# *Deliberation, Democracy, and Mechanisms for Cooperation*

KATY TABERO

# Deliberation, Democracy, and Mechanisms for Cooperation

by

Katy Fiona Tabero

This thesis is submitted for the degree of Doctor of Philosophy

Department of Economics

Durham University

2023

## Abstract

This thesis explores group decision-making and mechanisms to encourage cooperation through three experimental studies.

Study one uses a public goods game (PGG) with informal and formal sanction mechanisms to understand how team decision-making differs from individual decision-making in a democratic institutional setting. Teams consistently outperform individuals when sanctioning schemes are available, by selecting higher sanction rates when choosing the formal scheme and pro-socially targeting punishment toward low-cooperators when using the informal scheme. This improved decision-making appears to be a result of deliberation and has implications for using team decision-making to overcome moral hazards.

Building on this, study two examines team behaviour in a real effort experiment to understand the impact of democratic decision-making. Specifically, in one treatment teams may vote on whether to implement a policy that reduces the returns from free-riding within their group, while in the other treatment, this policy is randomly implemented. Teams exhibit significantly higher productivity when they are able to democratically decide whether to implement the policy, regardless of the vote outcome. While teams in these treatments also increase their time free-riding, the higher productivity compensates for this and so it does not harm overall production. As in the first chapter, this study highlights the benefits of autonomous team-decision making in improving cooperation.

Study three explores how a group may encourage cooperation to prevent a more costly problem in a two-stage PGG. Subjects complete real effort tasks that either reward them directly or improve the payoff schedule in the following stage, forming a second-order social dilemma. Free-riding does not dominate the pre-stage nor does cooperation decline as strongly as observed in other PGG, demonstrating how leveraging fewer resources to overcome related social dilemmas can make cooperation easier. Further, providing a simple cost- and ramification-free feedback mechanism considerably increases the level of cooperation observed.

# Contents

## List of Tables

## List of Figures

## Declaration

No part of the thesis has previously been submitted for a degree in this or any other institution.

Chapter 2 is based on a joint research paper with Prof. Kenju Kamei. An earlier draft of this paper titled, “The Individual-Team Discontinuity Effect on Institutional Choices: Experimental Evidence in Voluntary Public Goods Provision” is available as a discussion paper for the Keio-IES Discussion Paper Series 2022-015, Institute for Economics Studies, Keio University. Both authors contributed significantly to this project. I was heavily or solely responsible for the literature review, the programming, organization, and implementation of the experiments, the completion of data analysis (the direction of which was jointly discussed), and the organization, and implementation of the content analysis.

Chapter 3 is based on a joint research paper with Prof. Kenju Kamei. Both authors contributed significantly to this project. Both authors contributed significantly to this project. I was heavily or solely responsible for the literature review, the programming, organization, and implementation of the experiments, the completion of data analysis (the direction of which was jointly discussed), and the organization, and implementation of the content analysis.

Chapter 4 is based on a joint research paper with Prof. Kenju Kamei, Prof. Louis Putterman, and Prof. Jean-Robert Tyran. All four authors contributed significantly to the design and drafting of this project. I was heavily or solely responsible for the literature review, the programming, organization, and implementation of the experiments, the completion of data analysis, and the experiment design section.

## Acknowledgments

I am hugely grateful to my supervisors both at Durham and further afield. Specifically, I would like to thank Prof. Kenju Kamei for their guidance, collaborative efforts, and the opportunities presented throughout my PhD journey, especially given the difficult circumstances of the pandemic and their relocation midway through. Through our joint projects, I have become a much more confident and analytic researcher. I would also like to thank John Hey who not only hosted me at the York Exec Lab for two of the projects, but also took on the role of external supervisor and spent the time to welcome me to York. Despite only joining me for the latter half of my PhD process, I am also extremely grateful to Dr. Daniel Li and Dr. Andis Sofianos for taking over the roles of first and second supervisor. The discussions I have had with both of you have been immensely helpful (and interesting) in developing the projects, even at a late stage.

In addition to those mentioned, I am grateful to my collaborators, Prof. Louis Putterman and Prof. Jean-Robert Tyran, for their time, experience, and effort towards our collaborative projects. I feel very fortunate to have the opportunity to work with both. I would also like to thank those who have provided feedback or reviews over the course of my studies, including Dr. Leslie Reinhorn, Dr. Jinrui Pan, Prof. Nejat Anbarci, Prof. Pedro Dal Bó, those who attended my presentations at ESA Bologna 2022, the SEET Workshop in Valencia 2023, and internal seminars held at Durham University, and several anonymous referees.

Practically, my projects would not have been possible without help from those at the York Exec Lab and the Experimental and Behavioural Economics Lab of Newcastle University, including John Hey, Mark Wilson, Dr. Till Weber, Dr. Irene Mussio, Dr. Morgan Beeson and Dr. Matt Walker. Further I am thankful to administrative staff in the Finance Office at Durham University who made the organization and payments of experiment subjects as straight forward as possible and saved me countless hours of work.

I gratefully acknowledge the financial support I received from the Economic and Social Research Council (ESRC), in the form of both the Doctoral Training Partnership Studentship and the Doctoral Training Partnership Research Training Support Grant, funding from the Department of Economics at Durham University, and from a grant-in-aid from the Japan Center for Economic Research (awarded to Prof. Kenju Kamei), the grant-in-aid from the Murata Science Foundation (awarded to Prof. Kenju Kamei), and funding awarded for subject payment in Chapter 4 from Prof. Jean-Robert Tyran.

Last, but certainly not least, I am immensely grateful for the support of my family and friends who I could not have endured this journey without. I especially wish to thank my parents, Sylvia and Andrzej, my brother, Alex, and my partner Niall for always having confidence in me, even when I did not.

# 1. Motivation

Groups (or teams) are popular decision-making and/or work units in firms and beyond, whether making decisions to organize themselves, as with small businesses or households, or managing resources and the governance of others, as with management teams or councils. Yet, within a group, each member has an opportunity to free-ride on the effort of others by directing their resources towards private activities that benefit only them. While the problem of moral hazard is not a new one, finding structures or mechanisms to counter it (especially in the variety of circumstances it presents itself) remains an active area of research.

Building on previous social dilemma literature, this thesis uses three individual experimental studies to investigate how the decision-making unit, method of decision-making, and mechanisms available to encourage cooperation affect the decisions made or effort levels observed in a social dilemma environment. The first two studies address how deliberation and democracy impact the quality of the decisions made, while all three studies investigate mechanisms designed to discourage free-riding, including both deterrent and non-deterrent variations.

In the first study, decisions are compared between individuals and teams (small groups) as decision-making units in an institutional public goods game. Teams are often found to be more competitive or self-interested than individuals in experimental settings, attributed to group dynamics and stronger cognitive ability (see Section 2.2). As teams are such prominent decision-making units in the real world, it is important to understand under what circumstances they are more competitive and whether it is possible to encourage inter-team cooperation. Specifically, if it is the case that teams do behave as if more cognitively able, then it is possible that they are able to make better use of cooperative mechanisms where available. A laboratory experiment is used to test differences in preferences for (and usage of) sanction schemes depending on the decision-making unit, individual or team. Units vote whether to implement a formal or informal sanction scheme in a finitely repeated public goods game and then play under the endogenously selected rules. Teams are found to outperform individuals consistently when sanctioning schemes are available, regardless of which scheme is selected, highlighting the benefits of deliberative decision-making. When the formal sanction scheme is selected, teams vote for deterrent sanction rates much more frequently than individuals, shifting the privately optimal behaviour to full contribution to the group account. When the informal sanction scheme is selected, teams inflict costly punishment more frequently on low contributors than individuals, thereby reducing the relative frequency of "misdirected" punishment among teams. The results underscore first the effectiveness of both formal and informal sanction

schemes when well-utilized by the respective unit, and second the higher ability of teams to take advantage of such mechanisms to establish and sustain cooperation more quickly and efficiently.

The second study builds on the first to explore the impact of team decision-making in a novel collaborative real effort experiment. While the first chapter argues that teams make stronger decisions than individuals in an institutional setting, this chapter explores whether the autonomy to make such a decision in itself has an impact on behaviour, a phenomenon known as the "dividend of democracy." In the experiment, subjects are randomly assigned to a team of three and jointly solve a collaborative real effort task under a revenue-sharing rule in their group, consisting of two other teams. Throughout the task-solving period, each individual worker can privately and independently shirk by playing Tetris while benefiting from the efforts of others. Before beginning the task-solving phase, teams in one treatment may deliberate and vote on whether to implement a policy that punishes free-riding (by reducing the return from shirking) in their group, while in the other treatment, the policy is randomly implemented. Results show that some teams in the workplace will voluntarily reduce their members' private benefits to achieve the group optimum in a social dilemma, and further that such endogenous decision-making in itself enhances work productivity (per work time production). Teams exhibit significantly higher productivity when they are able to democratically decide whether to reduce the return from shirking by voting than when the policy implementation is randomly decided from above, irrespective of the actual policy implementation outcome. This demonstrates that democratic culture directly affects behaviour and has implications for workplace organization. Despite this, the workers under democracy also increase their shirking, possibly as a result of fatigue owing to their stronger work efforts. Yet, due to the higher levels of productivity experienced by the teams in this treatment, overall production is not reduced. Supporting the findings of the first study, this chapter highlights the benefits of worker participation in decision-making through communication and voting, not only in terms of their willingness to utilise sanction mechanisms to achieve a shared goal, but also in terms of external benefits to effort and productivity.

The third study develops on the first two by considering further mechanisms to improve group cooperation. A two-stage public goods game is used to show how smaller collective efforts can be used to leverage against larger social dilemmas, such as leveraging civic engagement to hold a government accountable (one frame used in the study). In this experiment, subjects may complete real effort tasks in a pre-stage in order to reduce inefficiency affecting the group account in the following main stage. Two types of tasks are available, one that directly rewards the individual, and another which improves the payoff schedule of the main stage public good game; as a result, there is a second-order social dilemma. Results show that, despite the strong incentive to complete tasks that

only benefit oneself, free-riding does not dominate the pre-stage nor does cooperation decline as strongly as observed in other standard public goods games. Further, providing a simple cost- and ramification-free feedback mechanism to a smaller subgroup within each group considerably increases the level of cooperation observed. While subjects were unable to communicate directly in this setting, feedback was used to signal the approval of other group members' actions and support a cooperative atmosphere within the subgroup, an element also found to be important in the second study whereby informal encouragement spurred greater team efforts.

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

## 2.1. Introduction

Teams have seen increasing popularity as a decision-making unit within organizations in the last half a century; this applies to both the public and private sector, and across a breadth of industries (see Lawler *et al.* 1992, 1995; Devine *et al.* 1999; Kersley *et al.* 2005). For example, Eurofound (2020) found that around 70% of workers in the EU27 claimed to work as part of a team. 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 proposed to improve 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).

A central issue still understudied in the literature on institutions is how teams form institutions and behave in a social dilemma when compared to individuals. Scholars studying workers' performances and interactions in social dilemmas 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 much to contribute to the group (exert costly effort). Standard theory suggests that socially optimal effort provision cannot be achieved in typical environments due to workers' free riding, whereby they pursue their own self-interest – a phenomenon called "1/N problem" (e.g., Alchian and Demsetz 1972). A large number of experiments have been conducted to examine worker behaviours in such voluntary provision of public goods when *individuals* are the decision-making unit (see, e.g., Ledyard [1995] and Chaudhuri [2011] for a survey). They show that real individuals' behaviours have some similarities to the theoretical suggestions: 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 behaviours with repetition (e.g., Fischbacher and Gächter 2010). However, the literature simultaneously shows that groups can sustain cooperation when the members can voluntarily monitor their peers' contribution behaviours (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 formal

(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* formal governance by voting, finding that without any guidance, groups can achieve high efficiency through such endogenous institution 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, little 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 in the literature on institutions. Hence, the neglect of teams' self-governance possibility is natural, and the use of *individuals* in a laboratory can be thought of as a simplification for experimentation in the literature. However, this assumption may not be correct according to the findings from another, but substantial, literature on group or team decision-making. This research area proposes the so-called "individual-team discontinuity effect" (simply "discontinuity effect," hereafter): teams may behave more efficiently than individuals (see, e.g., Charness and Sutter [2012], Kugler *et al.* [2012] and Kerr and Tindale [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; Sutter 2005; Kocher *et al.* 2006), ultimatum games (e.g., Robert and Carnevale 1997; Bornstein and Yaniv 1998), signaling games (e.g., Cooper and Kagel 2005), centipede games (e.g., Bornstein *et al.* 2004), trust games (e.g., Kugler *et al.* 2007), coordination games (e.g., Feri *et al.* 2010), monetary policy decisions (e.g., Blinder and Morgan 2005), and public goods games (e.g., Auerswald *et al.* 2018, Kamei 2019b). It is possible that teams also construct institutions differently from individuals in the voluntary provision of public goods.

This paper provides the first experiment to investigate whether teams outperform individuals in the context of a social dilemma when teams as a decision-making unit govern their assigned group *through communication and institution formation by voting*. Specifically, each group can use either a formal sanction scheme or an informal (peer-to-peer) sanction scheme. In a repeated public goods game, members of each team communicate with one another to make joint voting and contribution decisions. The institutional formation and their behaviours under constructed institutions are compared against the case where the units are individuals.

There are two possible mechanisms that predict that teams behave differently from individuals in the present context. The first mechanism is the so-called Condorcet's (1785) jury theorem and behavioural public choice theorem (e.g., Ertan *et al.*, 2009). This states that if the probability of an individual voting for an option in a binary choice is larger than ½, say 0.75, then the probability that

the option is enacted under majority voting is larger than the percentage of individual supporting votes, say 0.85 > 0.75. Likewise, if the probability of an individual voting for an option is smaller than ½, say 0.25, then the probability that the option is enacted under majority voting is even smaller than the percentage of individual supporting votes, say 0.15 < 0.25. This hypothesis is valid when members do not influence each other when voting (the "independence" condition in the theorem). The other mechanism is the so-called "truth wins." This hypothesis emphasizes the role of communication: teams can achieve more efficient outcomes through communication, learning and deliberation (e.g., Laughlin 2015, Friedkin and Bullo 2017).

The experiment results are more consistent with the "truth wins" idea among the two mechanisms. First, teams achieve much higher efficiency than individuals thanks to the former's effective use of the sanctioning institutions. In particular, given an option to construct a formal sanction scheme, individuals vote for inefficient, non-deterrent sanction rates much more than 50% of the time. By sharp contrast, teams vote for *deterrent* sanction rates, i.e., the rates that make free riding materially unprofitable, more than 50% of the time. This pattern is inconsistent with the Condorcet's jury theorem: if team decision-making had meant mere aggregation of individual preferences, team votes would have been more concentrated around small, non-deterrent sanction rates. The observed pattern therefore implies the role of influence among members and deliberation in team decision-making. When informal punishment is collectively enacted, its teams punish low contributors more frequently than individuals, which helps reduce the relative frequency of "misdirected" punishment, i.e., punishment of high contributors. A structural estimation uncovers that the percentage of pro-social punishers is larger among teams than individuals, again unlike the view that emphasizes the effect of team decision-making on aggregating preferences. Moral hazard in groups is a central issue in organizations as it can hurt productivity (e.g., Holmstrom 1982). While recent experiments suggest that it can endogenously be resolved by allowing agents to construct institutions (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018), the finding of the present study underlines the clear role of organizational structure in strengthening a group's ability to govern themselves, whether under formal or informal schemes. This would open up a new research direction in the field concerning the shape of efficient organizations.

The present paper is related to recent experiments studying team joint decentralized punishment behaviour: Auerswald *et al.* (2018) and Kamei (2021). On the one hand, Auerswald *et al.* (2018) let members in a team decide by voting how to punish peer to peer as a team in a finitely repeated public goods game, and found that teams punish less but contributed more than individuals. The present paper is significantly different from their paper in two important aspects. First, team

members in the present experiment *communicate* with each other to decide their team's informal punishment decisions. Such intra-team communication is unavailable in Auerswald *et al.* (2018). This difference in the decision process may change behaviour, because the literature on team decision-making emphasizes the role of communication, learning and deliberation. The experiment data indeed show a contrast in results between theirs and the present one: weaker collective punishment in Auerswald (2018) versus stronger collective punishment in the present paper than individuals. The hypothesis that emphasizes the mere effect of team decision-making in aggregating individual preferences discussed above explains the punishment patterns in Auerswald *et al.* (2018) quite well, but not those in the present experiment. Second, and equally important, the present experiment explores units' choices between the formal and informal sanction schemes, and investigates their behaviours under each of the enacted schemes. Auerswald *et al.* (2018) does not have the components of institutional choices. It is worthwhile investigating teams' institutional choices and their behaviours under endogenously-enacted formal schemes, because theoretical predictions are completely different for the formal and informal schemes, and people's behavioural tendencies in the informal scheme may not be perfectly applicable to the formal scheme (e.g., Falkinger *et al.* 2000; Kosfeld *et al.* 2009; Kamei *et al.* 2015). On the other hand, Kamei (2021) studied how team third-party punishment differs from individual third-party punishment in a one-shot, prisoner's dilemma design. Third-party punishment is driven by purely altruistic motives of uninvolved parties in Kamei (2021), while the present paper studies *involved* parties' decisions whose punitive acts may help raise their own material gains in future interactions.[1]

The rest of the paper proceeds as follows: Section 2.2 summarizes related literature and Section 2.3 describes the experimental design, Section 2.4 discusses hypotheses, and Section 2.5 reports experimental results. Section 2.6 briefly reports results from finite mixture modeling and Section 2.7 reports results of the communication content analysis. Section 2.8 concludes.

