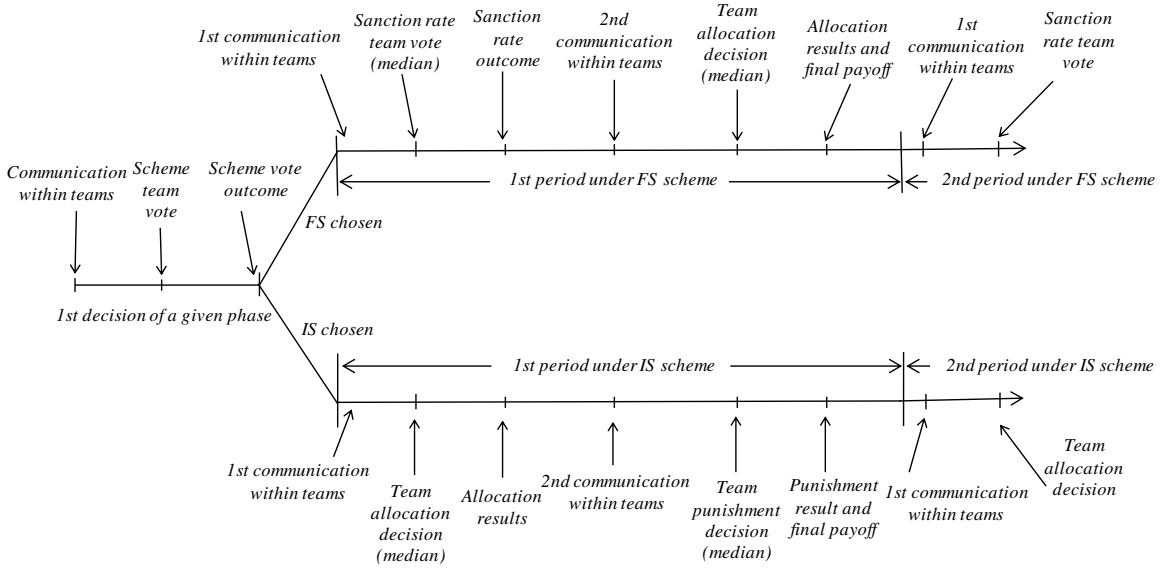
Figure 1: *Schematic Diagram*



(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

treatment. A total of 408 subjects (25 sessions) participated in the experiment. The experiment, except instructions, was programmed in the z-Tree software (Fischbacher, 2007). The schematic diagrams can be found in Figure 1. All subjects were recruited using solicitation emails sent through *hroot* (Bock *et al.*, 2014). All instructions were neutrally framed. Any loaded words,

such as cooperate, were avoided.[11] Communication, except the communication via electronic chat windows in the team treatments, was prohibited. At the end of the experiment, subjects were asked a number of demographic information questions, such as gender.

4. Experimental Results

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 choices of decision-making units, while Section 4.3 provides a comparison between individuals and teams 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 also seen in Phases 4 to 6 (Andreoni, 1988), and end-game defection was also 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 overall 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.[12]

Contribution trends differ drastically between individuals and teams when they voted on the sanctioning schemes. 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 16 points (80% of the endowment) in the final three phases. This contrasts with the teams' declining contribution trend in the T-No treatment. On the other hand, individuals in the I-Voting treatment did not 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 in each

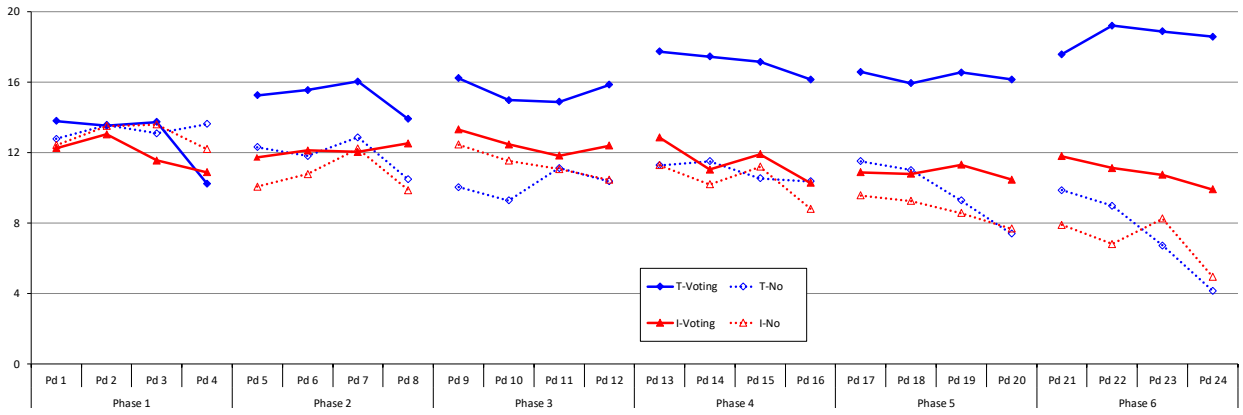
[11] In all treatment conditions, at the outset the basic structure of experiment was explained to the subjects, such as the number of periods, phases and the matching protocol (the fixed team and group composition), and the condition of Part 1. The instructions in Part 1 were the same for the I-No, I-Voting and I-Voting-ST treatments (the T-No, T-Voting and T-Voting-ST treatments). Subjects received the other set of instructions after the initial phase. The instructions for Part 2 differed by treatment. The gradual introduction of conditions helps reduce cognitive loads on subjects, and is often used in the PGG experiment with institutional choices (e.g., Ertan *et al.*, 2009; Kamei *et al.*, 2015).

[12] 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).

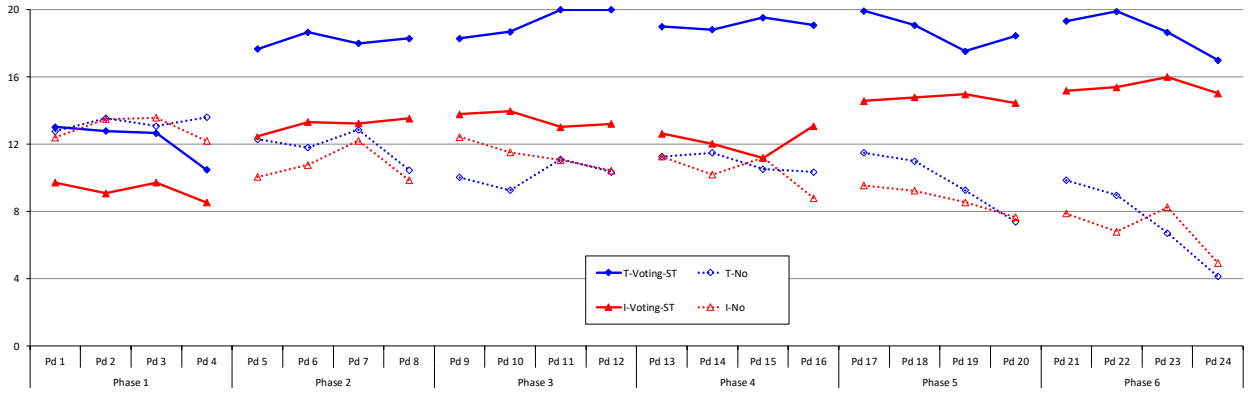
phase. The clear difference between the T-Voting and I-Voting treatments is consistent with the discontinuity-effect hypothesis, demonstrating its application in an institutional choice setting. So, why did teams perform better than individuals only when sanctioning schemes were available? An answer to this question may be that teams use punishment opportunities more efficiently than individuals, perhaps driven by the former's greater cognitive ability, a factor which will be analyzed and shown to be the case in Sections 4.2 and 4.3.