## 2.2. Related Literature

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) team decision-making.

First, there is extensive literature on social dilemmas contributed by not only economists but also scholars in neighboring fields (e.g., political science, psychology). One of the most frequently used set-ups in this area is a public goods game (PGG). In a PGG, individuals are allocated to a group

[1] Due to the stark difference in the incentive structure, there are two distinct experimental literatures on decentralized punishment; one for peer-to-peer punishment, and the other for third-party punishment.

of $N$ ($N > 2$), given a fixed endowment, and then decide how much to contribute to their group. Parameters are set such that members have private incentives to free ride, while contributing certain amounts is Pareto efficient. For years, such experimental PGGs 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. A typical contribution pattern is that some individuals initially attempt to cooperate; however, non-cooperation remains rife and features the expected downward trend of cooperation norms (e.g., Ledyard 1995; Chaudhuri 2011).

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 (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), and underlines 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 behavioural 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, intriguingly (b) groups prefer and sustain cooperation with *informal* mechanisms, such as peer-to-peer punishment, instead of relying on formal mechanisms, if doing so leads to a more efficient outcome. For example, Kamei *et al.* (2015) let individuals choose between formal and informal sanction schemes. 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 fixed cost, despite the standard theory prediction, the benefits of consistency, and reduced risk the formal mechanism offers (see also Fehr and Williams (2018)). To the authors' knowledge, all 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 an institutional setting when a group consists of multiple teams, and how teams make joint institutional decisions through communication and voting.

The second, closely related area is a substantial literature on team decision-making. Prior experiments 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 and Tindale 2004). One persistent finding is that teams display greater cognitive ability than individuals in logic or problem-solving activities, as has been seen, for example, in teams' quicker learning of the game-theoretic prediction of 0 in beauty contest games (e.g., Kocher *et al.* 2006 ; Kocher and Sutter 2005; Sutter 2005) or teams' stronger predictions in a replica of monetary policy decision-making (e.g., Blinder and Morgan, 2005). These kinds of sophisticated team behaviours 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 design 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 teams' better-disciplined use of informal punishment (Cinyabuguma *et al.* 2006).

Another 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 tendency may mean that teams form better institutions than individuals since they may be more willing to incur costs to enforce social norms, with the aim of enjoying long-term benefits. Having said that, it is debatable whether teams make better decisions than individuals when risk is involved. For example, "winner's curse" is worse for teams than individuals in an auction setup if teams are composed of individuals with distinct information (e.g., Cox and Hayne 2006; Sutter *et al.* 2009). Results from risk elicitation experiments are mixed (e.g., Baker *et al.* 2008). While Bateman and Munro (2005) found that teams are more risk averse than individuals, Rockenbach *et al.* (2007), and Harrison *et al.* (2013) did not find so. Ambiguous findings in risk elicitation may mean that risk attitudes depend on the environment and the degree of risk (e.g., Shupp and Williams 2008). Regarding rationality, individuals and teams both display similar tendencies to violate expected utility theory (Bone *et al.*, 1999, 2004; Bateman and Munro, 2005; Rockenbach *et al.*, 2007), unlike theoretical implication (e.g., Bone, 1998). Empirical studies suggest similar portfolio decisions and returns for teams and individuals (Prather and Middleton 2002), but more moderate betting behaviour among teams than individuals (Adams and Ferreira, 2010).

It is worth remarking that teams may display more selfish choices than individuals, due to the greater presence or influence of fear and greed in team settings (e.g., Wildschut *et al.* 2003; see also Ahn *et al.* [2001]). For example, 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 institutional design as certain treatments require communication which is crucial for cooperation (for example, see Brosig *et al.* [2003] and Kamei [2019b]). Some studies have also found evidence of teams behaving more cooperatively than individuals in a *repeated* environment when sanctioning institutions are absent. Hence, no clear predictions are possible for discontinuity effects. For example, in Wildschut *et al.* (2003) the individual-team discontinuity effect was minimized most when reciprocal strategies were practiced. Kreps *et al.* (1982) showed theoretically that this may be a rational strategy when units believe that their opponents will play a "tit-for-tat" or non-cooperative strategy. This is empirically supported in a repeated prisoner's dilemma game by Kagel and McGee (2016) and Cooper and Kagel (2022) 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), and Müller and Tan (2013) also report teams' more cooperative behaviour in a repeated common-pool resource problem, weakest-link/average-opinion game, and Stackelberg market game, respectively. Results are relatively mixed in a repeated PGG: while teams contributed significantly more than individuals in Auerswald *et al.* (2018) and Kamei (2019b), teams behaved almost identically to individuals in Cox and Stoddard (2018). Auerswald *et al.* (2018) also showed that teams may contribute more strongly than individuals when teams inflict informal punishment by voting. Their experiment is useful in predicting subjects' behaviours in the present experiment, and hence will be discussed in Section 2.4.

### 2.3. Experiment Design

The experiment is built on a linear PGG. Subjects play the games under one treatment condition (between-subjects design).[2] Six treatments are constructed using a 2×3 design (Table 2.1). The first dimension is the decision-making unit, either an individual or a three-person team. The second dimension is the institutional environment, i.e., there is a regime without sanction schemes, with modest sanction schemes, or with strong sanction schemes. The six treatments are named as "I-No (Individual, No Voting)," "I-Voting-M (Individual, Voting, Modest)," "I-Voting-ST (Individual,

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

Voting, Strong)," "T-No (Team, No Voting)," "T-Voting-M (Team, Voting, Modest)," and "T-Voting-ST (Team, Voting, Strong)."

This experiment uses formal and informal sanction schemes (FS and IS schemes hereafter) as available institutions, so that the results on subjects' decisions are easily comparable against prior related research. Specifically, the sanction scheme is designed based on Kamei *et al.* (2015). Each decision-making unit votes whether to enact an FS or IS scheme in their group, after which a majority rule is applied. A novel part of the design is that unlike all prior experiments on 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 units are *teams*. The treatments with individuals being as the decision-making units will act as a control treatment.

**Table 2.1:** *Summary of Treatments*

| Treatment name | Decision-making unit | Voting | Punishment strength | Number of groups (sessions) | Number of subjects |
|---|---|---|---|---|---|
| I-No | Individuals | No | n.a. | 12 (2) | 36 |
| I-Voting-M | Individuals | Yes | modest | 11 (2) | 33 |
| I-Voting-ST | Individuals | Yes | strong | 11 (2) | 33 |
| T-No | Teams | No | n.a. | 12 (7) | 108 |
| T-Voting-M | Teams | Yes | modest | 11 (6) | 99 |
| T-Voting-ST | Teams | Yes | strong | 11 (6) | 99 |
| Total | | | | 68 (25) | 408 |

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

A partner matching protocol is used in all 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 experiment. The number of periods is set at 24. The periods are grouped into six phases of four periods each (Figure 2.1). The number of periods is common knowledge to the subjects. Subject identity is kept anonymous throughout.

In each period, every decision-making unit is endowed with 20 points (62.5 points = 1 pound sterling), and then simultaneously decide how many points to allocate between their private and public accounts. Contribution amounts must be non-negative integers and not exceed 20. A marginal

per-capita return (MPCR) is set at 0.6. A MPCR of 0.6 is often used like a MPCR of 0.5 in PGG experiments with a group size of three or even four (e.g., Falk *et al.*, 2013; Kocher *et al.*, 2008; Lugovskyy *et al.*, 2017). In other words, when unit *i* contributes $c_{i,t}$ to the public account, she receives the following payoff $\pi_{i,t}$:

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

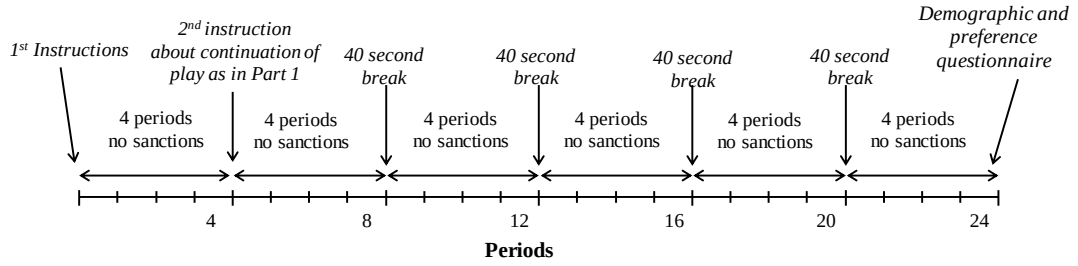
In the three treatments with teams, each member in a team *i* receives the team's payoff (to make the payoff consequence the same for team members in the team treatments and individuals in the individual treatments).[3] At the end of a given period, each unit is informed of (i) their own payoff and (ii)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

The structure of Phase 1 (also called "Part 1") is the same for all six treatments. In this phase, subjects repeat the public goods game without any sanctioning opportunities (No Sanction [NS] scheme, hereafter) four times with the same group membership. Phase 1 is intentionally included to help subjects learn the basic structure of the PGG and let them experience the dynamic free riding problem, typical to a public goods dilemma. Such exogenous periods are sometimes included in experiments on voting for the sake of gradual learning if the setup is complex (e.g., Ertan *et al.*, 2009; Dal Bó *et al.*, 2010; Zhang *et al.*, 2014; Kamei *et al.*, 2015; Fehr and Williams, 2018).

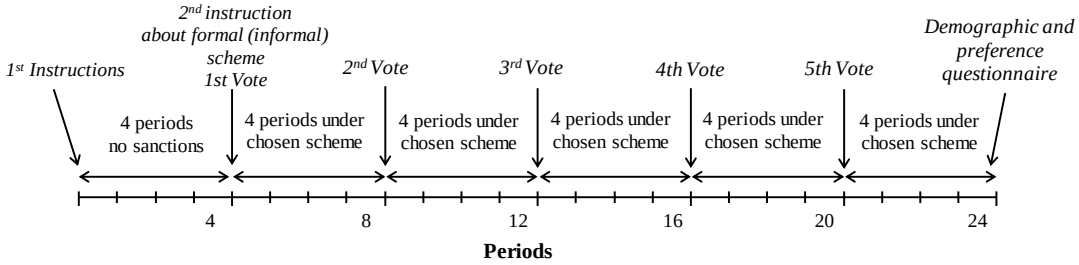
Phases 2 to 6 (collectively "Part 2," hereafter) differ by whether they can use sanction schemes, which is the second dimension of the 2×3 design, as summarized in Sections 3.2 and 3.3. The reason to set five phases (20 rounds) in Part 2 is to allow for the evolution of institutional choice and cooperation behaviour over time. Learning the experimental environments usually takes some time in a complex voting setup like the present one; and prior experiments also had many endogenous rounds to study how institutional choices evolve (e.g., Ertan *et al.*, 2009; Kamei *et al.*, 2015; Fehr and Williams, 2018; Zhang *et al.*, 2014; Gürerk *et al.*, 2006). Panels A and B of Figure 2.1 summarize the schematic diagrams.

[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).

**Figure 2.1:** *Schematic Diagram*




(A) I-No and T-No Treatments




(B) I-Voting-M, I-Voting-ST, T-Voting-M and T-Voting-ST Treatments

### *2.3.2. The Individual Treatments*

In the sanction-free I-No treatment, subjects play the PGG under the NS scheme for all five phases in Part 2 (Figure 2.1.A). There is a 40-second pause between the adjacent phases to control for the restart effects (Andreoni, 1998; Kamei *et al.*, 2015) that may be present in the voting treatments.

In the individual voting treatments (I-Voting-M and I-Voting-ST), each phase of Part 2 begins with the decision-making units voting for the FS or IS scheme – see Figure 2.1.B. Voting is mandatory and does not cost subjects. Whichever scheme receives the majority of votes (i.e., more than or equal to two votes) will be enacted in that group and will be used for all four periods of the phase. Sections 3.2.1 and 3.2.2 summarize the details of the FS and IS schemes, respectively.[4]

#### *2.3.2.1. The IS Scheme*

If a group selects the IS scheme, each period consists of two stages. The first stage is a contribution stage already described in Section 2.3.1. The second stage is an informal punishment stage. In the punishment stage, a decision-making unit *i* can reduce the payoff of each of the other two units (*j*) in their group by assigning punishment points $p_{i \to j} \in \{0, 1, 2, \ldots, 10\}$. While each punishment point given costs the recipient *x* points ($x > 1$), it costs the punisher one point. Thus, the

[4] The FS (IS) 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).

larger the parameter $x$ is, the larger impact a punishment has on the punished. $x = 3$ is used as modest punishment intensity in the I-Voting-M treatment, while $x = 5.5$ is used as strong punishment intensity in the I-Voting-ST treatment.

Following a prior experimental framework (e.g., Fehr and Gächter 2000, 2002), the punishment points allocated by others cannot make the recipients' earnings for that period negative. However, each decision-making unit always incurs the cost that the unit spends in imposing punishments. The payoff for unit $i$ in period $t$ playing IS can therefore be expressed as follows:

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

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

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

#### *2.3.2.2. The FS Scheme*

Group members also face two decision stages in each period when FS is in place. The first stage is a voting stage. In this stage, each unit in a group votes on the rate at which allocations to the *private* account are penalized. Voting is mandatory and cost-free. The available sanction rates (*SR*, hereafter) are 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, and 1.2. After voting, the median of three votes is enacted in the group. Subjects vote four times, once at the onset of each period for that phase (which means that a new sanction rate can be selected in each period) – see Figure 2.1.B.[5]

There are two costs in operating the FS scheme. First, there is a fixed administrative cost when using the FS scheme of 4 points per decision-making unit for that period (Kamei *et al.* 2015). Second, when a member $i$ is fined, the whole group incurs a variable cost of imposing the sanction, i.e., $y$ times the sanctioned amount to $i$. Here, $y$ is a unit cost to impose formal punishment, and the sanction amount is the amount not contributed by $i$ to the public account multiplied by the group sanction rate. Thus, the larger the sanction, the larger the variable cost the group incurs. The cost is equally shared among the three units in the group, meaning that each unit pays $(1/3)\cdot y = y/3$ of the

[5] The FS scheme is not a purely centralized system in which a principal imposes a scheme, but rather a decentralized system in which the agents decide a scheme by voting, whose features share some similarity to the IS scheme.

sanctions.[6] $y$ is adjusted to design such that the cost ratio is the same for the FS and IS schemes (further details below).

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

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

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. Note, however, that they still incur the fixed cost of 4 points per period in this situation.

Equation (3) suggests that for each sanction imposed, the punished must pay $1 + y/3$, while the two other units in the group pay $2y/3$ (= $y/3\times2$) in total as punishers. The cost ratio between them is thus $1 + y/3: 2y/3$. To make the FS scheme fully comparable to the IS scheme, the cost ratio is set the same as the IS scheme, namely, $1 + y/3: 2y/3 = x: 1$. This reduces to the following condition for $x$ and $y$:

$$y = 3/(2x - 1). \quad (4)$$

Following Equation (4), $y = 3/(2\cdot3 - 1) = 3/5$ is adopted for the I-Voting-M treatment ($x = 3$), while $y = 3/(2\cdot5.5 - 1) = 3/10$, is adopted for the I-Voting-ST treatment ($x = 5.5$).

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

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

### *2.3.3. The Team Treatments*

The T-No, T-Voting-M and T-Voting-ST treatments are identical to, respectively, the I-No, I-Voting-M and I-Voting-ST treatments (Figure 2.1), except that the units are *three-person teams*, not individuals. Three subjects playing as a team will jointly make a single decision as a decision-making unit. At the onset of the experiments, subjects are randomly assigned to a team of three, and the team composition does not change throughout the entire experiment. The teams are then randomly assigned to a group of three teams (thus each group consists of nine subjects) before the experiment commences.

The team's joint decision-making follows Kamei (2019b, 2021). Three members in a team communicate with each other for 60 seconds using a computer chat screen before making each team decision. Members are not allowed to communicate verbally, eliminating the risk of contamination of the experiment which may occur if players were able to overhear another team's discussions. The members are only able to communicate with other members of their own team. Anonymity is preserved, such that the subjects are identified by fixed Player IDs in the chat screen, and they are instructed that disclosing any information that may identify themselves or using offensive language is prohibited.[7]

A team's three joint decisions are determined using the median voting rule. This includes the allocation decisions in the PGG (all team treatments), and punishment decisions under the IS scheme and sanction rate votes under the FS scheme (T-Voting-M and T-Voting-ST treatments).[8] The specific procedure is as follows: The three members in a team first discuss strategies and decisions with their team. After the communication stage, each member privately and simultaneously submits their preferred decision (e.g., an amount they wish to contribute as the team's joint contribution decision).[9] The median of the three submissions becomes the team's decision. Each team member is informed of the submissions of their two other team members, anonymously and in a random order.