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. Interestingly, 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'

Figure 2: *Average Contribution Period by Period*



(A) Treatments with Modest Punishment Intensity



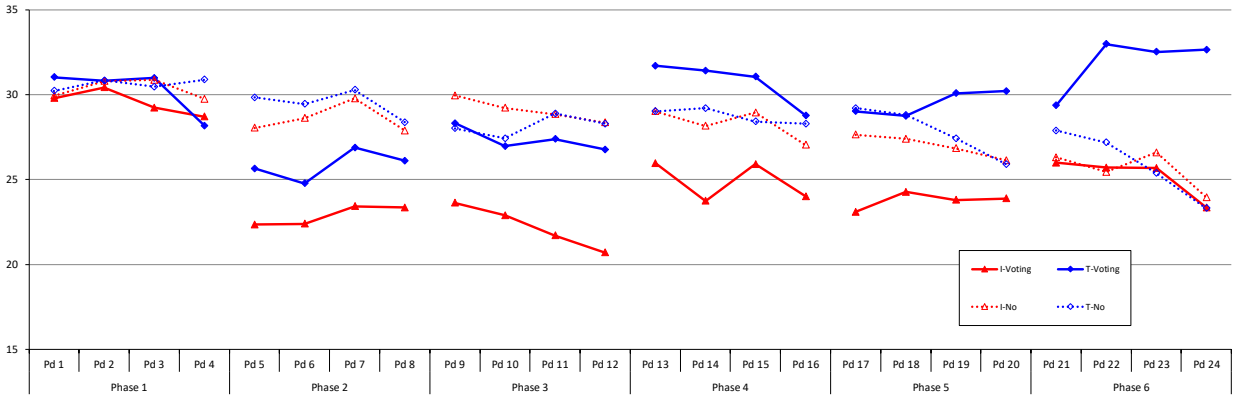
(B) Treatments with Strong Punishment Intensity

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

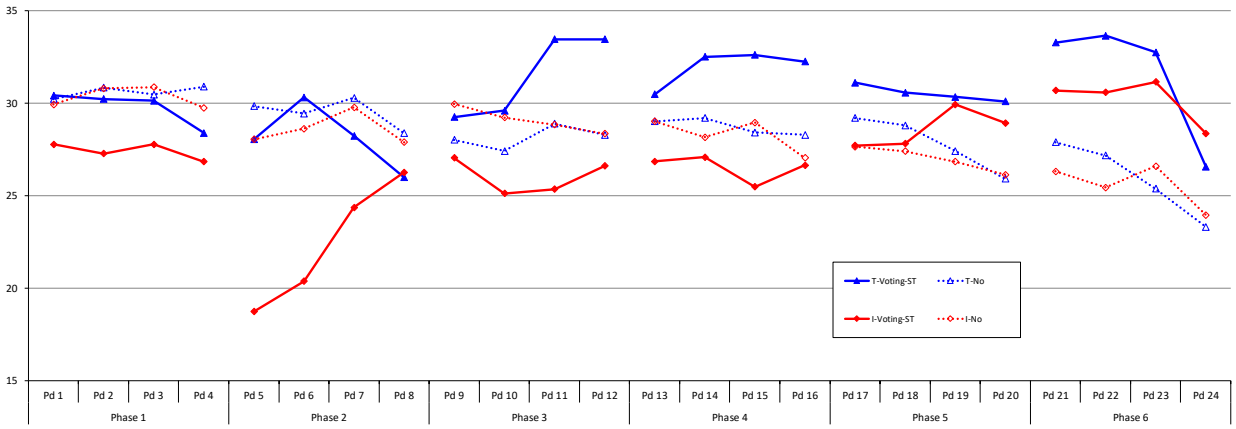
contribution behaviors are sensitive to the punishment effectiveness as has been shown by Anderson and Putterman (2006) and Nikiforakis and Normann (20087). Having said this, the difference in the average contribution was consistently large between individuals and teams even under the strong punishment intensity.

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 much 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

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. The average payoffs in the I-No (T-No) treatment were monotonic transformation of average contributions in Figure 2 based on Equation (1).

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 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 lead 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 available than otherwise (see H_0: (c) = (d) in Table 2).[13] 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.

A regression was also performed as a supplementary analysis, in order to additionally analyze the contribution trend in Part 2 (Appendix Table B.1). It first confirms that when the sanctioning schemes were unavailable, decision-making units, whether individuals or teams, decreased contribution amounts significantly over time. Second, by sharp contrast, teams increased contribution amounts significantly from phase to phase in both the T-Voting and T-Voting-ST treatments. Third, the contribution trend differs by punishment intensity when the decision-making units are individuals: The contribution follows an increasing (somewhat decreasing) trend in the I-Voting-ST (I-Voting) treatment. A regression also confirms that the payoff trend is similar to the contribution trend: declining trends for the I-No and T-No

[13] The same positive effect of voting can be found even if we do not pool the two team treatments – see Panel C of Appendix A.

treatments *versus* an increasing trend in the T-Voting treatment (the maintenance of high payoff in the T-Voting-ST treatment) – see Appendix Table B.2.

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-treatments (across-treatments) comparisons. The standard errors can be found in Panel A of Appendix A. More detailed across-treatment comparisons can be found in Panel C of Appendix A. 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 reported in Table 2 and the regression analyses in Appendix Tables B.1 and B.2, additional regression was conducted

to study treatment differences in efficiency using group-level average contribution or payoff as the dependent variable and treatment dummies as independent variables. As shown in Appendix Table B.3, it consistently confirms strong discontinuity effects under voting between individuals and teams. [#1] Only groups that had experienced both the FS and IS schemes in Part 2 were used. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

Lastly, a closer look at the data by scheme uncovers three further interesting patterns. First, in the I-Voting and I-Voting-ST treatments, while cooperation did not evolve when FS was in place, the individuals maintained strong cooperation norms when IS was instead in effect (panels (a) and (b) of Appendix Figure B.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, resulting from the low cooperation norms and administrative cost payments, individuals persistently earned much less when they had the FS scheme, relative to the I-No treatment (panels (a) and (b) of Appendix Figure B.2). The difference is significant between the I-Voting and I-No treatments (Table 2.II). 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 was not quick, requiring enough interaction experiences in the experiment (Gächter *et al.*, 2008).

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 B.1). Table 2 also shows that regardless of which scheme was in effect, teams contributed significantly larger amounts in the T-Voting and T-Voting-ST treatments than in the T-No treatment. 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* the phases (panels (c) and (d) of Appendix Figure B.2).

Result 1: *(a) Decision-making units (whether individuals or teams) learned to free ride over time*

[14] A group-level Mann-Whitney test finds that the average contribution in Phase 6 (in 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 $p = .0709$ (.0196). Likewise, the average payoff in Phase 6 (in 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 $p = .0709$ (.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.

when sanctioning schemes were unavailable. (b) With the sanctioning schemes, individuals in the I-Voting treatment prevented a collapse of cooperation norms, and individuals in the I-Voting-ST gradually learned to cooperate over time. Nevertheless, their contribution behaviors in the FS scheme were weak. (c) The impact of voting was strong for teams: Under each punishment intensity, teams learned to cooperate more strongly than individuals regardless of which sanction scheme (FS or IS) was available – in support of the discontinuity-effect hypothesis between individual and team decision-making.

4.2. Scheme Choice

The strong efficiency under the IS scheme found in Section 4.1 was not driven by a small number of groups. Despite standard theory 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.a). 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.b). 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 A).

A look at the across-phase trend indicates that decision-making units' preferences for the IS scheme are stable. Around half of the units preferred the IS scheme in the very first voting phase in the I-Voting, T-Voting and I-Voting-ST treatments (in the second voting phase for the T-Voting-ST treatment), after which the popularity of the IS scheme remained stable.

Realized relative effectiveness of informal and formal sanctions affected decision-making units' decisions to choose schemes. Seven, nine, eight and six groups experienced both the FS scheme in some phase(s) and the IS scheme in the other phase(s) as a result of voting. Using

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 .10 level, at the .05 level and at the .01 level, respectively.