[7] A subject receives a fine of 10 pounds with an apparent violation of this rule. No one disclosed any identifiable information, and only seven out of 306 subjects (2.28%) had to pay the fine with the rule of offensive language.
[8] To the authors' knowledge, there are several frequently-used methods to resolve a disagreement: (a) the computer randomly selects one choice (e.g., Kamei 2019b, 2021); (b) a default option is applied (e.g., Kagel and McGee 2016); (c) each teammate does not obtain any points (e.g., Feri *et al.* 2010); and (d) a majority rule is applied (e.g., Gillet *et al.* 2009). There is no consensus regarding which method is the best. A median voting rule was adopted in the present study when units vote among more than two options, so that teams' decisions reflect team members' average preferences, considering that individual preferences are always applied in the individual treatments. A majority voting rule was adopted for voting between FS and IS because then each unit's voting affects outcome when they are pivotal.
[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

A team's joint scheme choice (FS or IS) is based on a majority rule. As in the other team decision-making, each team member votes on which scheme they prefer after communication, with the team's majority choice (an option with at least two votes) being the team's joint voting decision.[10]

### 2.4. Hypothesis

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

The standard theory assumption also predicts that having IS does not alter equilibrium play from that in the NS scheme because punishment activities are costly (e.g., Fehr and Gächter 2000, 2002). Notice that, from Equation (2), $\partial \pi_{i,t} / \partial p_{i \to j} = -1 < 0$ for all *i* and *t*. Thus, it is materially beneficial for each unit to not punish one another ($p_{i \to j} = 0$), in which case their payoff would be unaffected when compared to the payoff in the allocation stage (Equation (1)).

In contrast, standard theory prediction (based on pure selfishness and players' correct beliefs) is different in the FS scheme from that in the NS or IS scheme (e.g., Falkinger *et al.* 2000; Kamei *et al.* 2015). Each unit's optimal contribution amount in the second stage of a given period depends on what sanction rate is realized. Notice that $\partial \pi_{i,t} / \partial c_{i,t}$ is calculated from Equation (3) as: $-0.4 + SR_t(1 + y/3)$. This suggests that units contribute nothing when the enacted $SR_t$ is 0.0 or 0.2 as then $\partial \pi_{i,t} / \partial c_{i,t} < 0$,[11] but they contribute the full endowment amount when $SR \geq 0.4$. In other words, $SR \in \{0.4, 0.6, 0.8, 1.0, 1.2\}$ is a deterrent sanction rate, while $SR \in \{0.0, 0.2\}$ is a non-deterrent sanction rate. Each unit obtains a payoff of 32 points (= 20 – 20 + 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. Therefore, each unit has a material incentive to vote for a deterrent sanction rate in the first stage for as long as that they are pivotal. A possibility of errors in others'

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 submitted the same vote in 278 out of 330 cases.

[11] When SR = 0.2, $\partial \pi_{i,t} / \partial c_{i,t} = -0.4 + 0.2(1 + y/3) = -0.2 + 0.2y/3 = -0.16 < 0$ under modest punishment intensity (y = 3/5); $= -0.18 < 0$ under strong punishment intensity (y = 3/10).

voting suggests that they always have a material incentive to do so since the likelihood that they are pivotal is never zero under each contingency (trembling-hand perfect equilibrium). Hence, the standard theory predicts that given an option to vote, everyone votes for the FS scheme rather than IS; and the FS scheme is then enacted in the group as a result of a majority rule applied. Under the FS scheme, every unit votes for a deterrent sanction rate in the first stage; and each unit contributes the full endowment amount to the public account in the second stage.

However, players' optimal decisions and the superiority of the FS over the IS scheme may change if the common knowledge of players' selfish preferences is dropped. For example, Kreps *et al.* (1982) show theoretically that cooperation can be sustained when players believe that there are some non-selfish types, e.g., players that act according to a "tit-for-tat" strategy, in the population, even though there are no such types in reality. In addition, prior experimental research has shown that some humans indeed have other-regarding preferences (see, e.g., Fehr and Schmidt 2006 and Sobel 2005 for a survey); and the overall contribution pattern usually displays one of a conditional contribution type, while conditional contribution preferences are quite heterogeneous among subjects. The prior research has also demonstrated that people can sustain contributions at high levels under certain conditions when the IS scheme is available (e.g., Fehr and Gächter 2000, 2002; Anderson and Putterman, 2006; Nikiforakis and Normann, 2008) due to peer-to-peer punishment inflicted driven by non-selfish preferences. The costly punishment activities and the maintenance of contributions can be rationalized successfully by, for example, the inequity-averse preference model (Fehr and Schmidt, 1999). Thus, people's institutional choices between the FS and IS schemes are not obvious for real human subjects as they can achieve high cooperation regardless of which scheme is selected, and they may prefer IS to FS to avoid a fixed administrative cost. Empirically, institutional choices are known to be affected by which scheme is more materially beneficial, as shown by prior experiments in the literature (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018).

The main aim of this paper is to investigate how teams utilize sanctioning institutions differently from individuals, and as a result how teams make contribution decisions differently from individuals under a given sanction scheme. As described above, a standard theoretical analysis that does not incorporate the internal aspects of team decision-making suggests the same behaviours for teams and individuals, since the theory treats teams the same as individuals as decision-making units.

***Hypothesis 1*** *(theory based on selfish preferences and common knowledge of rationality, without considering the internal aspects of team decision-making): The difference in the decision-making format, team- or individual decision-making, does not have any impact in the experiment.*

However, such analysis misses an important dimension for teams, namely, the preference aggregation process or intra-team dynamics. There are two different approaches that consider the preference aggregation process when three members make a joint team decision. The first one relates to the notion of 'wisdom of the crowds' and is captured by theories of preference aggregation such as the so-called Condorcet jury theorem, or what Ertan *et al.* (2009) call the "behavioural public choice theorem" (also see Hauser *et al.* 2014). Condorcet's (1785) jury theorem suggests that, assuming the event of individuals in a given population choosing the correct answer is independent (unconditional independence), the probability of the majority of individuals voting for a correct answer hinges on the individual correctness probability. That is, if the probability of an individual's answer being correct exceeds ½, and that this is the same for all individuals in the population, then the probability of the majority in a group voting for the correct answer is larger than that in which each individual votes correctly. Similarly, if general competence is below ½, then the probability of a majority voting for an incorrect answer is worse in a group than by individuals and increases with group size. In terms of this experiment, it can be interpreted that the decisions of individuals in the individual treatments would indicate the individual correctness probability. Should individuals be able to select the strategically correct answer more than 50% of the time, then it follows that teams of three formed of similar individuals will have a greater probability of voting for the strategically correct answer, whether a deterrent sanction rate or lack of punishment. This tendency is summarized as Hypothesis 2 below:

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

Hypothesis 2 explains well voting outcomes in prior experiments where each individual voted independently as a decision-making unit (Ertan *et al.*, 2009; Hauser *et al.*, 2014). It also explains Auerswald *et al.* (2018) who studied teams' punishment behaviour when its member decides on a single punishment decision by voting without communication. Individuals in the individual treatment of Auerswald *et al.* inflicted punishment around 34.5% of the time; and consistent with Hypothesis 2, teams in their team treatments punished much less frequently than the individuals. Similar to Ertan *et*

*al.* (2009), their data suggests that especially anti-social punishment is far less among teams than individuals.

As an anonymous referee pointed out, a punisher may incur some psychological costs in addition to monetary costs, since the punishment acts harm their peers.[12] One may reasonably argue that teams incur such psychological costs less than individuals (i.e., team decision-making dilutes the psychological costs) because each member in the former can share with two other team members the responsibility associated with the punishment decisions. At least in the present setup, however, it can be argued that the effects of preference aggregation dominate the effects of diluted responsibility as earlier experiments (Ertan *et al.*, 2009; Auerswald *et al.*, 2018) clearly support Hypotheses 2, despite possibly team decision-making encouraging punishment due to smaller psychological costs.

It should also be worth noting that Hypothesis 2 is valid only when the independence of the probability of individuals being correct holds. This independence assumption is unlikely to hold for the present experiment as the three team members have intra-team communication and deliberation before making each team decision. In this sense, team decision-making may be *more than* the aggregation of three members' preferences, and as such one can expect that Hypothesis 2 will not hold perfectly.

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

In the context of the present experiment, correct answers can be considered the options that lead to the highest utilities, since, as already discussed, voting patterns revealed in prior experiments on institutions showed the strong behavioural effects of material outcomes in driving units' choices of sanctioning institutions (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.*

[12] An analysis of communication logs of the present experiment in fact revealed some subjects' dislike of using punishment (see Kamei and Tabero 2022).

2009; Kamei *et al.* 2015; Fehr and Williams 2018). Thus, the learning model allows for one or more team members to persuade their teammates of the correct logic and so lead them to a better decision, i.e., selecting deterrent sanction rates when the FS scheme is in effect, and punishing low rather than high contributors when the IS scheme is in effect.[13] Recent experiments on problem-solving suggest asymmetry regarding influence and persuasion. For instance, He *et al.* (2022) found that more cognitively able and knowledgeable members can influence less knowledgeable members more strongly if they work together, thereby making it easier for the latter to find correct answers while the former is little affected by incorrect suggestions made by the less able member (see Schulze and Newell (2016) and Bonner *et al.* (2002) for similar findings). Prior experiments on team decision-making also support the role of deliberation and learning as a mechanism of improved decision-making under certain conditions (e.g., Cooper and Kagel, 2022). In sum, while one may still see examples of teams that converge to poor strategies, for example when most members are incorrect, it is expected on average that deliberation and learning will allow teams to discover an optimal strategy more quickly than individuals.

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

## 2.5. Experimental Results

The experiment was conducted at the University of York (see Appendix A.D.1 for the implementation). Section 2.5.1 provides an overview of the decision-making units' average behaviours and examines treatment differences in contributions and payoffs. Section 2.5.2 investigates scheme voting behaviour, while Sections 2.5.3 and 2.5.4 compare the units in utilizing the sanctioning institutions.

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

Groups experienced typical free riding dynamics in the no-voting treatments (Figure 2.2). 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, end-game defection was evident in period 24. The average contribution across all periods was 10.19 points (50.9% of the endowment) in the I-No treatment. Likewise, the average contribution of teams was also modest, 10.57 points (52.9% of

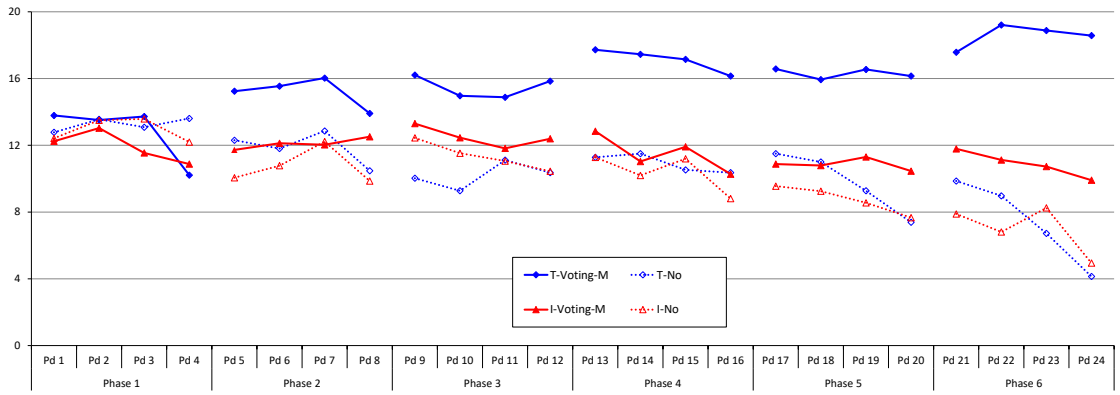
[13] Well-targeted informal punishment plays a vital role in achieving a high payoff (e.g., Hermann *et al.*, 2008).

the endowment), in the T-No treatment, and the contribution dynamic followed a declining trend, similar to that of individuals in the I-No treatment. Similar trends for individuals and teams are unsurprising because of the floor effect, typical to the serious free-riding dynamics in a repeated PGG.

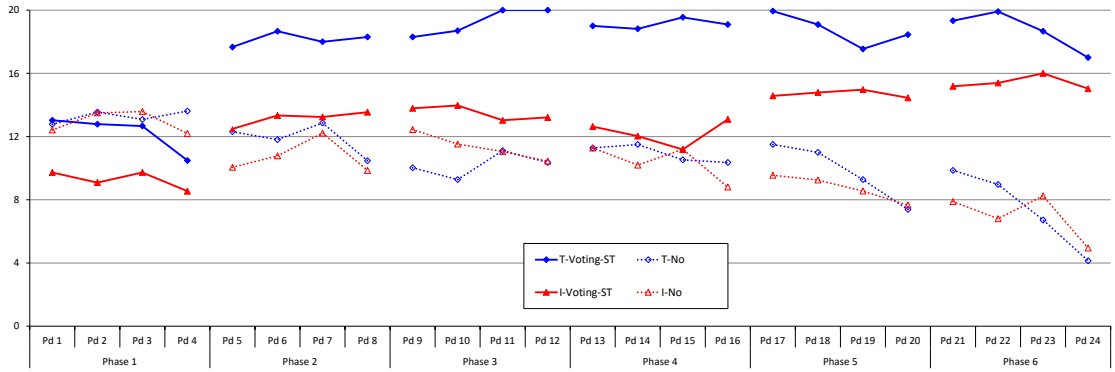
A key comparison in this study is units' decisions to contribute under voting. Contribution trends differ drastically between individuals and teams in the voting conditions, clearly at odds with Hypothesis 1. The difference was especially large under the mild punishment intensity (Figure 2.2.A). Teams in the T-Voting-M treatment learned to cooperate gradually from phase to phase. Remarkably their average contributions were more than 80% of the endowment in the final three phases. By comparison, individuals in the I-Voting-M treatment did not follow as strong a learning pattern, although they did not learn to free ride either. The individuals' average contributions hovered between 10 and 12 points. The clear difference between the T-Voting-M and I-Voting-M treatments is consistent with the discontinuity-effect hypothesis. When the punishment intensity was strong, cooperation evolved at a further higher level among teams – see Figure 2.2.B, i.e., close to the full contribution level in each Part 2 phase. With strong punishment, individuals (in the I-Voting-ST treatment) were able to gradually learn to cooperate. However, the difference in the average contribution was consistently large between individuals and teams.

Figure 2.3 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 cooperate, shown in Figure 2.2.A, triggers negative emotional responses from their peers (e.g., Casari and Luini 2009; Gächter *et al.* 2008). As a result, individuals received lower payoffs in the I-Voting-M than in the I-No treatment in all phases except Phase 6 (Figure 2.3.A).

**Figure 2.2:** *Average Contribution Period by Period*



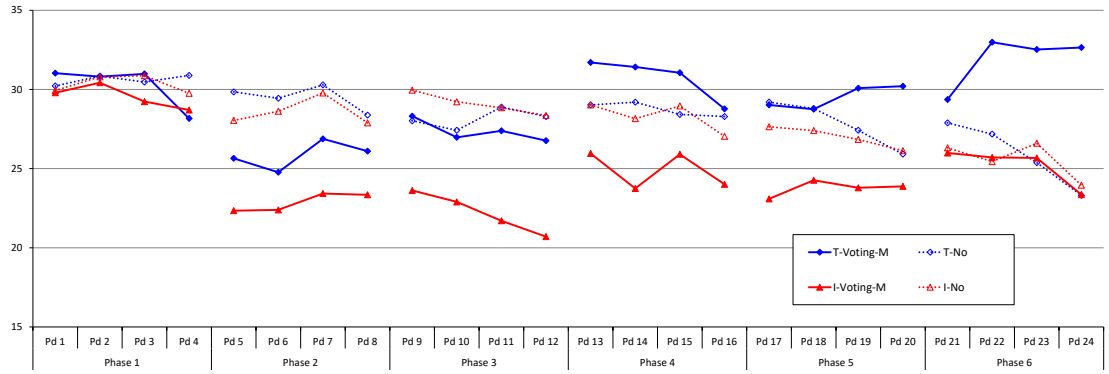

(A) Treatments with Modest Punishment Intensity



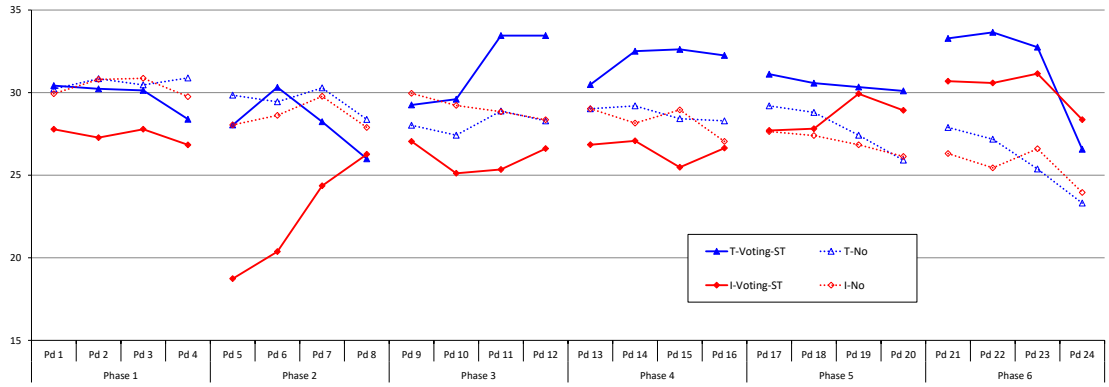

(B) Treatments with Strong Punishment Intensity

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

**Figure 2.3:** *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.

Second, teams also experienced such negative welfare losses in the T-Voting-M treatment (Figure 2.3.A). However, the negative impact 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 (Figure 2.2.A), this implies that, in later phases, teams did not need to discipline their group members through costly punishment.

Third, likewise, when the punishment intensity was strong, having the sanctioning schemes led to similar negative welfare consequences in groups. However, the duration in which groups suffered from losses was shorter relative to the treatments with modest punishment (Figure 2.3.B). In other words, the availability of strong punishment induced the members to learn to cooperate quickly, thereby helping reduce the welfare loss due to punishment activities.

A series of non-parametric tests were performed to judge treatment differences statistically (Table 2.2), which confirms most of the patterns seen in Figures 2.2 and 2.3. 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-M and T-Voting-ST treatments, teams' contribution behaviours 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 2.2).[14] Third, individuals earned significantly less in Part 2 than in Part 1 of the experiment in the I-Voting-M treatment, but not in the I-Voting-ST treatment.[15]

A regression analysis was additionally conducted as a robustness check by controlling for the panel structure, as the treatment differences are important. Two estimations were performed by changing the dataset (Table 2.3): one using all observations in Part 2, and the other using the second half of the experiment, as it took some time for units to stabilize behaviours. Two patterns are worth mentioning. First, the contribution behaviour of teams is strong for each punishment strength. Second, the individual-team discontinuity effect is significant for each punishment strength. These two patterns are significant, regardless of which dataset is used, all data or only the second half of the experiment.

[14] The same positive effect can be found even if the two team treatments are not pooled (Panel C of Appendix A.A).

[15] A regression was also performed as a supplementary analysis to analyze the contribution trend in Part 2 (Appendix A.B, Table B.1). It confirms that when the sanctioning schemes were unavailable, units, whether individuals or teams, decreased contributions significantly over time. By contrast, teams increased contributions significantly from phase to phase in both the T-Voting-M and T-Voting-ST treatments. A regression also confirms that the contribution trend differs by punishment intensity when the units are individuals: an increasing (somewhat decreasing) contribution trend in the I-Voting-ST (I-Voting-M) treatment. It further shows that the payoff trend is similar to the contribution trend: declining trends for the I-No and T-No treatments *versus* an increasing trend in the T-Voting-M treatment (the maintenance of high payoff in the T-Voting-ST treatment) – see Appendix A.B, Table B.2.

**Table 2.2:** *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-M, I-Voting-ST) | 10.60 | 12.68 | 0.2914 | 10.24 | 0.8313 | 15.04 | 0.2790 | 0.2330 |
| (b1) I-Voting-M | 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-M, T-Voting-ST) | 12.53 | 17.67 | 0.0001*** | 18.02 | 0.0002*** | 17.30 | 0.0166** | 0.1054 |
| (d1) T-Voting-M | 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-M, I-Voting-ST) | 28.48 | 25.27 | 0.0575* | 23.38 | 0.0086*** | 27.09 | 0.0304** | 0.1252 |
| (b1) I-Voting-M | 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-M, T-Voting-ST) | 30.02 | 29.90 | 0.9353 | 29.71 | 0.8092 | 30.10 | 0.1701 | 0.1252 |
| (d1)T-Voting-M | 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. Wilcoxon signed rank (Mann-Whitney) tests were conducted for within(across)-treatments comparisons, using group means of contributions and payoffs. For example, 12 matched pairs of group means were used to calculate *p* (= 0.0414) to compare average contributions between Phase 1 and Phases 2-6 in row (a) of Panel I. See Panel A of Appendix A.A for the standard errors. See Panel C of Appendix A.A for more detailed across-treatment comparisons. "Indiv Voting" includes the I-Voting-M and I-Voting-ST treatments. "Team Voting" includes the T-Voting-M and T-Voting-ST treatments. [#1] Only groups that had experienced both the FS and IS schemes in Part 2 were used. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table 2.3:** *Treatment Differences in Contribution and Payoff*

| Dependent variable: | Avg. contribution of group *i* in period *t* | | | | Avg. payoff of group *i* in period *t* | | | |
|---|---|---|---|---|---|---|---|---|
| Period: | All periods | | 2nd half of the experiment | | All periods | | 2nd half of the experiment | |
| | ($5 \le t \le 24$) | | ($13 \le t \le 24$) | | ($5 \le t \le 24$) | | ($13 \le t \le 24$) | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| (a) I-Voting-M | 1.93* | 1.51 | 2.39** | 1.85 | -3.92*** | -4.16*** | -2.35*** | -2.90** |
| dummy | (1.11) | (1.60) | (1.19) | (1.72) | (0.54) | (0.80) | (0.71) | (1.22) |
| (b) I-Voting-ST | 4.16*** | 2.11 | 5.41*** | 2.65 | -0.96*** | -2.68** | 1.48*** | -1.06 |
| dummy | (0.63) | (1.65) | (0.55) | (1.89) | (0.34) | (1.26) | (0.25) | (1.48) |
| (c) T-No dummy | 0.40 | -0.82 | 0.68 | -0.45 | 0.32 | -0.79 | 0.54 | -0.47 |
| | (1.12) | (1.25) | (1.17) | (1.50) | (0.90) | (1.01) | (0.93) | (1.22) |
| (d) T-Voting-M | 6.89*** | 5.93*** | 8.63*** | 7.63*** | 1.36 | 0.17 | 3.75*** | 2.60** |
| dummy | (1.19) | (1.06) | (0.99) | (1.00) | (1.69) | (1.40) | (1.24) | (1.16) |
| (e) T-Voting-ST | 9.16*** | 7.30*** | 10.17*** | 8.08*** | 3.02*** | 1.04 | 4.40*** | 2.13 |
| dummy | (0.52) | (0.73) | (0.42) | (0.75) | (1.05) | (1.11) | (1.23) | (1.41) |
| Vote number {= Phase | -0.18 | -0.14 | -0.32 | -0.25 | 0.44 | 0.45 | -0.11 | -0.03 |
| – 1 = 1, 2, 3, 4, 5} | (0.18) | (0.19) | (0.36) | (0.40) | (0.27) | (0.29) | (0.35) | (0.38) |
| Periods within phase | -0.32*** | -0.29*** | -0.50*** | -0.47*** | -0.18 | -0.17 | -0.41*** | -0.40*** |
| {= 1, 2, 3, 4} | (0.08) | (0.09) | (0.11) | (0.12) | (0.12) | (0.14) | (0.13) | (0.14) |
| Constant | 10.98*** | 7.93*** | 11.21*** | 9.87*** | 26.83*** | 23.80*** | 28.41*** | 27.31*** |
| | (0.85) | (2.68) | (1.65) | (3.12) | (1.08) | (2.97) | (1.56) | (2.96) |
| Control[#1] | No | Yes | No | Yes | No | Yes | No | Yes |
| # of observations | 1,360 | 1,220 | 816 | 732 | 1,360 | 1,220 | 816 | 732 |
| # of groups | 68 | 61 | 68 | 61 | 68 | 61 | 68 | 61 |
| Wald $\chi^2$ | 469.28 | 1169.75 | 750.57 | 1474.19 | 200.16 | 340.15 | 96.92 | 132.43 |
| Prob > Wald $\chi^2$ | 0.000*** | 0.000*** | 0.000*** | 0.000*** | 0.000*** | 0.000*** | 0.000*** | 0.000*** |
| Two-sided *p*-value from Wald test | | | | | | | | |
| [discontinuity effects under voting:] | | | | | | | | |
| $H_0$: (a) = (d) | 0.001*** | 0.020** | 0.000*** | 0.003*** | 0.002*** | 0.009*** | 0.000*** | 0.001*** |
| $H_0$: (b) = (e) | 0.050** | 0.000*** | 0.000*** | 0.001*** | 0.000*** | 0.004*** | 0.015** | 0.027** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [impact of punishment strength:] | | | | | | | | |
| $H_0$: (a) = (b) | 0.049** | 0.762 | 0.014** | 0.708 | 0.000*** | 0.187 | 0.000*** | 0.205 |
| $H_0$: (d) = (e) | 0.050** | 0.237 | 0.116 | 0.672 | 0.390 | 0.637 | 0.707 | 0.788 |

*Notes*: Group-average observations were used. Linear regressions with robust standard errors clustered by session ID. Group random effects were also included to control for the panel structure. The numbers in parentheses are the standard errors. [#1] Control variables include group-average period 1 contribution amounts, the percentage of female subjects in the group, and the percentage of students with economics major in the group. In the even-numbered columns, only groups in which all members answered the three demographic questions were used as data. The coefficient estimates of the controls were omitted to conserve space as these are not related to the research questions of the paper. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Lastly, a closer look at the data by sanction scheme uncovers further patterns. First, in the I-Voting-M and I-Voting-ST treatments, cooperation did not evolve when FS was in place, although individuals maintained strong cooperation norms when IS was instead in effect (Appendix A.B, Figure B.1). Hence, the individuals' overall low level of contribution seen in Figure 2.2 is partly attributable to their selection of sanction rates and/or contribution behaviours under the FS scheme. Second, due to the low cooperation norms and administrative cost payments, the individuals persistently earned much less under the FS scheme, relative to the I-No treatment (Appendix A.B, Figure B.2), and the difference is significant (Table 2.2). Third, under the IS scheme, individuals in the I-Voting-M (I-Voting-ST) treatment received lower payoffs than those in the I-No treatment in Phases 2 to 5 (Phase 2 to 3), due to losses from intensive punishment activities. This implies that learning to cooperate with informal punishment requires a sufficient repetition of interactions, as Gächter *et al.* (2008) demonstrated.

The picture is markedly different in the team treatments. Whether in the FS or IS scheme, cooperation was sustained at significantly high levels, relative to the T-No treatment (Appendix A.B, Figure B.1, Table 2.2). Teams also quickly responded to the informal punishment received from their peers. Although payoff losses due to punishment were large in Phases 2 and 3 (in Phase 2) with the IS scheme in the T-Voting-M (T-Voting-ST) treatment, they achieved high payoffs after these phases. Despite administrative cost payments, teams in the T-Voting-ST treatment did earn more than those in the T-No treatment across *all* phases (Appendix A.B, Figure B.2).

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

*cooperation more strongly and quickly than individuals regardless of which sanction scheme was chosen.*

### *2.5.2. Scheme Choice*

The strong efficiency under the IS scheme unlike the prediction based on pure selfishness was not driven by a small number of groups. Despite the 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-M, T-Voting-M, I-Voting-ST, and T-Voting-ST treatments, respectively (Table 2.4.I). As a result of majority voting, groups adopted the IS scheme similar percentages of the time, i.e., 47.3%, 58.2%, 54.6%, and 40.0% of the time in the corresponding treatments (Table 2.4.II). These percentages are all significantly different from 5% (5% is a probability that is usually assumed for an error to happen), which means that units' voting for the IS scheme and the vote outcomes were driven by systematic motives – see Table 2.4 again. Group-level Mann-Whitney tests also indicate that scheme choice behaviours did not differ between individuals and teams (Panel K of Appendix A.A).

Realized relative effectiveness of FS and IS schemes affected voting. Seven, nine, eight, and six groups experienced both the FS and IS schemes at least once in the I-Voting-M, T-Voting-M, I-Voting-ST and T-Voting-ST treatments, respectively. Using these groups, Figure 2.4 demonstrates that units were more likely to vote for the scheme under which they had previously experienced higher payoffs, while there is a large variation for units' voting, perhaps driven by strong heterogeneity in subjects' cooperation and punishment tendencies (e.g., Fischbacher *et al.* 2001; Fischbacher and Gächter 2010; Kamei 2014). This resonates with the idea that people's choices are guided by material outcomes (e.g., Ertan *et al.* 2009; Kamei *et al.* 2015),[16] and it may be a general phenomenon as the role of realized payoffs has been demonstrated in another setup, e.g., voting on leadership (e.g., Güth *et al.* 2007).

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-M 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,

[16] To supplement this finding, a regression analysis was conducted regarding how units' voting in Phase 6 (the final phase) may be influenced by relative payoff ratios they experienced before that phase. As shown in Appendix A.B, Table B.3, 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).

respectively.[17] The average contributions of groups that always selected FS were 15.16, 19.54, 18.29, and 19.67 points in the I-Voting-M, I-Voting-ST, T-Voting-M, and T-Voting-ST treatments, respectively.

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

**Table 2.4:** *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-M | 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-M | 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*** |

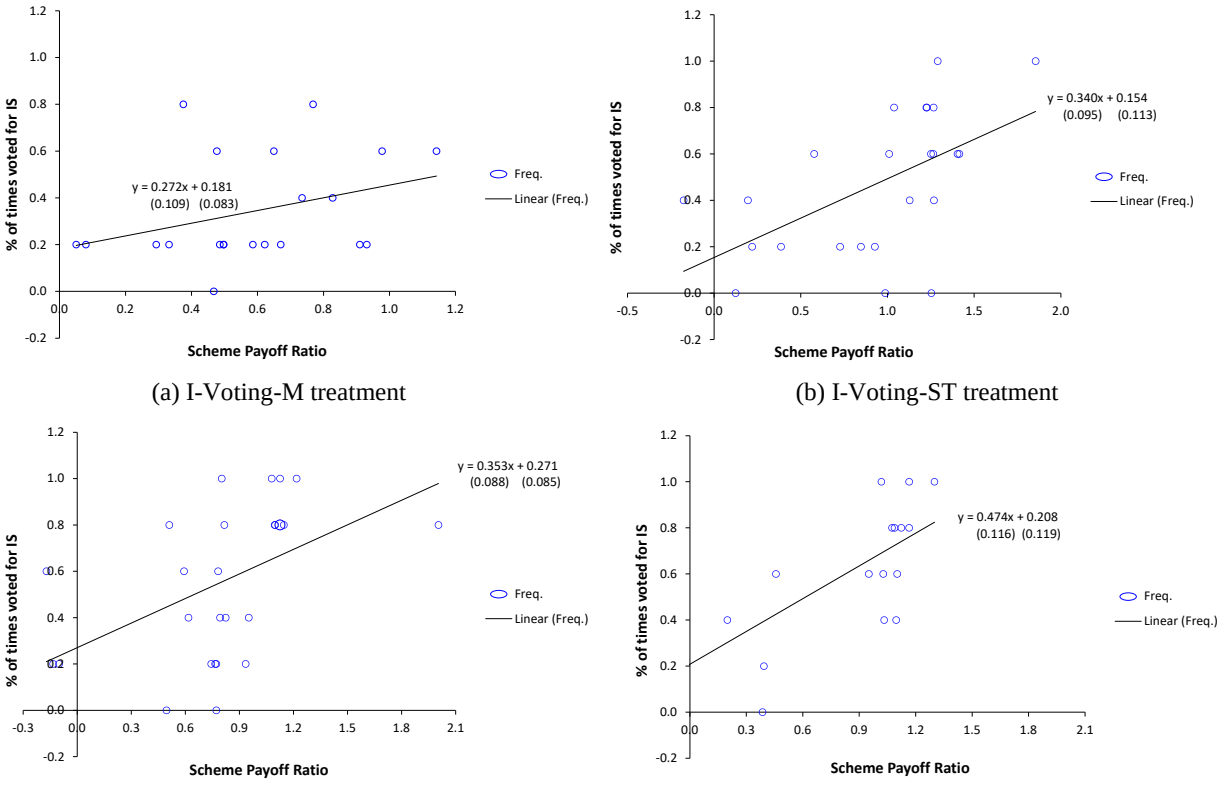
II. Percentages of Times that the IS Scheme was Selected in Groups

| | Phase 2 | Phase 3 | Phase 4 | Phase 5 | Phase 6 | Overall | *p*-value for Wilcoxon signed rank tests [#1] |
|---|---|---|---|---|---|---|---|
| I-Voting-M | 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-M | 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, Wilcoxon signed rank tests were performed using the group-average observations. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, 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-M, I-Voting-ST, T-Voting-M, and T-Voting-ST treatments, respectively. The average contribution of the group that exclusively selected IS in the I-Voting-M treatment was 11.2 points.