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 strongly guided by material incentives (e.g., Ertan *et al.*, 2009; Kamei *et al.*, 2015).[15,16]

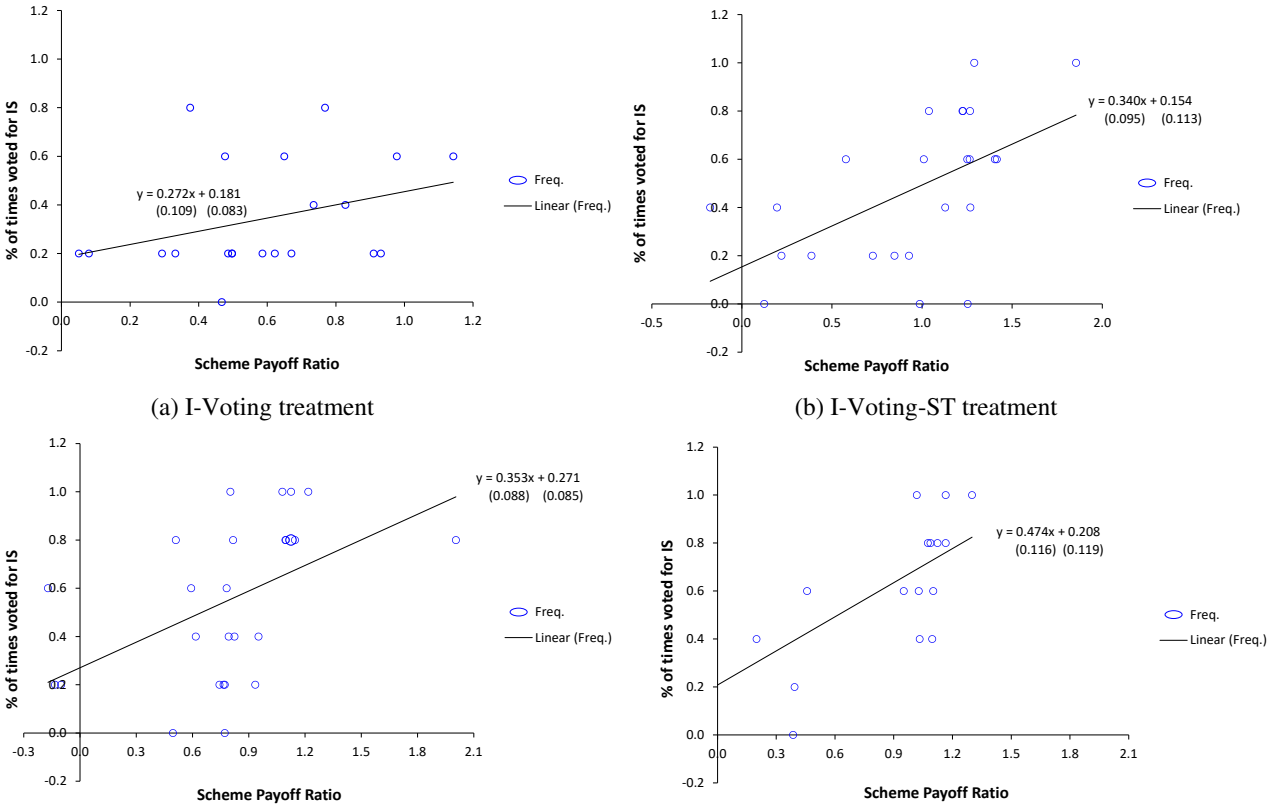
Figure 4 also indicates two more interesting patterns. First, there is a large variation for decision-making units' voting behaviors: the correlations between units' scheme choices and experienced relative payoff ratios are far from perfect. This implies strong heterogeneity in subjects' cooperation and punishment tendencies (e.g., Fischbacher *et al.*, 2001; Fischbacher and Gächter, 2010; Kamei, 2014). Second, the punishment intensity markedly influenced the relative effectiveness of informal sanctions. Under the modest punishment intensity, a considerable majority of the decision-making 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 activities. However, the informal punishment became more effective under the strong punishment intensity. The percentages of the subjects who on average earned less

[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 B.4, the relative payoff ratio is a significantly positive predictor for their selection of the IS scheme both in the individual voting and team voting treatments (when data are pooled irrespective of the punishment intensity). The role of experiences is also supported by an analysis of communication logs (Section 6).
[16] It should be acknowledged that the direction of causality may be the opposite. Those who sorted into the IS scheme may have more cooperative inclinations, and hence may have ended up earning higher payoffs with informal punishment.

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 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.[17] The average contributions of groups that always selected FS were

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 as a result of voting. The horizontal axis (x-axis) is calculated by a given decision-making 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 decision-making unit voted for IS during the experiment and takes a value between 0 and 1. The size of each point indicates its frequency (almost all points have a frequency of 1). The numbers in parentheses in the linear 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 = .046, .009, .004, and .009, respectively.

[17] 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.

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.

What factors, other than material concerns, might affect the groups' scheme choices? To find some nuanced evidence, bi-variate correlations between a group's average vote outcome and their average attribute variables were calculated. Three reasonable, but intriguing patterns emerged (Appendix Table B.5). First, subjects' preferences for fairness drove their scheme choices. Subjects provided their views on fairness for each scheme in the post-experiment questionnaire.[18] A calculation found that the fairer group members on average perceived the IS scheme relative to the FS scheme, the more likely the 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 believe that people are trustworthy, the more frequently a given group implemented the IS scheme.

Third, cognitive ability may have also 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.

Result 2: *(a) Despite standard theory predicting the superiority of the FS scheme, around half of the groups adopted the IS scheme. (b) Decision-making units tended to vote for the scheme under which they had previously experienced higher payoffs. (c) Almost all groups that selected one scheme (either FS or IS) for all phases achieved successful cooperation in that scheme. (d) Groups that perceived the IS scheme as fairer were more likely to implement that scheme. Groups with stronger trust in others' cooperation behaviors were more likely to implement the IS scheme.*

4.3. Discontinuity Effects in Utilizing the Sanctioning Institutions

Analyses so far have found that while scheme choices were similar for individuals and teams (Section 4.2), teams cooperated much more strongly than individuals did (Section 4.1).

[18] Subjects were asked to rate the fairness of each of the three schemes (NS, IS and FS) on a five-point scale: from 1 (very unfair) to 5 (very fair).

This subsection explains that this discontinuity effect was driven by the difference in the ways in which decision-making units utilize the sanctioning institutions.

4.3.1. Voting and Contribution Behaviors in the FS Scheme

Consistent with the standard theory prediction, units' decisions to contribute under the FS scheme were strongly influenced by their group's sanction rate. A regression analysis finds that decision-making units were significantly more likely to contribute large amounts, the higher the sanction rate their group had implemented (odd-numbered columns of Appendix Table B.7). Having a deterrent sanction rate effectively improves units' decisions to contribute (even-numbered columns of Appendix Table B.7). The larger impact of having stronger punishment is consistent with prior research on formal sanction 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.

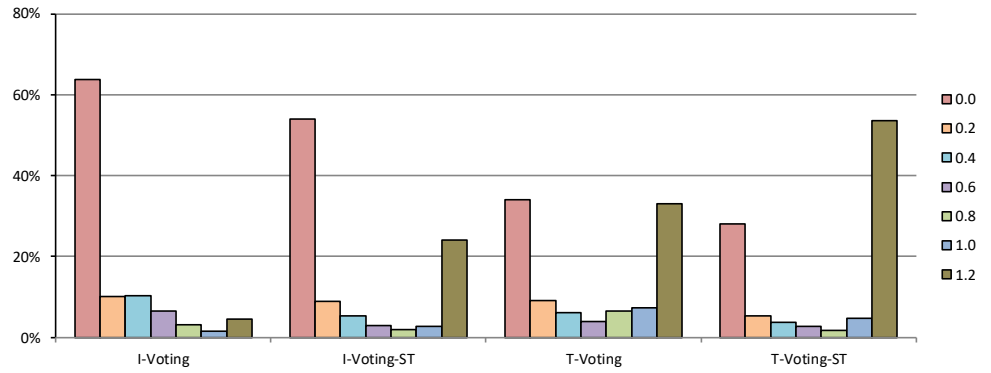
However, the popularity of sanction rates differs clearly between individuals and teams. The sanction rate of 0.0 was the focal point among the individuals. Strikingly, individuals in the I-Voting and I-Voting-ST treatments voted for the zero sanction rate on 63.79% and 54.00% of the occasions, respectively (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).[19,20]

By sharp contrast, in the team treatments, preferences for the highest sanction rate – 1.2 per point allocated to the private account – were quite strong (Figure 5.A). Especially in the T-Voting-ST treatment, teams voted for the highest rate on 53.54% of the occasions. At the same time, teams voted for the zero sanction rate only 34.06% and 28.03% of the time in the T-Voting and T-Voting-ST treatments, respectively. With the majority rule, 31.52% (26.09%) and 62.12% (14.39%) of the group's vote outcomes were the highest (zero) sanction rate in the T-Voting and T-Voting-ST treatments, respectively.[21]

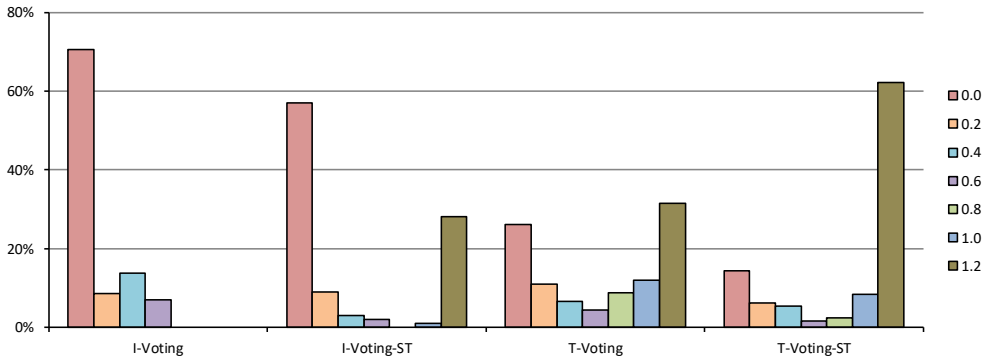
[19] In Kamei *et al.* (2015), almost all individuals successfully constructed deterrent schemes. The difference between this study and Kamei *et al.* could be due to the difference in the research site: the USA versus England. Alternatively it could be due to the difference in the group size – three in this study versus five in Kamei *et al.* (2015).
[20] 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 behavioral public choice theorem (e.g., Ertan *et al.*, 2009; Hauser *et al.*, 2014).
[21] 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 $p = .0080$ ($p = .0319$), according to a group-level Mann-Whitney test, when pooled data are used – see Panel F of online Appendix A.

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.[22,23] The difference in the severity of selected sanction rates well explains the stronger contribution behaviors of teams (Figure 2, Appendix Figure B.1).

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



(A) Distributions of Decision-Making Units' Voting



(B) Distributions of Vote Outcomes

Decision-making units' decisions to contribute interestingly differ 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

[22] The average realized sanction rates are significantly different at two-sided $p = .0116$ between individual versus team voting when pooled data are used (see Panel F of Appendix A)

[23] Figure B.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.

cooperative groups can be assumed to select stronger sanction rates, making mutual cooperation easier (Appendix Table B.7). If this interpretation 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 (Table 4).

The maintenance of group cooperation norms leads to large long-term payoffs. Hence, the teams' stronger behavioral responses to given sanction rates suggest that, with the FS being enacted, teams may be more far-sighted and less myopic loss averse than individuals (e.g., Sutter, 2007, 2009; Bougheas *et al.*, 2013).

Result 3: *(a) 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-Voting-ST) treatment was 0.11 (0.39). (b) Teams contributed significantly more than individuals for given sanction rates. Particularly, the former contributed much more than the latter when non-deterrent sanction rates were in effect.*

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

	(a) Individual Voting			(b) Team Voting			(c) Mann-Whitney tests[#1]		
Sanction rate	(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)	.0110**	.1415	.0274**
0.4 or above	17.67 (5.38)	16.72 (6.29)	18.34 (4.27)	19.10 (1.51)	18.50 (2.14)	19.42 (0.72)	.0268**	.1467	.0949*

Notes: Numbers in parenthesis are standard errors based on group averages. [#1] Two-sided *p*-values for group-level Mann-Whitney tests. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

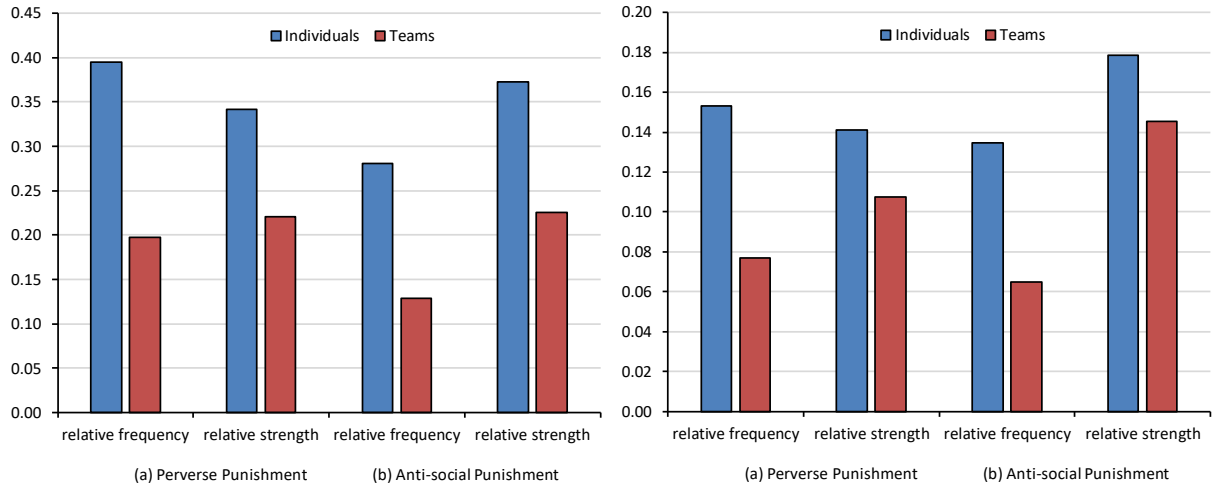
Lastly, as was done for the groups' scheme choices in Section 4.2, bi-variate correlations between a group's average sanction rate and their attribute variables were calculated in order to explore what factors (other than material concerns) might have affected their selection of sanction rates. While no discontinuity effects were found with this analysis, the calculation reveals (a) female subjects' possible dislike of using punishment, (b) economics students' rational voting behaviors, 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 – see Appendix Table B.6 for details.

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 B.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* (see the absolute negative deviation variable). Second, contributing more than another member also attracted punishment by that member to some degree (see the positive deviation variable), but such anti-social punishment is weaker than pro-social punishment – the difference is statistically significant according to Wald tests. 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).

However, intriguingly, behavioral responses to punishment received differ by the decision-making unit (see again Table B.8). First, 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 formed contribution decisions based on their group's last-period contribution decisions, implying that they tend to simply conform to group norms. On the other hand, while such conformity explains teams' contribution dynamics on average, they also 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.

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.

The difference in the behavioral response to punishment created overall distinct punishment distribution by the decision-making unit. Figure 6 reports the relative strength and frequency of anti-social (perverse) punishment to pro-social (non-perverse) punishment. It reveals that pro-social (non-perverse) punishment was more dominant among teams than individuals.[24] This pattern again supports the discontinuity effect hypothesis, meaning that teams utilize informal punishment opportunities more effectively than individuals to encourage cooperation.

5. Structural Estimations of Punishment Types under the IS Scheme

The main finding demonstrated in Section 4 is 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. Specifically, deterrent sanction rates were enacted much more frequently among teams than individuals under the FS scheme. Under the IS scheme, the relative frequency and strength of anti-social punishment were both smaller among teams than individuals.

While the percentages of behavioral 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 behavioral differences under the IS scheme more accurately by using a finite mixture modeling approach since it allows estimation of the distribution of types from observed punishment patterns (e.g., McLachlan and Peel, 2000; Moffatt, 2016).[25]

Finite mixture modeling assumes a set of possible behavioral types in advance and then

[24] Due to the small sample size, the difference is only significant at $p = .0544$ for the relative frequency if a one-sided group-level Mann-Whitney test is used for pooled data (i.e., the two individuals treatments versus the two team treatments).