**Figure 2.4:** *Scheme Choice and Relative Payoff Ratio*




(a) I-Voting-M treatment (b) I-Voting-ST treatment

(c) T-Voting-M 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 (Seven, nine, eight, and six groups in the I-Voting-M, T-Voting-M, I-Voting-ST and T-Voting-ST treatments, respectively). The horizontal axis (x-axis) is calculated by a given unit's average payoff under the IS scheme divided by their average payoff under the FS scheme across all periods. The vertical axis (y-axis) is the percentage of times the unit voted for IS and takes a value between 0 and 1. The size of each point indicates its frequency. The numbers in parentheses in the linear equation (OLS) in each panel are robust standard errors clustered by group ID. The slopes in the linear lines in panels a, b, c, and d are significantly positive at two-sided $p$ = 0.046, 0.009, 0.004, and 0.009, respectively.

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

Which hypothesis, Hypothesis 2 or Hypothesis 3, drove Result 1? Section 2.5.3 attempts to answer this question by investigating units' use of the sanction scheme in the detail.

#### *2.5.3.1. Voting and Contribution Behaviours in the FS Scheme*

Units' decisions to contribute under the FS scheme were strongly influenced by their group's sanction rate. A regression analysis finds that units were significantly more likely to contribute large amounts, the higher the sanction rate their group had implemented (Appendix A.B, Table B.4). Having a deterrent sanction rate effectively improves units' decisions to contribute (Appendix A.B,

Table B.4). The larger impact of having stronger punishment is consistent with prior research on formal sanctioning institutions (e.g., Falkinger *et al.* 2000; Kamei *et al.* 2015), which suggests that a centralized solution of the free riding problem is to enforce an incentive mechanism in a society or organization.

Result 1 was driven by the difference in voting. As shown in Figure 2.5, the popularity of sanction rates differs markedly between individuals and teams. First, the sanction rate of 0.0 was the focal point among the individuals. Strikingly, individuals in the I-Voting-M and I-Voting-ST treatments voted for the zero sanction rate on 63.79% and 54.00% of the occasions, respectively (Figure 2.5.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 2.5.B).[18]

Given this revealed voting patterns of individuals, Hypothesis 2 predicts that if team decision-making were mere aggregation of three members' preferences, then teams would vote for non-deterrent sanction rates, and therefore collectively enact non-deterrent formal schemes in their groups, more frequently than individuals. However, clearly contrary to this hypothesis, teams voted for the zero sanction rate only 34.06% and 28.03% of the time in the T-Voting-M and T-Voting-ST treatments, respectively. Instead, consistent with Hypothesis 3, teams used voting much more efficiently than individuals: teams voted for deterrent sanction rates (0.4 or above) on 56.88% and 66.67% of the occasions in the T-Voting-M and T-Voting-ST treatments, respectively. In particular, teams' preferences for the highest sanction rate – 1.2 per point allocated to the private account – were strikingly strong (Figure 2.5.A). In the T-Voting-ST treatment, teams voted for the highest rate on 53.54% of the occasions. With the majority rule, 31.52% (26.09%) and 62.12% (14.39%) of the vote outcomes were the highest (zero) sanction rate in the T-Voting-M and T-Voting-ST treatments, respectively.[19] The average realized group sanction rates were 0.64 and 0.89, both of which are deterrent, in the T-Voting-M and T-Voting-ST treatments, respectively, while these were much smaller

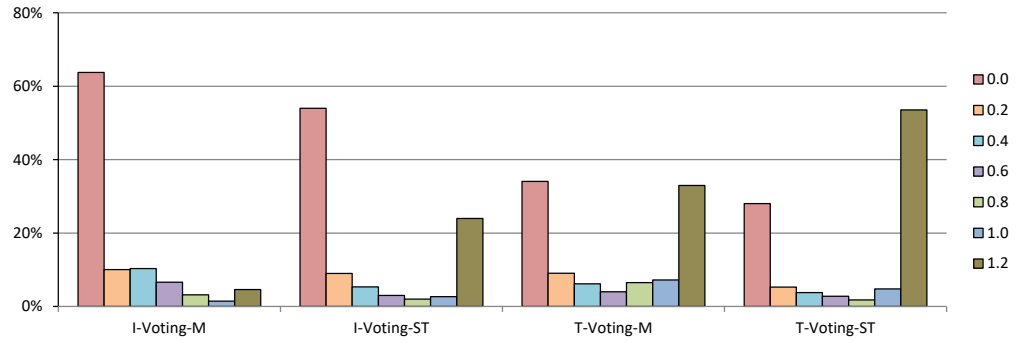
[18] The outcome of the zero sanction rate is somewhat larger than the percentage of the voters who preferred it (e.g., 70.69% > 63.79%). As already discussed in Section 2.4, this is due to the majority voting system because it tends to outnumber the preferences of minorities – a phenomenon driven by the Condorcet's (1785) jury theorem or the behavioural public choice theorem (e.g., Ertan *et al.* 2009; Hauser *et al.* 2014).

[19] The percentages of cases in which a group selected the zero (highest) sanction rate in Phases 2 to 6 are significantly different between individual and team voting at two-sided $p = 0.0080$ ($p = 0.0319$), according to a group-level Mann-Whitney test, when pooled data are used – see Panel F of Appendix A.A.

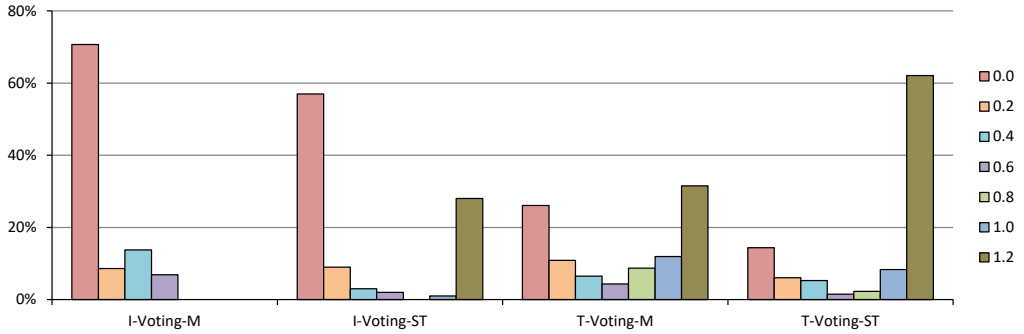
in the individual treatments, i.e., 0.11 and 0.39 in the I-Voting-M and I-Voting-ST treatments, respectively.[20] The difference in the selected sanction rates well explains the stronger contribution behaviours of teams in the voting treatments (Figure 2.2, Appendix A.B, Figure B.1).[21]

Result 1 was partly affected by units' reaction to group sanction rates. Strikingly, on average, teams contributed significantly more than individuals, whether sanctions were deterrent or not (columns

**Figure 2.5:** *Voting on Sanction Rates and Vote Outcome*




(A) Distributions of Decision-Making Units' Voting




(B) Distributions of Vote Outcomes

[20] The average realized sanction rates are significantly different at two-sided $p = 0.0116$ between individual versus team voting when pooled data are used (see Panel F of Appendix A.A).
[21] 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.

**Table 2.5:** *Average Contribution by Sanction Rate under the FS scheme*

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

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

a.i, b.i, and c.i of Table 2.5). As the maintenance of cooperation norms leads to large long-term payoffs, the teams' stronger responses to sanction rates suggest that, with the FS being enacted, teams may be more far-sighted and less myopic loss averse than individuals (Sutter 2007, 2009; Bougheas *et al.* 2013), and these are consistent with the positive effects of deliberation and learning that the truth win mechanism proposes.

**Result 3**: *(a) While individuals voted for the zero sanction rate more than 50% of the time, teams did so much less than 50% of the time, inconsistent with Hypothesis 2. Instead, teams voted for deterrent sanction rates more than 50% of the time, consistent with Hypothesis 3. (b) As a result of the majority rule applied, teams on average enacted deterrent sanction rates, but individuals failed to do so. (c) Teams contributed significantly more than individuals for given sanction rates.*

#### *2.5.3.2. Contribution and Punishment Behaviours in the IS Scheme*

Decision-making units inflicted costly punishment based on the distribution of contributions in their group (Table 2.6). First, the smaller the amount a unit *j* contributed to the public account relative to *i*, the more strongly *i* punished *j*. Second, contributing more than another member also attracted punishment by that member to some degree, but such anti-social punishment is significantly weaker than pro-social punishment. These two patterns, which hold for all treatments, are in line with the prior research (e.g., Fehr and Gächter 2000; Kamei and Putterman 2015).

Figure 2.6 reports the relative strength and frequency of anti-social (perverse) punishment to pro-social (non-perverse) punishment. It reveals that (a) pro-social (non-perverse) punishment was more prevalent than anti-social (perverse) punishment among individuals, and that (b) pro-social

(non-perverse) punishment was even more dominant among teams than individuals.[22] The teams' better targeted punishment behaviours are consistent with both Hypothesis 2 and Hypothesis 3. A simulation exercise in the next subsection (Section 2.5.4) reveals that Hypothesis 3 is more reasonable to explain informal punishing behaviours in the IS scheme.

**Table 2.6:** *Determinants of Punishment Decisions under the IS scheme*

Dependent variable: punishment point assigned from decision-making unit *i* to *j* in period *t*

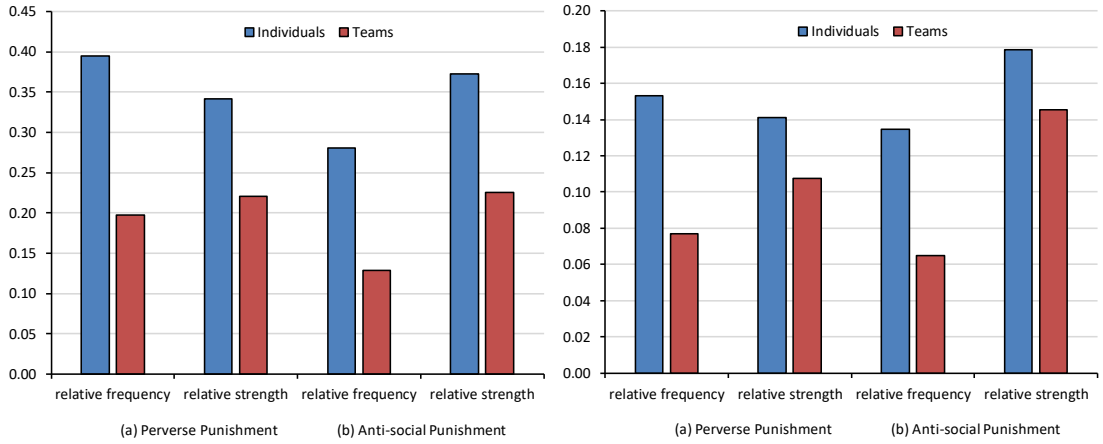
| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (a) Positive deviation in period *t* {= | 0.30*** | 0.30*** | 0.27*** | 0.28*** |
| $\max\{(c_{j,t} - c_{i,t}), 0\}$} | (0.04) | (0.04) | (0.72) | (0.07) |
| (b) Absolute negative deviation in period *t* | 0.63*** | 0.65*** | 0.63*** | 0.68*** |
| {= $\max\{(c_{i,t} - c_{j,t}), 0\}$} | (0.04) | (0.05) | (0.07) | (0.08) |
| (c) Average contribution in their group in | -0.33*** | -0.32*** | -0.24*** | -0.24*** |
| period *t* | (0.03) | (0.04) | (0.04) | (0.05) |
| Interaction: (a) × strong punishment | --- | --- | 0.12 | 0.12 |
| dummy | | | (0.10) | (0.10) |
| Interaction: (b) × strong punishment | --- | --- | 0.16 | 0.09 |
| dummy | | | (0.11) | (0.11) |
| Interaction: (c) × strong punishment | --- | --- | -0.19*** | -0.26*** |
| dummy | | | (0.05) | (0.05) |
| Interaction: (a) × team treatment dummy | --- | --- | -0.03 | -0.06 |
| | | | (0.09) | (0.09) |
| Interaction: (b) × team treatment dummy | --- | --- | -0.09 | -0.13 |
| | | | (0.09) | (0.09) |
| Interaction: (c) × team treatment dummy | --- | --- | -0.08** | -0.08** |
| | | | (0.03) | (0.04) |
| Interaction: (a) × strong punishment | --- | --- | 0.24 | 0.16 |
| dummy × team treatment dummy | | | (0.27) | (0.26) |
| Interaction: (b) × strong punishment | --- | --- | -0.07 | -0.08 |
| dummy × team treatment dummy | | | (0.23) | (0.23) |
| Interaction: (c) × strong punishment | --- | --- | 0.17*** | 0.26*** |
| dummy × team treatment dummy | | | (0.05) | (0.05) |
| Period within Phases {=1, 2, 3, 4} | -0.23* | -0.18 | -0.23* | -0.18 |
| | (0.13) | (0.13) | (0.13) | (0.13) |
| Constant | -0.48 | -1.36 | -0.53 | -1.51* |
| | (0.68) | (0.86) | (0.69) | (0.87) |

[22] Due to the small sample size, the difference is only significant at $p = 0.0544$ for the relative frequency if a one-sided Mann-Whitney test is used based on group averages of all treatments. However, a finite mixture modeling analysis in Section 2.6 reveals significantly different punishment strategies between individuals and teams.

| Control[#1] | No | Yes | No | Yes |
|---|---|---|---|---|
| # of observations | 2,640 | 2,528 | 2,640 | 2,528 |
| # of left-censored observations | 2,216 | 2,126 | 2,216 | 2,126 |
| # of right-censored observations | 30 | 27 | 30 | 27 |
| Log likelihood | -1708.67 | -1610.99 | -1693.07 | -1589.56 |
| Wald $\chi^2$ | 336.48 | 317.21 | 347.58 | 334.00 |
| Prob > Wald $\chi^2$ | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** |
| Two-sided *p*-value for Wald test for $H_0$: (a) = (b) | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** |

*Notes:* Decision-making unit random effects tobit regressions. The numbers in parentheses are standard errors. Observations in periods 5 to 24 are used. [#1] Control variables include unit-average period 1 contribution amounts, the percentage of female subjects in the unit, and the percentage of students with economics major in the unit. In the even-numbered columns, only individuals (teams) in which the individual (all three team members) answered the three demographic questions were used as data. The coefficient estimates of the controls were omitted to conserve space.*, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Figure 2.6:** *Relative Strength and Frequency of Perverse/Anti-Social Punishment*




(A) I-Voting-M versus T-Voting-M 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.

### *2.5.4.Effects of Preference Aggregation (A Simulation Exercise)*

Units' voting on sanction rates was consistent with Hypothesis 3 (Section 2.5.3.1), but units' informal punishment behaviour was consistent with both Hypotheses 2 and 3 (Section 2.5.3.2). One may ask how teams would have utilized the sanction schemes should they have made their team

punishment decisions *without communication*. Answering this specific question helps sort out which hypothesis better explains the informal punishment behaviour. This question was studied by Auerswald *et al.* (2018) and can also be answered by applying the evidence from prior experiments, such as Ertan *et al.* (2009). The prior research findings consistently suggest that, without communication, team decision-making makes punishment weaker, as Hypothesis 2 suggests. While one may argue that punishing as a team may raise their punitive inclinations by reducing psychological costs to harm others through diluted responsibility, these earlier studies did not indicate such effects. Thus, without intra-team communication, arguably, team decision-making can be treated as aggregation of individual preferences.

In order to examine the role of preference aggregation in team decision-making, a simulation exercise was conducted by utilizing the data of individual punishment decisions collected in the individual treatments. The observations in the two treatments (I-Voting-M, I-Voting-ST) are pooled in this exercise to obtain a general pattern with a large dataset, as the behavioural patterns are largely similar for the two treatments (Section 2.5.3). Two kinds of simulations were performed: one for voting on sanction rates under FS, and the other for informal punishment decisions under IS. The computer simulation randomly constructed a team of three from individuals in the individual treatments with replacement 600 times under the FS scheme (1,000 times under the IS scheme); and each "hypothetical" team's joint decision was then calculated using the median of three individual votes (informal punishment points given), following the preference aggregation method used in the team treatments of this study.

Figure 2.7 reports the simulation results. It shows similar tendencies seen in Auerswald *et al.* (2018) and Ertan *et al.* (2009) in that preference aggregation makes the individuals' tendencies more extreme. First, as shown in Panel A, the hypothetical teams vote for the zero sanction rate even more than individuals in the FS scheme. As a result, the average sanction rate enacted by the group consisting of three hypothetical teams is only 0.09. This effect of preference aggregation in the simulation is sharply contrasted with the real teams' selection of stronger sanction rates in the two team treatments (see the white bars). Second, and similarly, the hypothetical teams informally punish their peers much less than individuals, whether pro-socially or anti-socially. However, real teams in the team treatments pro-socially punished their peers more strongly than individuals (Panel B.i). The pattern on anti-social punishment is also intriguing. Anti-social punishment by hypothetical teams is almost non-existent (Panel B.ii). This simulation result resembles the patterns seen in the Ertan *et al.* (2009) and Auerswald *et al.* (2018). On the other hand, real teams' anti-social punishment was similar

to individuals' (see again Panel B.ii). This seems to suggest that those with anti-social inclinations to punish may have strong influence during the intra-team communication to decide team punishment.