[25] While the popularity of sanction rates in the FS scheme (Figure 5) explains treatment differences well already, some units voted on sanction rates in an indecisive manner (e.g., voted for deterrent rates in some periods but voted 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; a type who votes based on relative earnings in the past under the IS versus FS schemes). However, almost all models were unable to be estimated (failed to converge) due to too a small sample size for such indecisive subjects.

assigns a probability measure over the types to each subject so that the likelihood is maximized. Table 5 reports the estimation results.[26] 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 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, columns 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.[27] This implies that the issue of misdirected punishment is ubiquitous whether among individuals or teams.

The estimation results by the respective treatment provide 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.[28] Hence, under weak punishment intensity, there is strong evidence that team decision-making effectively prevents

[26] 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 by Moffatt (2016). See the footnote of Table 5.

[27] Regarding misdirected punishment, no consistent patterns were seen by the definition: "anti-social" punishment was less frequent among teams than individuals, while, conversely, "perverse" punishment was more frequent among the former than the latter.

[28] 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 $p = 0.391$ ($p = 0.148$).

decision-making 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 columns 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 columns 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 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. 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 (column B.iii). The corresponding percentage is around 45-52%, in the I-Voting-ST treatment (column A.iii), meaning 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 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 χ^2	118.77	103.89	43.70	162.02	128.48	41.56
Prob > Wald χ^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 χ^2	120.31	61.92	14.56	123.22	70.73	39.78
Prob > Wald χ^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 .10 level, at the .05 level and at the .01 level, respectively.

Can teams' effective use of punishment fully explain their sustained contribution behaviors? One may assume that, as was seen in Result 3.b, it is possible for decision-making units to change *contribution* strategies dependent on the format (individual or team decision making), which may account for different contribution dynamics between individuals and teams. As a further analysis, units' contribution types were structurally estimated using the finite mixture modeling approach.

Appendix Table B.9 summarizes the detailed estimation result. It shows, first, that when sanctioning institutions are absent, almost all units are classified as cooperative types regardless of the decision-making format. In particular, 77.8% of individuals and 91.7% of teams are estimated to have decided how to contribute conditional upon their group members' last-period contribution amount in the I-No and T-No treatments, respectively (Model I of Table B.9). Second, and interestingly, such cooperative types, particularly the "reciprocator" type, are similarly predominant both for individuals and teams when the IS scheme is in effect. Only 4.9% of individuals and 10.2% of teams decided contribution amounts based on the size of punishment they received from peers in the previous period (Model II of Table B.9). This estimation result is parallel to the standard game-theoretical hypothesis, in that the theoretical prediction does not change by the availability of peer-to-peer punishment. Hence, the performance differences between phases with versus without the IS scheme, as well as teams' strong cooperation under the IS scheme, can be attributed to the different levels of cooperation attitudes induced by punishment activities. As reciprocators contribute conditional upon their peers' contributions, they can on average contribute high or low depending on the witnessed norms in their group. Recall that the levels of teams' contributions were much higher *from the very first period* of

almost all phases when the punishment institutions were available (Figure 2, Figure B.1). Thus, reciprocators would have helped sustain these high cooperation norms in the team treatments, and conversely reciprocated the lower contributions seen in the individual treatments.

Third, however, the estimated distributions under the FS scheme differ markedly from those under the IS scheme (Model III of Table B.9). Only 35.9% of individuals and 7.2% of teams are categorized as reciprocators. Instead, 36.6% of individuals and 51.3% of teams decided contribution amounts in response to the sanction rate collectively enforced in the current period. This nicely explains the mechanism behind the performance differences between individuals and teams: Individuals were less likely than teams to enact strong sanction rates (Result 3.a), which could in turn *demoralize* the individuals from contributing to their groups. Hence, it can be concluded that teams' efficient voting is key for their strong contribution behavior. It should be worth noting here that strong cooperation norms fostered by teams might have spilled over to the period with non-deterrent sanction rates, consistent with the so-called behavioral spillover phenomenon, as teams also contributed strongly under non-deterrent sanction rates (Result 3.b, Table 4).

6. Team Communication Dialogues

While empirical analyses performed thus far were based on decision-making units' decision data, teams' communication dialogues contain richer information that may explain the reasoning behind team decisions. As a final analysis, teams' communication dialogues were carefully analyzed following the standard coding procedure in the current experimental literature (e.g., Cason and Mui, 2015; Kagel and McGee, 2016; Leibbrandt and Sääksvuori, 2012). In particular, two research assistants (RAs) were hired as independent coders. The two RAs did not know each other through the entire coding process. They were also not explained any substance of the research, such as the research aim or the subject pool, to avoid demand effects. Instead, they were simply provided with the instructions, teams' communication dialogues, and the list of codes, and were then asked to assign as many relevant codes as possible to each dialogue. The full list of codes is available in online Appendix C.2. Once the two RAs finished coding all of the groups' logs, 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 do the same reconsideration process. This *reconsideration* process was first used, and confirmed

its effectiveness to catch any errors in initial coding, by van Elten and Penczynski (2020). Online Appendix C.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 in the present paper to judge the reliability of coding (e.g., Cason and Mui, 2015; Leibbrandt and Sääksvuori, 2012). Kappas were calculated as 0.28, 0.29 and 0.38 on average for the initial coding in the T-No, T-Voting and T-Voting-ST treatments, respectively. The reconsideration process improved the Kappas. After the coders' independent reconsideration, the Kappas became 0.88, 0.90 and 0.87 in the T-No, T-Voting and T-Voting-ST treatments, respectively. Appendix C.3 includes the Kappa value for each individual code, indicating that almost all codes have high Kappa values. Regression analyses in the following subsection utilize codes whose Kappa is above 0.4. 0.4 is often used as a criterion for reliability of codes, for example, in Landis and Koch (1977), Bougheas *et al.* (2013) and Cason *et al.* (2012). In the present paper, 95% of codes have Kappa values greater than 0.4.

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 in the coding exercise 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 (e.g., believing that own payoff is maximized mathematically by having the zero sanction rate and zero contribution)."

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 if it is equal to or greater than 0.4; non-deterrent if it is less than 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 team members recognize the relationship between sanction rates and material incentives in the game. The sanction rate should be set equal to or greater than 0.4 to induce other teams to contribute fully to the public account. A rational team 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. These four codes were on average marked 6.5%, 6.8%, 10.7% and 3.9% per period per team, 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 harms efficient institutional formation. Second, C6 is a significantly 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 log 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. An example of a team's log is as follows:

Member ID1: whey did that team put 5
Member ID2: don't they legit just make less money
Member ID1: yeh
Member ID2: by doing that
Member ID2: ??????????????????????????????
Member ID2: im so confused
Member ID1: need a high fine rate again to try and discourage them
Member ID3: they are making all lose money
Member ID2: lol
Member ID 1: same best if all three teams work together
Member ID 2: I actually have no clue
Member ID 1: I like we aren't competing with them
Member ID 3: we have to put 1.2
Member ID 1: yeah deffo agree
Member ID2: definitely

This result collaborates with the fact that the sanction rate of 1.2 was the most popular among the deterrent sanction rates (Figure 5). It should be noted here that Kamei *et al.* (2015) also found that given an option to vote, most groups enacted the strongest sanction rate even when clearly beyond the deterrent level.

In summary, it can be concluded that teams' frequent voting for strongly deterrent sanction rates were driven by their negative reactions and intolerance towards low contributions, and their learning about its impact (recall that strong punishment smoothly altered the teams' uncooperative behaviors as evidenced in Figures 2 and 3).

6.2. Informal Punishment Decisions in the IS Scheme

Units, whether individual or teams, inflicted punishment not only pro-socially but also anti-socially (Section 4.3.2). Four codes are considered in the coding exercise 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- versus pro-social punishment classification (Hermann *et al.*, 2008). As in the earlier analyses, four more codes (F5 to F9) are also considered in this analysis based on the perverse versus non-perverse punishment definition (Cinyabuguma *et al.*, 2006). The analysis result shown in this subsection is based on Codes F1 to F4. Results are similar when Codes F5 to F9 are instead used (Appendix C.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' costly punishment activities are indeed 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 also the coefficient estimates are much larger than for F2 and F4, respectively). Therefore, it can be concluded that punishment motives are heterogeneous (Kamei, 2014), and units have clear intentions to punish pro-socially, or anti-socially, under certain conditions, parallel to the observations from the decision data.

The regression results reveal three further reasonable patterns. First, emotion (Code F16: "Suggests punishment as an emotional response") drives punishment, consistent with the findings from neuroscience research (e.g., de Quervain *et al.*, 2004). 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, positive punishment costs (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	---
# of left-censored observations (0.0)	205	---	94	---	111	---
# of right-censored observations (1.2)	303	---	91	---	212	---
Log likelihood	-446.718	---	-195.108	---	-216.428	---
Wald χ^2	136.33	---	75.53	---	78.49	---
Prob > Wald χ^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 C.4.a. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .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	---
# of left-censored observations (0)	535	---	315	---	220	---
# of right-censored observations (20)	5	---	3	---	2	---

Log likelihood	-363.288	---	-208.870	---	-86.680	---
Wald χ^2	172.55	---	150.91	---	n.a.	---
Prob > Wald χ^2	0.000	---	0.000	---	n.a.	---

Notes: Tobit regressions. Decision-making unit random effects were included to control for the panel structure. Codes associated with the definition of anti-social/pro-social punishment (F1, F2, F3, F4) were used in this table. The regression includes all F codes (except F5 to F8) 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 C.4.b. It should be noted that the alternative definition of punishment is perverse or non-perverse (Section 4.3.2). A regression result with codes associated with the definition of perverse/non-perverse punishment (F5, F6, F7, F8) is omitted to conserve space since it generates qualitatively similar results – See Appendix C.4.b for the result. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

6.3. Contribution Decisions

While both contribution levels and dynamics differed drastically according to the presence of the sanctioning schemes (Figure 2, online Appendix Figure B.1), coding analyses, summarized in Table 7, suggest qualitatively similar patterns for all treatments. First, units with unconditional willingness to cooperate contributed large amounts (variable i). Apart from such altruistic motives, some units also aimed to encourage other units to cooperate, or to avoid discouraging already cooperative teams, through contributing large amounts (variable ii). Second, however, some units 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).

Recall, however, that the analysis in Section 5 revealed a different distribution of units' contribution types in the FS scheme compared with the NS or IS scheme. The former decided contribution amounts in response to the sanction rate collectively enacted in their group. Thus, in order to explore the reasoning in greater depth, the following two codes are considered in the regression analysis:

D9: "Discusses contribution to avoid fines e.g. suggests high contribution to avoid fines."
D10: "Discusses contribution based on material motives (i.e., contribute large amounts if the enforced sanction rate is deterrent; contribute little if it is non-deterrent)."

The estimation result shown in column (2) of Table 7 indicates that units' desire to avoid receiving fines, rather than material calculations, drove their strong contribution behaviors. This means that positive effects of formal 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 levels of cognitive ability to understand the material incentive structure in the game.

The earlier analysis in Section 5 also revealed that units' contribution types were roughly similar for the NS and IS schemes (Table B.9). One may wonder whether their reasoning was also relatively the same for the two conditions. In order to explore whether informal punishment opportunities may have affected decisions to contribute, four codes specific to the IS scheme, i.e., beliefs and recent experiences regarding being punished, are considered in the analysis. The estimation result indicates that units who discussed their experiences being pro-socially punished in the last period (and hence cared about such incidents) tended to increase contributions in the current period. However, except for this positive tendency, none of the other codes has a significant coefficient estimate (see column (3) of Table 7). This reinforces the conjecture made in Section 5 that the mere presence of IS may raise groups' cooperation levels, and units' reciprocal tendencies detected in variables ii and iv successfully sustained the positive cooperation norms in the group.

Table 7: *Reasoning behind Units' Contribution Decisions*

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

	(1) No scheme		(2) Under FS scheme		(3) Under IS scheme	
Codes included in the regression:	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
v. Confusion, errors, mistakes (codes A12, D11, E14 dummies)	-0.650	0.771	-4.381*	2.392	0.205	1.578
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	---	---
viii. Contribute based on belief being punished (code E5 dummy)	---	---	---	---	0.886	1.303
ix. Decrease contribution if not punished in previous rounds (code E7 dummy)	---	---	---	---	-1.075	1.250
x. Increase contribution if pro-socially punished in previous rounds (code E8 dummy)	---	---	---	---	2.412*	1.323
xi. Decrease contribution if anti-socially punished in previous rounds (code E10 dummy)	---	---	---	---	1.955	1.702
# of observations	1,128	---	672	---	648	---

# of left-censored observations (0)	170	---	26	---	17	---
# of right-censored observations (20)	253	---	536	---	473	---
Log likelihood	-2636.596	---	-507.868	---	-588.738	---
Wald χ^2	749.45	---	212.5	---	254.42	---
Prob > Wald χ^2	0.000	---	0.000	---	0.000	---

Notes: Tobit regressions. Decision-making unit random effects were included to control for the panel structure. 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 C.4.c. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

6.4. Scheme Choice

The remaining analysis is on communication dialogues related to team 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 teams 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 C.4.d includes the detail of the estimation results.

7. Conclusion

Team decision-making is ubiquitous in real-world organizations, 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 these 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 in organizations 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. Using individuals as the

decision-making unit could be a nice simplification if the following implicit 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 choices and behaviors under the selected institutions between individuals and teams to confirm this assumption. 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 and therefore overcome free riding in groups more effectively. In the experiment, decision-making units, teams or individuals, were given an 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 enacted deterrent sanction rates by voting much more frequently than individuals. When peer-to-peer punishment was instead selected, teams inflicted costly punishment more frequently on low contributors than individuals.

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 (each group in a team treatment had nine subjects). It should be acknowledged that the sizes of teams and/or groups could be much larger in real organizations. 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 definitely be a useful robustness check to study the same research questions by changing the group size and/or team size in the framework of this paper. For another example, the three members in a team 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 piece is the most standard setup in the current experimental literature on team decision-making (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 effects 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, and (if yes) how it depends according to the team composition (e.g., gender composition). Of course, needless to say, the finding of this research would also open up further 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 compared with individuals, even if they face the same incentive structure. This means that the conventional assumption taken in the theory of the firm may not be accurate for real human workers.

References

Aboramadan, Mohammed, 2020. "Top management teams characteristics and firms performance: literature review and avenues for future research." *International Journal of Organizational Analysis,* 29(3), 603-628.

Ahn, T. K., Elinor Ostrom, David Schmidt, Robert Shupp, and James Walker, 2001. "Cooperation in PD games: Fear, greed, and history of play." *Public Choice*, 106(1/2), 137-155.

Alchian, Armen, and Harold Demsetz, 1972. "Production, Information Costs, and Economic Organization." *American Economic Review*, 62(5), 777-795.

Anderson, Christopher, and Louis Putterman, 2006. "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism." *Games and Economic Behavior*, 54(1), 1-24.

Andreoni, James, 1988. "Why Free Ride? Strategies and Learning in Public Goods Experiments." *Journal of Public Economics*, 37, 291-304.

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

Bainbridge, Stephen, 2002. "Why a Board? Group Decisionmaking in Corporate Governance." *Vanderbilt Law Review*, 55(1).

Bednar Jenna, Yan Chen, Tracy Liu, and Scott Page, 2012. "Behavioral Spillovers and Cognitive Load in Multiple Games: an Experimental Study." *Games and Economic Behavior*, 74, 12-31.

Blinder, Alan, and John Morgan, 2005. "Are Two Heads Better than One? Monetary Policy by Committee." *Journal of Money, Credit and Banking*, 37(5), 789 -811.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch, 2014. "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review*, 71, 117-120.