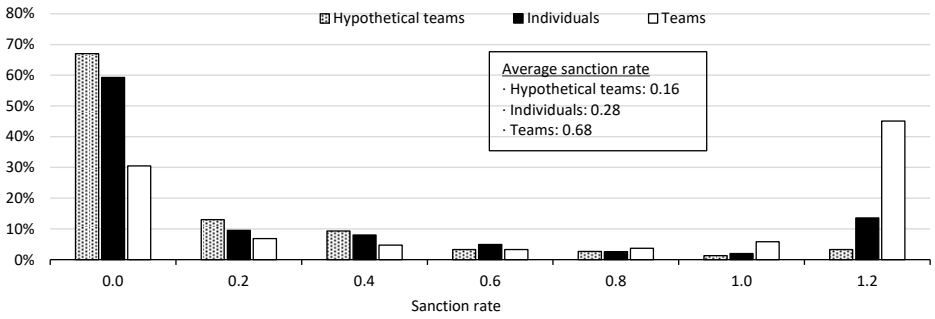
In sum, the simulation exercise confirms that the behavioural pattern seen in the present experiment is more consistent with Hypothesis 3.

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

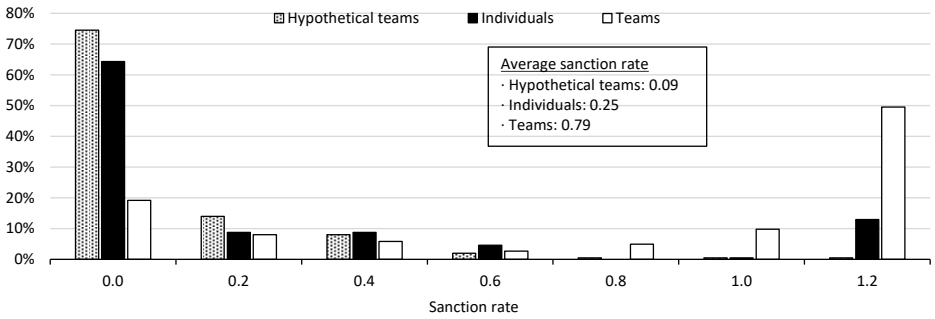
The main experimental finding of the previous section was that (a) teams are able to sustain cooperation at a higher level than individuals when they can vote on sanctioning institutions, and (b) the teams' high efficiency is driven by their effective use of punishment. The detailed analyses in Section 2.5 indicated that the punishment patterns are consistent with Hypothesis 3, rather than Hypothesis 2. To further analyze the discontinuity effect, finite mixture modeling was used to structurally estimate what percentages of individuals and teams punished pro-socially or anti-socially in the IS scheme.

Finite mixture modeling assumes a set of possible behavioural types in advance and then assigns a probability measure over the types to each subject so that the likelihood is maximized (McLachlan and

**Figure 2.7:** *Distribution of Hypothetical Teams' Punishment Decisions (Simulation Results)*

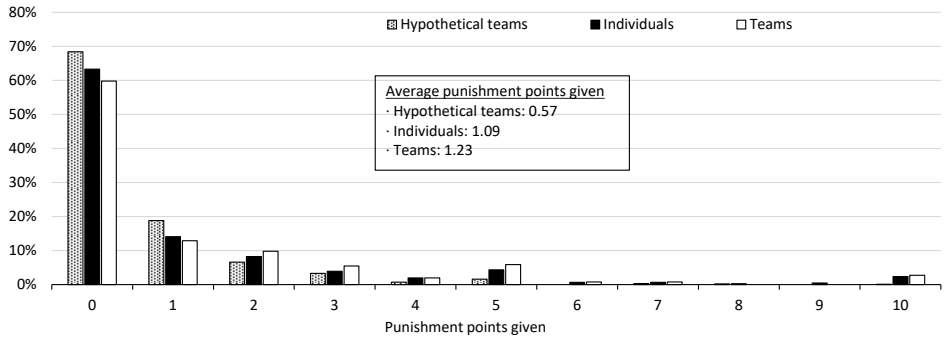



(i) Distributions of Decision-Making Units' Voting

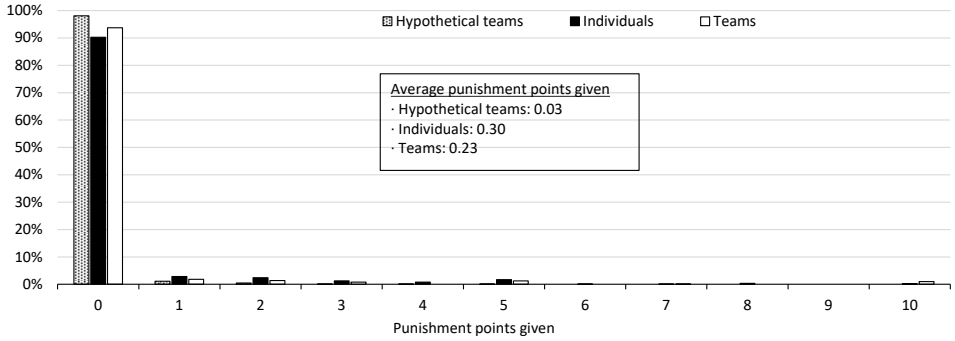



(ii) Distributions of Vote Outcomes

A. Voting on Sanction Rates in the FS scheme




(i) Pro-social Punishment Points Given




(ii) Anti-social Punishment Points Given

B. Informal Punishment Points Given in the IS Scheme

Peel 2000; Moffatt 2016). Table 2.7 reports the estimation results.[23] Two models were estimated by assuming different sets of three punishment types, as there are two approaches to define punishment patterns. 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 2.5.3.2. The selfish type is defined as a player who does not inflict punishment throughout.

Consider Models A.i and B.i to see behavioural differences between individuals and teams with a larger dataset. The results show that a larger percentage of teams, relative to individuals, are categorized as punishers who sanction 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:

[23] Typical to a maximum likelihood method, estimation results may depend on what starting values are assumed. In each model of Table 2.7, starting values were chosen to achieve the highest log likelihood.

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.[24] This implies that the issue of misdirected punishment is ubiquitous whether among individuals or teams. The estimated distributions of types are again inconsistent with Hypothesis 2 (team decision-making limits punishment acts as individual inclinations to punish are modest for the present subject pool), similar to the result in Section 2.5.4.

**Result 4:** *On average, a significantly larger percentage of teams, relative to individuals, were categorized as inflicting punishment on low contributors, while "misdirected" punishment types were similarly observed both for individuals and teams.*

The estimation results by the respective treatment further revealed that the percentages of pro-social or non-perverse punishers do not shrink by team decision-making, contrary to Hypothesis 2. Under the modest punishment intensity, the estimated percentages of pro-social (non-perverse) punishers *do not* differ much between individuals and teams. Under strong punishment intensity, the percentages of pro-social or non-perverse punishers are much larger in teams than individuals.

**Table 2.7:** *Estimated Percentages of Punishment Types in the IS Scheme*

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

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

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

## 2.7. **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 Appendix A.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). Appendix A.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 A.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.

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

As discussed in Section 2.5.3.1, 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 behaviour:

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 2.5.3.1 at the same time found that teams selected stronger sanction rates much more frequently than individuals (Figure 2.5). Thus, two additional codes were also considered to explain possible sources for this efficient voting behaviour 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 2.8.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 2.5). 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 behaviours as evidenced in Figures 2.2 and 2.3).

### *2.7.2.Informal Punishment Decisions in the IS Scheme*

Units, whether individual or teams, inflicted punishment not only pro-socially but also anti-socially (Section 2.5.3.2). Four codes are considered in the coding exercise to investigate motives behind these punitive behaviours:

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 A.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 2.8.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 2.8:** *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 $\chi^2$ | 136.33 | --- | 75.53 | --- | 78.49 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. The regression includes all C codes and G codes with Kappa being above 0.4, phase dummies, and the Period within phases variable as independent variables. The full estimation result can be found in Appendix Section A.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 $\chi^2$ | 172.55 | --- | 150.91 | --- | n.a. | --- |

| | | | | | | |
|---|---|---|---|---|---|---|
| Prob > Wald $\chi^2$ | 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 Appendix Section A.C.4.b. It should be noted that the alternative definition of punishment is perverse or non-perverse (Section 2.4.3.2). 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 A.C.4.b for the result. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

### *2.7.3. Contribution Decisions*

While both contribution levels and dynamics differed drastically according to the presence of the sanctioning schemes (Figure 2.2, Appendix A.B, Figure B.1), coding analyses, summarized in Table 2.9, 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).

We also consider codes specifically related to contributions under either scheme. To capture potential motivations under the FS scheme, 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 2.9 indicates that units' desire to avoid receiving fines, rather than material calculations, drove their strong contribution behaviours. 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.

Lastly, we consider how informal punishment opportunities may have affected decisions to contribute by introducing 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 of 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 2.9). This suggests that the mere presence of IS may help to raise group cooperation levels, and that units' reciprocal tendencies (detected in variables ii and iv) successfully sustained the positive cooperation norms.

**Table 2.9:** *Reasoning behind Units' Contribution Decisions*

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

| Codes included in the regression: | (1) No scheme | | (2) Under FS scheme | | (3) Under IS scheme | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| i. Contribute high always (codes A2, D1, E1 dummies) | 4.954*** | 0.617 | 8.916*** | 2.049 | 3.015*** | 1.201 |
| ii. Contribute high to encourage others to cooperate (codes A3, D3, E3 dummies) | 5.428*** | 0.612 | 2.380 | 2.594 | 5.558*** | 1.570 |
| iii. Contribute low always (codes A4, D2, E2 dummies) | -4.098*** | 0.625 | -6.524*** | 2.107 | -7.058*** | 1.143 |
| iv. Contribute low out of distrust (codes A5, D4, E4 dummies) | -4.429*** | 0.714 | -12.415*** | 2.700 | -7.788*** | 1.608 |
| 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 $\chi^2$ | 749.45 | --- | 212.5 | --- | 254.42 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

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

### *2.7.4.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 2.4, members discussed prior experiences/contributions/behaviours under IS and FS schemes in order to decide which sanctioning scheme to vote for (Code B11). Appendix A.C.4.d includes the detail of the estimation results.

## 2.8. Conclusion

Team decision-making is ubiquitous whether in the public or private sphere. The literature in the theory of the firm has so far assumed that team decision-making is equal or inferior to individual decision-making due to imperfect information, monitoring issues, and agency costs (e.g., Alchian and Demsetz 1972; Marschak and Radner 1972). In addition, team decision-making has received no attention in the experimental literature in an institutional choice setting to date. While during the last two decades numerous scholars have studied members' institutional choices and self-governance possibilities by letting them vote in experiments (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter

*et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018), no studies used teams as the decision-making unit (voter). Using individuals as the decision-making unit could be a nice simplification if the following assumption is correct: teams make the same institutional choices as individuals on the condition that the former hold the same information and face the same incentive structure as the latter. However, to the authors' knowledge, there is no research to compare institutional formation and behaviours under endogenously selected institutions between individuals and teams, although there is large literature on team decision-making in economics, psychology and management (e.g., Charness and Sutter 2012, Kugler *et al.* 2012, Kerr and Tindale 2004). Moreover, little research has been conducted to study the role of team decision-making in the empirical literature in management and organizations.

This paper demonstrated, for the first time, that teams may be more able than individuals to form efficient institutions by voting through communication and therefore overcome free riding in groups more competently. In the experiment, decision-making units, teams or individuals, were given a voting opportunity to either construct a formal sanction scheme or to use informal punishment in a public goods dilemma. Teams made single voting decisions through communication. The results showed that teams achieved surprisingly higher levels of group contributions than individuals. 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 much more frequently than individuals. The difference in voting is remarkable: while the majority of individuals in the individual treatments voted for the weakest non-deterrent rate, teams voted for deterrent sanction rates more than 50% of the time. When peer-to-peer punishment was instead selected, teams inflicted costly punishment more frequently on low contributors than individuals. The high effectiveness of team decision-making cannot be explained by the mere aggregation of individual preferences: for example, teams would vote for the zero sanction rate more frequently than individuals, were individuals in the individual treatments assigned to a team of three and their preferences aggregated. It may therefore be the case that teams achieve better choices than individuals through deliberation and learning.

The present paper is related to the large literature on the theory of the firm. Here, delegation of the decision rights to employees is often discussed as a way to motivate workers[25] and to theoretically resolve intra-firm inefficiency due to the informational advantage of the workers (e.g., Gibbons and Roberts 2013). An extreme organizational form of this is employee- or worker-ownership/management

[25] Charness *et al.* (2012) experimentally showed that delegating a compensation choice to employers in a firm induces them to put in more effort.

(e.g., Pencavel, 2001; Gibbons *et al.*, 2013). There is a recent rise in employee empowerment, engagement and participation, and greater interest in employee welfare in firms in many countries.[26] Collective selection of an appropriate incentive scheme in the group, whether formal or informal, may be crucial for a firm with a stronger decentralized organizational form because the employees then have more power and discretion, and are thus arguably more susceptible to the so-called 1/N problem (Alchian and Demsetz 1972) than without delegation. While in the theory of the firm team decision-making is usually treated as a coordination problem in which the same processes involved in individual decision-making are used, but feature additional complexities relating to imperfect information, monitoring, and agency costs in the team (e.g., Alchian and Demsetz 1972; Marschak and Radner 1972; see also Gibbons *et al.* [2013] for a survey),[27] the present experiment demonstrated that given the ability to deliberate and vote, teams propose better incentive schemes than individuals in a decentralized system. This suggests that, in making organizational decisions, the firm should let work teams or divisions deliberate optimal institutions, rather than just surveying and collating individual preferences.[28] This view is similar to Bainbridge (2002) who argue that teams may be superior decision-makers to individuals when humans are boundedly rational, due to the teams' increased ability to store and process information. The discontinuity effect detected in this study suggests a need to bolster existing theory, perhaps explicitly incorporating the beneficial communication, deliberation, and influence process inside a team.

Having said that, it should be acknowledged that the design used for the present experiment lacks an important managerial element in this theoretical literature, namely monitoring, for the sake of simplicity. The experiment used an environment in which units establish incentive schemes, either IS or FS, by voting; but in such a democratic workplace, as Barron and Gjerde (1997) discussed, workers themselves may be (at least in part) responsible for managing the selected scheme, say, by undertaking a role in monitoring fellow workers' contributions. This presents an issue because peer monitoring is subject to second-order free riding.[29] While the present experiment assumes that members' contributions are always observed without noise in the endogenous IS scheme, and formal punishment is inflicted accurately on non-contributors in the endogenous FS scheme, investigating how the cost

[26] Even in some countries (such as Japan) where firms traditionally had more hieratical structures, a shift to an employee-centered workplace is encouraged to enhance productivity by the government as work style reform.
[27] Prior research in management has thus explored effective ways to coordinate and share information held by workers in organizations (e.g., Grant 1996).
[28] The superiority of teams over individuals can also be applied to club goods, such as agricultural collectives which sign up to be centrally governed to meet a goal as a collective and to be penalized for breaching their contract/not meeting their targets (the so-called "climate club" is a similar example in this category).
[29] Kamei *et al.* (2019) studied precisely this issue in the context of political accountability, i.e., monitoring (civic engagement) is needed to function the formal sanctioning state, but monitoring is subject to second-order free riding.

and format of such schemes, such as the use of "peer pressure" (Barron and Gjerde, 1997), affects the superiority of teams remains for future research.

This paper also contributes to empirical research on management, organizational economics, and personnel economics that studies team decision-making and team production. 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.* 2009; Aboramadan 2020; Certo *et al.* 2006). However, it is difficult to draw causal inferences from these studies for various reasons, for example, because there is possible selection bias in the management team formation, and many studies rely on self-assessed/reported questionnaires. Prior empirical research in personnel economics also argues that team production and worker participation (such as that in production sites) lead to better work performance than individual production (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. The present experiment suggests a strong beneficial effect of having team decision-making per se in terms of institutional formation in the workplace.

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. There are many directions for further research. For example, this study set both the team size and group size to three. The sizes of teams and/or groups could be much larger in real organizations, however. It would therefore be a useful robustness check to study the same research questions by changing the group size and/or team size. As another example, the three team members communicated with each other anonymously, i.e., without being allowed to disclose their identifiable information, to jointly make a single decision in the experiment. While this design setup is standard in the current experimental literature (e.g., Charness and Sutter [2012], Kugler *et al.* [2012] and Kerr and Tindale [2004]), in the typical workplace environment (excluding some anonymous online work) members of a team are fully or partially aware of the identity of each other. It would be worthwhile studying how the discontinuity-effect phenomenon differs by the anonymity condition within teams. Another important direction of further research is to study possible discontinuity effects when conflicts among team members prevail (e.g., Glaetzle-Ruetzler *et al.*, 2021). The present study assumes for simplicity that the three members in a team received the same payoffs, but there are many real-world situations where team members receive different payoffs. Lastly, of course, the finding of this research also opens up further avenues for theoretical research, as according to the

finding of the present experiment, teams (as decision-making units) make different institutional choices through deliberation and learning compared to individuals, even when facing the same incentive structure.