Bornstein, Gary, Tamar Kugler, and Anthony Ziegelmeyer, 2004. “Individual and group decisions in the centipede game: Are groups more ‘‘rational’’ players?” *Journal of Experimental Social Psychology*, 40, 599-605.

Bornstein, Gary, and Ilan Yaniv, 1998. “Individual and Group Behavior in the Ultimatum Games: Are Groups More “Rational” Players?” *Experimental Economics*, 1, 101-108.

Bougheas, Spiros, Jeroen Nieboer, and Martin Sefton, 2013. “Risk-taking in social settings: Group and peer effects.” *Journal of Economic Behavior & Organization*, 92, 273-283.

Brosig, Jeannette, Joachim Weimann, and Axel Ockenfels, 2003. “The Effect of Communication Media on Cooperation.” *German Economic Review*, 4(2), 217-241.

Carmeli, Abraham, Zachary Sheaffer, and Mayrav Yitzack Halevi, 2009. “Does participatory decision-making in top management teams enhance decision effectiveness and firm performance?” *Personnel Review*, 38(6), 696-714.

Casari, Marco, and Luigi Luini, 2009. “Cooperation under alternative punishment institutions: An experiment.” *Journal of Economic Behavior & Organization*, 71(2), 273-282.

Cason, Timothy, and Vai-Lam Mui, 1997. “A Laboratory Study of Group Polarisation in the Team Dictator Game.” *Economic Journal*, 107, 1465-83.

Cason, Timothy, and Vai-Lam Mui, 2015. “Rich communication, social motivations, and coordinated resistance against divide-and-conquer: A laboratory investigation.” *European Journal of Political Economy*, 37, 146-159.

Cason, Timothy, Anya Savikhin, and Roman Sheremeta, 2012. “Behavioral Spillovers in Coordination Games.” *European Economic Review*, 56, 233-245.

Cason, Timothy, Roman Sheremeta, Jingjing Zhang, 2012. “Communication and efficiency in competitive coordination games.” *Games and Economic Behavior*, 76, 26-43.

Certo, Trevis, Richard Lester, Catherine Dalton, and Dan Dalton, 2006. “Top Management Teams, Strategy and Financial Performance: A Meta-Analytic Examination.” *Journal of Management Studies*, 43(4), 813-839.

Charness, Gary, and Mattias Sutter, 2012. “Groups Make Better Self-Interested Decisions.” *Journal of Economic Perspectives*, 26, 157-176.

Chaudhuri, Ananish, 2011. “Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature.” *Experimental Economics*, 14(1), 47 - 83.

Cinyabuguma, Matthias, Talbot Page, and Louis Putterman, 2006. “Can second-order punishment deter perverse punishment?” *Experimental Economics*, 9, 265-279.

Cohen, Jacob, 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement*, 20, 37-46.

Cohen, Susan, and Diane Bailey, 1997. “What Makes Teams Work: Group Effectiveness Research from the Shop Floor to the Executive Suite.” *Journal of Management*, 23(3), 230-290.

Cooney, Richard, 2004. “Empowered self-management and the design of work teams.” *Personnel Review*, 33(6), 677-692.

Cooper, David, and John Kagel, 2005. “Are two heads better than one? Team versus individual play in signaling games.” *American Economic Review*, 95(3), 477-509.

Cox, Caleb, and Brock Stoddard, 2018. “Strategic thinking in public goods games with teams.” *Journal of Public Economics*, 161, 31-43.

de Quervain, D.J., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Urich Schnyder, Alfred Buck, and Ernst Fehr, 2004, “The Neural Basis of Altruistic Punishment.” *Science* 305(5688), 1254-1258.

Denant-Boemont, Laurent, David Masclet, and Charles Noussair, 2007. “Punishment, counterpunishment and sanction enforcement in a social dilemma experiment.” *Economic Theory*, 33, 154-167.

Delarue, Anne, Geert Van Hootegem,Stephen Procter,Mark Burridge, 2007. “Teamworking and organizational performance: A review of survey-based research.” *International Journal of Management Reviews*, 10(2), 127-148.

Devine, Dennis, Laura Clayton, Jennifer Philips, Benjamin Dunford, and Sarah Melner, 1999. “Teams in Organizations: Prevalence, Characteristics, and Effectiveness.” *Small Group Research*, 30(6), 678-711.

Ertan, Arhan, Talbot Page, and Louis Putterman, 2009. “Who to punish? Individual decisions and majority rule in mitigating the free rider problem.” *European Economic Review*, 53, 495-511.

Eurofound and Cedefop, 2020. “European Company Survey 2019: Workplace practices unlocking employee potential.” *European Company Survey 2019 Series*, Publications Office of the European Union, Luxembourg.

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebmer, 2000. “A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence.” *American Economic Review*, 90(1), 247-264.

Fehr, Ernst, and Simon Gächter, 2000. “Cooperation and Punishment in Public Goods Experiments.” *American Economic Review*, 90(4), 980-994.

Fehr, Ernst, and Simon Gächter, 2002. “Altruistic punishment in humans.” *Nature*, 415, 137-140.

Fehr, Ernst, and Schmidt Klaus, 2006. “The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories,” in Kolm S.-G. and Ythier J. M. (eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, pp. 615-91. North Holland.

Fehr, Ernst, and Tony Williams, 2018. “Social Norms, Endogenous Sorting and the Culture of Cooperation.” University of Zurich Department of Economics Working paper No. 267.

Feri, Francesco, Bernd Irlenbusch, and Matthias Sutter, 2010. “Efficiency gains from team-based coordination—large-scale experimental evidence.” *American Economic Review*, 100, 1892-912.

Fischbacher, Urs, 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10(2), 171-178.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr, 2001. "Are people conditionally cooperative? Evidence from a public goods experiment." *Economics Letters*, 71(3), 397-404.

Fischbacher, Urs, and Simon Gächter, 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." *American Economic Review*, 100(1): 541-56.

Gächter, Simon, Elke Renner, and Martin Sefton, 2008. "The Long-Run Benefits of Punishment." *Science*, 322(5907), 1510.

Gibbons, Robert, NikoMatouschek, and John Roberts, 2013. "Decisions in Organizations" (Chapter 10) included in *The Handbook of Organizational Economics* (edited by R. Gibbons and J. Roberts), pp. 373-431, Princeton University Press.

Gillet, Joris, Arthur Schram, and Joep Sonnemans, 2009. "The tragedy of the commons revisited: the importance of group decision-Making." *Journal of Public Economics*, 93, 785-97.

Grant, Robert, 1996. "Toward a Knowledge-based theory of the firm." *Strategic Management Journal*, 17, 109-122.

Grosse, Stefan, Louis Putterman, and Bettina Rockenbach, 2011. "Monitoring in Teams: Using Laboratory Experiments to Study a Theory of the Firm." *Journal of European Economic Association*, 9(4), 785-816.

Gunnthorsdottir, Anne, Daniel Houser, and Kevin McCabe, 2007. "Disposition, history and contributions in public goods experiments." *Journal of Economic Behavior & Organization*, 62(2), 304-15.

Gürerk, Ozgür, Bernd Irlenbusch, and Bettina Rockenbach, 2006. "The competitive advantage of sanctioning institutions." *Science*, 312(5770), 108-111.

Guzzo, Richard, Marcus Dickson, 1996. "Teams in Organizations: Recent Research on Performance and Effectiveness." *Annual Review of Psychology*, 47, 307-338.

Hamilton, Barton, Jack Nickerson, and Hideo Owan, 2003. "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation." *Journal of Political Economy*, 111(3), 465-497.

Hauser, Oliver, David Rand, Alexander Peysakhovich, and Martin Nowak, 2014. "Cooperating with the future." *Nature*, 511, 220-223.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, 2008, "Antisocial punishment across societies." *Science*, 319, 1362-1367.

Holmstrom, Bengt, 1982. "Moral Hazard in Teams." *Bell Journal of Economics*, 13(2), 324-340.