# 3. Free Riding, Democracy and Sacrifice in the Workplace: A Real Effort Experiment

## 3.1.Introduction

Maintaining motivation among workers is often difficult when private interests conflict with group interests in the workplace (e.g., Bolton and Dewatripont, 2004) — a typical example of this is moral hazard in teams (e.g., Alchian and Demsetz, 1972; Holmstrom 1982). Democratic culture may help mitigate the conflict by not only enhancing their self-determination and intrinsic motivation to cooperate (e.g., Deci and Ryan, 1985, 2000), but by also providing workers with opportunities to signal their willingness to cooperate with their peers through democratic processes (e.g., Connelly *et al.*, 2011; Bergh *et al.*, 2014), thereby making it easy to achieve the group optimum. In such environments, workers may decide to voluntarily sacrifice their private gains for the sake of group interests. But precisely what motivates workers' sacrificial behaviours? How large could the effects of endogenous decisions per se on productivity potentially be?

How to overcome moral hazard in teams is an important, sought-after question in economics and management. A large body of research spanning several decades has found that workers have difficulty cooperating with each other when free riding incentives are sufficiently strong in a social dilemma (e.g., Ledyard, 1995; Zelmer, 2003). Specifically, prior experimental research suggests that while some people demonstrate conditional willingness to cooperate, groups usually cannot sustain cooperation for various reasons, e.g., their cooperation behaviours are heterogeneous (e.g., Fischbacher *et al.*, 2001), they are easily discouraged by seeing their peers free ride (e.g., Fischbacher and Gächter, 2010); or many tend to cooperate but by less than others (e.g., Thöni and Volk, 2018). This echoes theoretical research that describes why moral hazard arises among workers when their effort levels are not perfectly observable (e.g., Alchian and Demsetz, 1972; Holmstrom, 1982).[30] Both the theoretical and empirical literature therefore discuss that some institutional solutions, such as competition (e.g., internal job ladder, tournament), punishment and rewards, monitoring, and sorting, are required to assist collaboration and cooperation in the workplace (see, e.g., Prendergast 1999 for personnel economics, and Chaudhuri 2011 for experimental, literature). This study contributes to the large body of literature by investigating workers' behavioural reactions to a reduction in incentives to shirk, the impact of

[30] The difficulty in sustaining cooperation has also been widely discussed in the theoretical literature in the voluntary provision of public goods (e.g., Samuelson, 1954; Bergstrom *et al.*, 1986).

democratic decision-making in a workplace setting, as well as the reasoning behind sacrificial behaviours in the workplace.

Collectively sacrificing one's benefits through fostering customs, conventions, or rules with the aim of resolving conflicting interests has been conceptually discussed in literature in the social sciences (such as anthropology) and biology as key features of humans. Examples include costly participation in religious groups and rituals, or recreational activities in societies (e.g., dance and festivals), food sharing (e.g., turtle hunting by islanders for funerary rituals), holding redistributive feasts, and attending group raids and defence (see, e.g., Smith and Bliege, 2000; Hawkes and Bliege, 2002; Sosis and Alcorta, 2003; Sosis and Bressler, 2003; Hagen and Bryant, 2003; Iannaccone, 1992). The mechanism is described as follows: sacrificing serves as a costly signal of one's own quality (e.g., Gintis *et al.*, 2001; Bliege and Smith, 2005), thus helping to coordinate with others to cooperate and bolster a cooperative atmosphere in dilemma situations.[31][32]

Parallel to these arguments, several laboratory experiments used public goods games or prisoner's dilemma games to study costly human sacrificing tendencies (e.g., Aimone *et al.*, 2013; Brekke *et al.*, 2011; Grimm and Mengel, 2009). The findings are that some groups (individuals) do collectively (voluntarily) sacrifice their private returns, thereby enhancing welfare. However, to the best of the authors' knowledge, sacrificing has not been studied in the workplace context using a naturally-occurring, real effort task, although recently there has been a theoretical attempt to characterize the effects of sacrificing in the workplace (Bisetti *et al.*, 2022).[33]

While sacrifice has received less attention in the workplace so far, it is becoming more and more relevant due to a surge in remote working (potentially boosting shirking) triggered by the Covid-19 crisis and technological advances. A broad range of examples of unobserved

[31] In general, many actors' decisions are characterized as costly signaling in modern societies. Examples include the job market, in which applicants invest in education or other qualifications to indicate their quality (Spence, 1973), or at the firm level by which firms indicate their quality to other firms, the market, or other stakeholders through investment in high profile board members, awards, alliances, or underpricing (see Bergh *et al.* 2014 for a review and examples).

[32] Empirically, people are known to choose transaction partners in dilemma situations based on factors that inform the quality of that partner. Elfenbein *et al.* (2012), using a novel data set composed of more than 160,000 eBay listings, successfully demonstrated that in online marketplaces, buyers tend to purchase products tied to charity, and thus sellers have incentives to use a charity program (e.g., eBay's Giving Works program) as a quality signal.

[33] Bisetti *et al.* (2022) propose a self-reporting mechanism in which a team's pay is based on their observed joint output and their team's self-reported performance. They prove that a team has the incentive to under-report their group's performance (sacrifice wages for all in the team) as a punishment to free-riders, thereby enabling them to achieve higher welfare.

shirking activities and countermeasure policies are readily available in the modern workplace. For example, cyberloafing is a typical and costly issue whereby employees covertly use their computer or internet access for personal use during work time. The issue is especially serious when they are not in an office. The employer may decide to introduce measures to counter employees' cyberloafing, for example, by monitoring their use of the internet, imposing internet restriction policies and penalties for breaching them, or placing technical restrictions on employees' access to certain non-work websites.[34] While such policies can simply be imposed from above by managerial staff or teams, the policies can also be enacted through decentralized decision-making. For instance, a factory may produce mechanical parts by assigning workers to several teams to take advantage of specialization. When their environment is democratic and they recognize that cyberloafing undermines productivity, they may democratically decide to enact a restriction policy across the teams, with an aim of improving the performance in the factory if they believe that their productivity impacts their material benefits such as their wages, bonuses, or rewards. Similar scenarios are common across various employment relationships, e.g., a branch in a consulting firm, or a sales office for products (e.g., cars). Another related example is "moonlighting" by which employees work multiple jobs, sometimes simultaneously and/or without the permission of their main employer.[35] For example, an employee may commit to working five days per week while secretly working for another firm to earn more by shirking the main job. Alternatively, an employee may hold a secondary side job that takes place outside of their primary work hours, but spend time during those hours contributing to their secondary job, such as responding to e-mails, advertising, or checking their website. This behaviour is quite relevant given the increase in remote working in recent years, which makes monitoring more difficult. Policies to make working on the side difficult and materially unbeneficial (e.g., through using a screen-capture tool and work-time tracking) may be considered if such free riding significantly undermines production in the main workplace.

This paper conducts an experiment with a novel "collaborative" real effort task. In the experiment, worker subjects are randomly assigned to a team of three, and three teams constitute a group. The real effort task requires each team to jointly calculate the number of 4s in a matrix

[34] Strengthening monitoring increases the probability that cyberloafing is detected and penalties are assigned, thereby reducing workers' incentives to cyberloaf. As will be described soon, for the sake of simplicity, the present paper considers a policy to reduce material returns from shirking deterministically in the workplace in the experiment.

[35] Moonlighting is increasingly common in some countries because it is encouraged by the government. For example, lifetime employment was a common practice in Japan traditionally. However, the Japanese Ministry of Health, Labour and Welfare published the "Guidelines for Promotion of Side Work" and deleted the description of prohibition of subsidiary business from "The Model Rules of Employment" in 2018.

whose cells contain 1s, 2s, 3s, or 4s. At the onset of the experiment, each team member is assigned a number, player 1, 2, or 3, such that they have different numbers from each other. The matrix that player *k* is allocated includes only number *k*s while the other three numbers are blacked out. Each member counts their assigned numbers, shares the counting outcome, and jointly calculates the final answer, on the condition that their remuneration is based on revenue-sharing in the group. To mimic the conflict between work and shirk (or another activity) in the real workplace, each member is allowed to privately and independently play a computer game, Tetris. Before the task-solving phase begins, a policy that reduces the incentive to play Tetris ("reduction policy," hereafter) is implemented in a group either *democratically* (by voting) or *autocratically* (randomly by the computer without voting). The two treatments (democratic, or autocratic) are designed using a between-subjects design.

This experiment is novel, particularly in three aspects. First, this study provides the first experiment to measure the so-called "dividend of democracy" when the decision-making and work units are teams. Prior research has shown that democracy in implementing a pro-social policy boosts cooperation in experimental games, such as public goods or prisoner's dilemma games, as it directly affects people's own behaviour and beliefs on their peers' cooperativeness (e.g., Tyran and Feld, 2006; Dal Bó *et al.*, 2010; Sutter *et al.*, 2010; Kamei, 2016). Scholars have recently started to study the applicability of such a dividend of democracy in a workplace setting by using a design with real effort tasks, but the results surprisingly showed that democracy per se may not have strong effects in real effort settings (e.g., Dal Bó *et al.*, 2019; Kamei and Markussen, forthcoming; Melizzo *et al.*, 2014). While all prior experiments on democracy used individuals as the decision-making unit, the present study uses teams as the decision-making unit of policy-making and task-solving for the first time, and find a significant dividend of democracy on work productivity (per-work-time production) as consistent with the earlier research work in experimental games. Teams are increasing more popular in firms as decision-making units, as discussed in Kamei and Tabero (2022). It is worthwhile studying the role of democratic culture, as the literature suggests that teams behave differently from individuals under certain conditions (e.g., see Charness and Sutter [2012], Kugler *et al.* [2012] and Kerr and Tindale [2004] for a survey) and that team decision making differs significantly from individual decision making because the former features a coordination problem that involves complexities relating to imperfect information, monitoring, and agency costs (e.g., Alchian and Demsetz, 1972; Marschak and Radner, 1972).

Second, the experiment is the first to investigate workers' sacrifice decisions and their reasoning in a real effort environment. While prior research used experimental games such as public goods games to propose that some individuals will reduce their private gains in dilemma situations, showing that such decisions lead to a Pareto improvement empirically (e.g., Aimone *et al.*, 2013; Brekke *et al.*, 2011; Grimm and Mengel, 2009), its validity in the workplace setting is unclear as little research used naturally-occurring, real effort in their experiments. Equally important is that no research explores what may drive workers to sacrifice their private gains, because no data is available regarding their thinking. Subjects in the present experiment decide whether to reduce their private gains through communication within their team as a team decision. This design enables us to collect a unique incentive-compatible dataset to study the reasoning behind sacrifice decisions. A well-established coding exercise is applied to the communication logs in order to uncover reasoning effectively.

Third, this study provides significant methodological contributions with the newly used "collaborative" counting task and gaming as a real activity. While much research has been conducted using real effort tasks, a significant issue has been reported by Araujo *et al.* (2016) that workers' incentive elasticity of outputs may be too small with the real effort tasks. Recently, Corgnet *et al.* (2015) and Kamei and Markussen (forthcoming) allowed subjects to use, respectively, internet browsers and comedy videos, as real leisure activities. Both of the papers showed that such activities enhance incentive elasticity in experiments. The present paper adds to the literature by using gaming as a real, but controlled, leisure activity for the first time in a computerized real effort experiment. Further, the members of each team jointly work on a *collaborative* counting task. While an individual counting zeros task is widely used in the literature (e.g., Falk *et al.*, 2006; Abeler *et al.*, 2009; Kamei and Markussen, forthcoming), the use of a collaborative version is the first attempt in the literature, to the authors' knowledge. This design is meaningful as collaboration is a central aspect of teamwork in many firms. Notice the difference in the game structure between the standard counting task and the collaborative counting task. The collaborative one is a coordination game: they earn from the team task only when all three members work by spending time counting and communicating accurately and effectively.

The experiment results reveal some teams' preferences for sacrifice and evidence of a dividend of democracy. 40.9% of teams voted to reduce the incentive to play the game, and as a result, the reduction policy was enacted for 38.7% of groups. Teams that were involved in democratic decision-making exhibited significantly higher work productivity, i.e., performance per minute of working, than those in the regime where the computer randomly decided policy

implementation, whether the reduction policy was imposed or not. This means that the democratic culture per se directly affected behaviour. Having said that, the workers under democracy reduced work time compared to those under autocracy, presumably due to more quickly accumulated fatigue of the former. Nevertheless, the former did not decrease team production overall thanks to the enhanced work productivity.

A coding exercise on their sacrifice decisions reveals that the units that planned to exclusively work on task-solving, believed that the reduction policy would deter others from shirking, or those that had supportive team atmospheres supported the reduction policy. It also uncovers the value of signaling through sacrificial decisions to encourage collaboration: teams who believed that other teams would complete tasks following the vote performed strongly.

The rest of the paper proceeds as follows: Section 3.2 summarizes the experimental design, and Section 3.3 reports the results. Section 3.4 provides insights obtained from an analysis of communication dialogues, and Section 3.5 concludes.

## 3.2. Experimental Design

The experiment is designed using a collaborative real effort task devised for this study. At the onset of the experiment, worker subjects are randomly assigned to a team of three. The three members are then randomly assigned ID numbers, 1, 2, or 3, so that each member receives a different number from one another. Anonymity is retained such that they do not know the identity of the other members (e.g., faces, names, gender). Let us call the player who is assigned number $k \in \{1,2,3\}$ "player $k$." The team composition and the assigned ID numbers do not change for the entire experiment (partner matching). Three teams further constitute a group (each group thus has nine members). The group composition also does not change throughout. Section 3.2.1 explains the nature of the collaborative team real effort task, after which Section 3.2.2 explains the structure of the experiment, a summary of treatments, the remuneration system, and the sacrifice policy that could be implemented in each group. Appendix B.A summarizes the experimental procedure and includes instructions used in the experiment.

### *3.2.1. A Collaborative Real Effort Task*

Three members in a respective team collaboratively solve a variant of the counting task ("collaborative counting task"). The original "counting task" (e.g., Falk *et al.*, 2006; Abeler *et al.*, 2009; Kamei and Markussen, forthcoming) is an individual real effort task in which subjects independently count the number of 0s in a matrix that contains 0s and 1s. To the authors' knowledge,

no collaborative version of the counting task has been devised and used in any prior experiments. In the new collaborative counting task, the three team members are provided with a 15×15 matrix, each cell of which has a randomly generated integer between 1 and 4 (each integer is independently drawn with a probability of 25%), and are then asked to submit the number of 4s. Collaboration is required to find the correct answer, because only number *k*s appear on the computer screen of player *k*, while the other three numbers are blacked out – see Figure 3.3.1 for a screen image for player 1. Each team can find the correct answer if player *k* counts the number of *k*s correctly and shares it with their teammates, and the team calculates the number of 4s accurately after that. For example, if the numbers of 1s, 2s, and 3s are, respectively, 32, 14, and 43, then the correct answer (the number of 4s) is: 225 – 32 – 14 – 43 = 136. A calculator is available on each subject's computer screen. How to calculate the number of 4s, and by whom, is up to each team's discretion. When the team decides on and wants to submit the answer, all three members must submit the team's joint answer on their own computer screens. Hence, in the submission stage as well they must communicate with each other about their team's decision to answer correctly. In the case of disagreement, a member can submit a different answer from the others.[36] However, the answer will then be counted as incorrect. Once all three members submit an answer, a new 15×15 matrix with randomly generated 1s, 2s, 3s, and 4s in each cell is assigned to the team, and the process repeats.