Ichniowski, Casey, Kathryn Shaw, and Giovanna Prennushi, 1997, "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines." *American Economic Review*, 87, 291-313.

Isenberg, Daniel, 1986. “Group polarization: A critical review and meta-analysis.” *Journal of Personality and Social Psychology*, 50(6), 1141-1151.

Kagel, John, 2018. “Cooperation through communication: Teams and individuals in finitely repeated Prisoners’ dilemma games.” *Journal of Economic Behavior & Organization,* 146, 55-64.

Kagel, John, and Peter McGee, 2016. “Team versus Individual Play in Finitely Repeated Prisoner Dilemma Games”. *American Economic Journal: Microeconomics*, 8(2), 253-76.

Kamei, Kenju, 2014. “Conditional Punishment.” *Economics Letters*, 124(2), 199-202.

Kamei, Kenju, 2016. “Democracy and resilient pro-social behavioral change: an experimental study.” *Social Choice and Welfare*, 47(2), 359-378.

Kamei, Kenju, 2019a. “Cooperation and endogenous repetition in an infinitely repeated social dilemma.” *International Journal of Game Theory*, 48(3), 797-834.

Kamei, Kenju, 2019b. “The power of joint decision-making in a finitely-repeated dilemma.” *Oxford Economic Papers*, 71(3), 600-622.

Kamei, Kenju, 2021. “Teams do inflict costly third party punishment as individuals do: experimental evidence.” *Games*, 12(1), 22.

Kamei, Kenju, and Louis Putterman, 2015. “In Broad Daylight: Fuller Information and Higher-Order Punishment Opportunities can Promote Cooperation.” *Journal of Economic Behavior & Organization*, 120, 145-159.

Kamei, Kenju, Louis Putterman, and Jean-Robert Tyran, 2015. “State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods.” *Experimental Economics*, 18, 38-65.

Kerr, Norbert, and Scott Tindale, 2004. “Group Performance and Decision Making.” *Annual Review of Psychology,* 55, 623-655.

Kersley, Barbara, Carmen Alpin, John Forth, Alex Bryson, Helen Bewley, Gill Dix, and Sarah Oxenbridge, 2005. “Inside the Workplace: Findings from the 2004 Workplace Employment Relations Survey.” [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/34310/05-1057-wers5-2004-inside-the-workplace-first-findings.pdf> [Accessed 25 October 2018].

Kocher, Martin, and Mattias Sutter, 2005. “The Decision Maker Matters: Individual versus Group Behavior in Experimental Beauty-Contest Games.” *Economic Journal*, 115, 200-223.

Kosfeld, Michael, Akira Okada, and Arno Riedl, 2009. “Institution Formation in Public Goods Games.” *American Economic Review*, 99(4), 1335-55.

Kreps, David, Paul Milgrom, John Roberts, Robert Wilson, 1982. “Rational cooperation in the finitely repeated prisoners' dilemma.” *Journal of Economic Theory*, 27(2), 245-252.

Kugler, Tamar, Edgar Kausel, and Martin Kocher, 2012 "Are groups more rational than individuals? A review of interactive decision making in groups." *WIREs Cognitive Science*, 3, 471-482.

Kugler, Tamar, Gary Bornstein, Martin Kocher, and Mattias Sutter, 2013. "Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy." *Journal of Economic Psychology*, 28, 646-57.

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

Lawler, Edward, Susan Albers Mohrman, and Gerald Ledford, 1992. *Employee Involvement and Total Quality Management: Practices and Results in Fortune 1000 Companies*. San Francisco: Jossey-Bass.

Lawler, Edward, Susan Albers Mohrman, and Gerald Ledford, 1995. *Creating High Performance Organizations: Impact of Employee Involvement and Total Quality Management*. San Francisco: Jossey-Bass.

Ledyard, John, 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, Andreas, and Lauri Sääksvuori, 2012. "Communication in intergroup conflicts." *European Economic Review*, 56(6), 1136-1147.

Marschak, Jacob, and Roy Radner, 1972. *Economic Theory of Teams*. New Haven: Yale University Press.

McLachlan, Geoffrey, and David Peel, 2000. *Finite Mixture Models*. New York: Wiley.

Moffatt, Peter, 2016. *Experimetrics: Econometrics for experimental economics*, Macmillan International Higher Education.

Moscovici, Serge, and Marisa Zavalloni, 1969. "The group as a polarizer of attitudes." *Journal of Personality and Social Psychology*, 12(2), 125-135.

Müller, Wieland, and Fangfang Tan, 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 and Economic Behavior*, 82, 658-74.

Nicklisch, Andreas, Louis Putterman, and Christian Thöni, 2021. "Trigger-happy or precisionist? On demand for monitoring in peer-based public goods provision." *Journal of Public* Economics, 200, 104429.

Nikiforakis, Nikos, and Hans-Theo Normann, 2008. "A Comparative Statics Analysis of Punishment in Public-Good Experiments." *Experimental Economics*, 11, 358-369.

Ostrom, Elinor, James Walker, and Roy Gardner, 1992. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review*, 86(2), 404-417.

Pfeffer, Jeffrey, 1998. "Seven Practices of Successful Organizations." *California Management Review,* 40(2), 96-124.

Robert, Christopher, and Peter Carnevale, 1997. “Group Choice in Ultimatum Bargaining.” *Organizational Behavior and Human Decision Process*, 72, 256-279.

Schopler, John, and Chester A Insko, 1992. “The discontinuity Effect in Interpersonal and Intergroup Relations: Generality and Mediation.” In W. Stroebe and M. Hewstone (Eds.), *European review of social psychology*, 3, 121-151). Oxford, England: John Wiley & Sons.

Schopler, John, Chester A Insko, Kenneth A Graetz, Stephen M Drigotas, and Valerie Smith, 1991. “The generality of the individual-group discontinuity effect: Variations in positivity-negativity of outcomes, players' relative power, and magnitude of outcomes.” *Personality and Social Psychology Bulletin*, 17(6), 612-624.

Schopler, John, Chester A Insko, Kenneth A Graetz, Stephen M Drigotas, Valerie Smith, and Kenny Dahl, 1993. “Individual-group discontinuity: Further evidence for mediation by fear and greed.” *Personality and Social Psychology Bulletin*, 19(4), 419-431.

Sobel, Joel, 2005. “Interdependent Preferences and Reciprocity.” *Journal of Economic Literature*, 43(2), 392-436.

Sunstein, Cass, 2007. *Republic.com 2.0*, Princeton University Press.

Sutter, Mattias, 2007. “Are Teams prone to myopic loss aversion? An Experimental Study on Individual Versus Team Investment Behavior.” *Economics Letters*, 97, 128-132.

Sutter, Mattias, 2009. “Individual behavior and group membership: comment.” American *Economic Review*, 99, 2247-2257.

Sutter, Matthias, Stefan Haigner, and Martin Kocher, 2010. “Choosing the Carrot or the Stick? – Endogenous Institutional Choice in Social Dilemma Situations.” *Review of Economic Studies*, 77(4): 1540-1566.

Traulsen, Arne, Torsten Röhl, and Manfred Milinskim, 2012. “An economic experiment reveals that humans prefer pool punishment to maintain the commons.” *Proceedings of the Royal Society B*, 279, 3716-3721.

Tyran, Jean-Robert, and Lars Feld, 2006. “Achieving Compliance when Legal Sanctions are Non-deterrent.” *Scandinavian Journal of Economics*, 108(1), 135-156.

van Elten, Jonas, and Stefan Penczynski, 2020. “Coordination games with asymmetric payoffs: An experimental study with intra-group communication.” *Journal of Economic Behavior & Organization*. 169, 158-188.

Wildschut, Tim, Brad Pinter, Jack L Vevea, Chester A Insko, and John Schopler, 2003. “Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect.” *Psychological Bulletin*, 129(5), 698-722.

Zhang, Boyu, Cong Li, Hannelore Silva, Peter Bednarik, and Karl Sigmund, 2014. “The evolution of sanctioning institutions: an experimental approach to the social contract.” *Experimental Economics*, 17(2), 285-303.