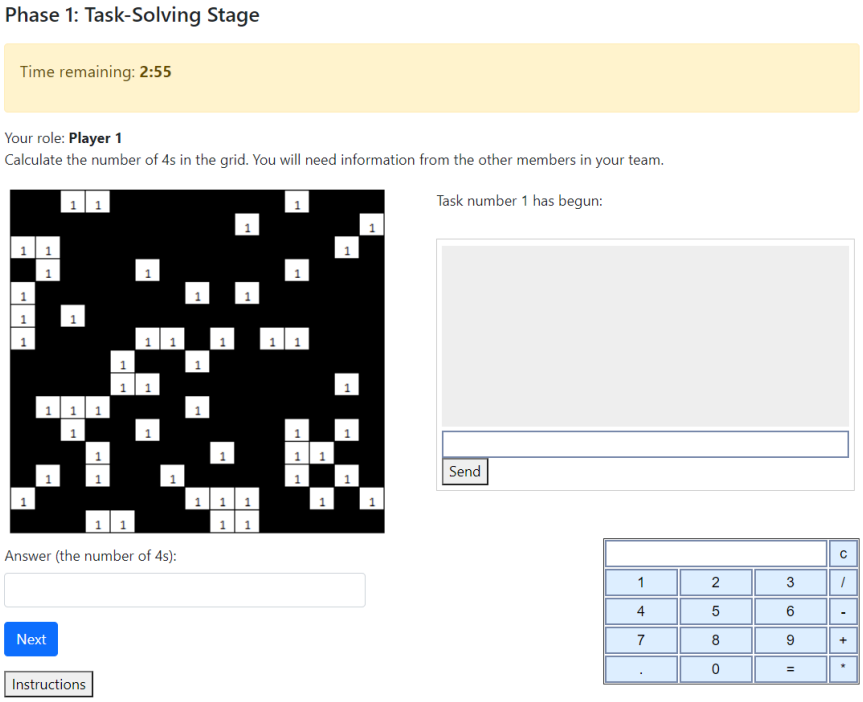
Free-form communication is available using an electronic chat window during the entire task-solving process (see Figure 3.3.1 again; Appendix B.A also includes the screen image of the chat window). This design piece helps the researchers study the reasoning behind members' behaviours, post-experiment. While any sort of communication, such as discussing strategy to solve the problems, sharing the number of *k*s, or chatting about unrelated matters, is allowed, subjects are prohibited from using any kind of offensive language or sharing any information that compromises anonymity.[37]

The more questions a team answers correctly, the higher the earnings they can generate in their group. Each correct answer is rewarded with 180 UK pence in the experiment. How the 180 pence are distributed within the team or the group is explained in Section 3.2.1.

[36] This very rarely happened in the experiment. All three members submitted the same answers in 96.9% of teams' submissions in the experiment (3,176 out of 3,278 completed tasks in the 62 experiment sessions). The authors read through all the communication dialogues and their submitted answers, and found that almost all disagreements are errors or typos. The mean number and the mode of disagreements across all teams that disagreed were, respectively, 1.72 and 1. The size of the error rate is unsurprising because the average number of attempts for these teams was 24.14 questions, above the average of 17.81 for the experiment, which might increase potential errors in typing.

[37] The authors read through the communication dialogues and found no team to have broken the anonymity rule.

**Figure 3.1**: *A Screen for Collaborative Counting Task*




*Notes*: A screen image for player 1. The numbers of 2s, 3s and 4s are blacked out on the screen that player 1 sees. The 15×15 matrix in this figure is for illustration only.

### *3.2.2. The Experiment*

There are two treatments that vary by changing the process to decide whether to enact a policy to curb members' shirking or not. A between-subjects design is used to avoid behavioural spillover (e.g., Bednar *et al.*, 2012) or possible spill-over effects of democracy (e.g., Kamei, 2016). The experiment begins with a practice phase, which is the same for all subjects in the experiment. The main task-solving phase begins after the practice phase and differs by the treatment.[38] The practice phase plays a role in not only familiarizing subjects with the collaborative counting task, but also providing them with an opportunity to try the task and learn their ability to solve it.

In the practice phase, each team performs the collaborative counting task for three minutes.[39] While they can answer as many questions as they wish, they are not informed whether they answer

[38] The practice phase and the main task-solving phase are called "phase 1" and "phase 2" in the experiment instructions.

[39] To avoid cognitive overload, subjects are provided with instructions for the practice phase only at the beginning of the experiment. Instructions for the main task-solving phase are distributed once the practice ends. Such gradual learning approach is often taken in experiments (e.g., Ertan *et al.*, 2009; Kamei *et al.*, 2015; Kamei and Tabero, 2020).

each question correctly during the three-minute period. They are instead informed of the number of correct answers at the end of the practice phase. Remuneration is based on revenue sharing in the team. This means that the money a team earns is equally divided among the three team members (each member receives 60 = 180/3 UK pence for a correct response). Each team does not interact with the other two teams in their group in this practice; nor are they informed of the performances of the other teams.

In the main task-solving phase, each team performs the collaborative counting task for a much longer duration – 35 minutes – with a revenue sharing rule in their *group*. This means that the credit of each correct answer (180 UK pence) is equally shared among the three teams, i.e., nine individuals as each team has three members. The marginal per-capita return is calculated as 20 (= 180×1/9) UK pence.

There are two more distinct aspects in the main task-solving phase. First, unlike the practice phase, each member can privately shirk by playing Tetris. They can do so by simply pressing the "Game" button (Figure 3.2.a). The screen is then switched to the Tetris site (Figure 3.2.b). No one, including their teammates, are made aware of a member's shirking unless the member voluntarily reports their behaviour using the electronic chat window. Further, the shirker earns a return by staying in the Game screen: 18 pence per minute spent in the Game screen.[40] They can return to the work site from the Game site at any time. Workers are *not* allowed to work while playing Tetris, whose requirement enables the researchers to quantify shirking versus work time as their work decisions. It should be noted here that the design of gaming was carefully made to enhance external validity, as workers often have alternative activities available when shirking in the workplace rather than being inactive. An advantage of using gaming over internet browsing (Corgnet *et al.*, 2015) as an alternative activity is the high level of control: workers may use internet browsers differently as their preferences are heterogeneous. This feature shares similarities with Kamei and Markussen (forthcoming) that adopted comedy video clips as an alternative activity. However, using a game is better than video clips because implementation is difficult with the use of the latter. While headsets were provided to each subject in Kamei and Markussen (forthcoming), the authors acknowledged that even a small

[40] This return can be thought of as material returns that can be obtained from shirking in the real workplace. Shirkers may build their social network using social media or by exchanging emails during work time, develop skills to benefit future job prospects, complete personal tasks, or even moonlight privately as in the real-world example described in the introduction of the paper. Such activities may not only provide intrinsic satisfaction but may also provide material benefits. A similar designing approach was chosen in Kamei and Markussen (forthcoming) where an activity alternative to solving a real effort task is to watch a funny video. Subjects in Kamei and Markussen (forthcoming) received a small return per minute watching the videos.

ripple of laughter and sounds could contaminate the data. In contrast, gaming is a purely independent, quite leisure activity.

Notice that with the gaming option, the incentive structure of the team task in main phase is one of the so-called "stag-hunt game" if they are highly skilled. Each team member can earn a small material gain with certainty by deviating from collaboration. However, they earn a large team payoff when all three team members work on the counting task, if each of them can count numbers sufficiently quickly.

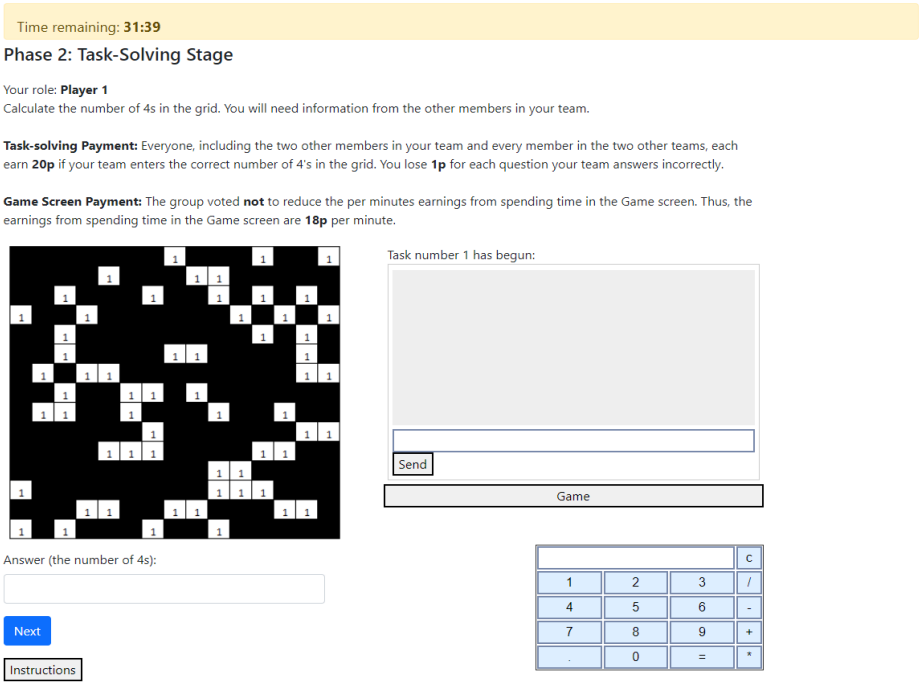
Second, there is a penalty of three pence per incorrect answer in the main task-solving phase. This penalty is imposed on the team that commits the error, not the whole group. Such penalties are commonly used in the real workplace; for example, poor performance or mistakes can result in monetary or social sanctions, increased threat of dismissal (through escalation procedures or informal threats), or reduced pay where performance related wages or bonuses are in place (see McNamara *et al.*, 2022; Doellgast and Marsden, 2018; Gibbons and Henderson, 2013, for examples). The penalty is equally shared among the three members in the team (i.e., one penny is deducted from the payoff per team member). In short, the payoff of member *i* in team *k* can be expressed as Equation (1):

$$\pi_{k,i}(c_k, ic_k, g_i) = 20[\sum_{n=1}^{3} c_n] - ic_k + r \cdot g_i, \quad (1)$$

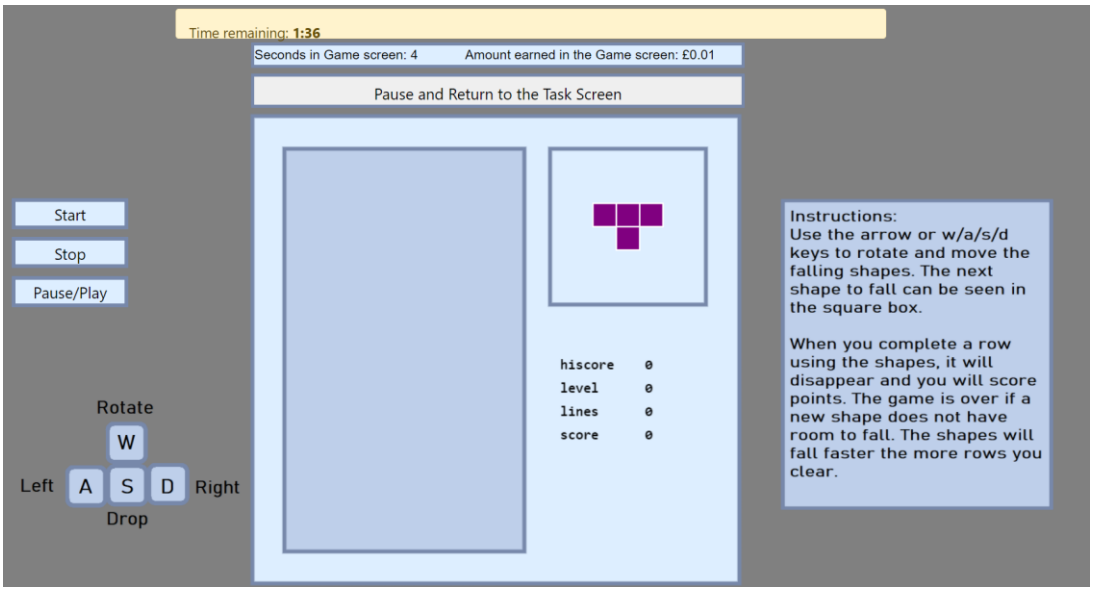
where $c_k$ and $ic_k$ are the numbers of, respectively, correct and incorrect answers by team *k*, $g_i$ is the time [minutes] that member *i* spends in the Game screen, and *r* is per minute return from shirking. Notice that their work time is $35 - g_i$ as they are not allowed to work while playing Tetris. Using the revenue sharing rule per group and the alternative leisure opportunity, the aim is to model the work environment as a tension between task-solving and gaming (i.e., social dilemma). As intended, gaming was a privately optimal option for almost all teams in the experiment sessions – see Section 3.3.

Worker subjects are not informed of how many questions they answer correctly during the 35-minute task-solving phase. Instead, at the end of the task-solving phase they learn (a) the total number of correct and incorrect responses of their own team and (b) the total number of correct responses in their group. This setup is realistic; for example, in manufacturing, the manager will learn how many defectives they have among mechanical parts produced in a given day, only after quality checks at specified intervals.

**Figure 3.2**: *A Screen Image for Collaborative Counting Task in the Main Task-Solving Phase*




(a) Work site




(b) Game site

At the beginning of the main task-solving phase, the return from staying in the Game screen ($r$) could decrease from 18 to 16 pence per minute. Notice that the size of the incentive change is very small at only two pence. This means that the reduction policy can be thought of as a non-deterrent

sanction policy, i.e., a policy that *does not* alter the privately optimal behaviours of workers in a group (e.g., Tyran and Feld, 2006; Kamei, 2016). As briefly reported in Section 3.3., this interpretation turns out to be correct in the experiment: gaming was a privately optimal choice for almost all subjects, whether the reduction policy was in place or not, due to the strong incentives to free ride on other teams' work efforts.

The process to implement the reduction policy differs by treatment. In the EXO treatment, the policy is imposed in each group by the computer randomly (i.e., with a probability of 50%). By contrast, in the ENDO treatment, the policy is implemented based on majority voting by the three teams.[41] The voting procedure follows three steps:

Step 1. The three members in each team are given three minutes to discuss, using an electronic chat window (e.g., Kamei 2019b; Luhan *et al.*, 2007), whether they want to reduce the per minute earnings from staying in the Game screen. The communication contents are not revealed to any other team. See Appendix B.A.3 for a screen image of this step.

Step 2. After the three-minute discussion, the three members each submit their preferred decisions. If the three submit the same decision, it becomes their team vote. However, in the case of disagreement they can submit whatever they prefer, in which case whichever receives at least two members' support is implemented as their team vote.

Step 3. The reduction policy is implemented in the group based on majority voting. Specifically, it is implemented (not implemented) if it receives two or three supporting (opposing) team votes. All subjects in the group are informed of the vote outcome and the number of supporting votes.

Notice that as the reduction policy, despite the size of the reduction being small, may encourage teams to work harder through decreasing the material incentives to shirk, thereby leading to a higher payoff, groups may decide to sacrifice such private returns by voting. As summarized in Table 3.1, there are four possible institutional outcomes in this study.

[41] While another realistic voting method is a unanimity rule (consensus), this study adopted majority voting because the interpretation of data becomes complex when the unanimity rule is in use as it possibly involves strategic voting among voters (e.g., Battaglini *et al.*, 2010; Kamei, 2019a).

**Table 3.1:** *Treatments, Distribution of Votes and Institutional Outcomes*

| Treatment and institutional outcome | Condition in which the policy is/isn't implemented | # subjects | # of subjects in pro-reduction teams | # of subjects in anti-reduction teams |
|---|---|---|---|---|
| **ENDO treatment** | Voting | 279 | 114 | 165 |
| (i) Policy was implemented | At least two teams vote for the policy | 108 | 75 | 33 |
| (ii) Policy was not implemented | At least two teams vote against the policy | 171 | 39 | 132 |
| **EXO treatment** | By the computer | 273 | --- | --- |
| (i) Policy was implemented | Randomly (50% probability) | 123 | --- | --- |
| (ii) Policy was not implemented | Randomly (50% probability) | 150 | --- | --- |
| Total | --- | 552 | 114 | 165 |

*Note:* The numbers in the "# of subjects in pro-reduction teams" and "# of subjects in anti-reduction teams" columns are based on the results of voting in the experiment.

### *3.2.3. Theoretical Predictions*

Theoretical predictions can be derived by setting a utility function for the player and then finding their utility-maximizing behaviour. As shown in online Appendix B.B, a calculation suggests that teams work harder with than without the reduction policy in a given institutional condition (ENDO or EXO), and that the positive effect is stronger in the ENDO than in the EXO treatment, for the following reasons. First, the positive effect of the reduction policy holds theoretically for the EXO treatment because the policy reduces the material incentives of shirking. As the reduction policy is imposed randomly in each group, in theory there are no differences in individual characteristics between the groups where the policy is imposed or not. Thus, only the material incentives matter in this treatment due to the lack of selection. Second, the positive effect is also applicable to the ENDO treatment, not only due to the beneficial effects of incentive changes, but also possible selection effects through voting. The reduction policy is enacted in the ENDO treatment only when the majority of teams support the policy. Considering that teams who are *better* at solving the collaborative counting task can be assumed to incur *smaller* effort costs for a given effort level, the beneficial effects of the policy on hard work exceed enhanced effort costs more easily for such higher-skilled teams. This means that higher-skilled teams are more likely to enact the reduction

policy by voting, and to perform strongly in the ENDO treatment. In other words, the impact of the reduction policy is detected more strongly in the ENDO treatment due to selection.

It should be worth remarking here that, theoretically, the positive effect of the reduction policy does not emerge when task-solving is too costly for teams. If the return from shirking as a team is much larger than the marginal return from working, members in selfish teams will just stay in the Game screen even when the sorting effects are present in the ENDO treatment.

The main hypothesis of the paper is on the dividend of democracy summarized below:

**Hypothesis**: *Teams put more effort into task-solving in the ENDO than in the EXO treatment, even after controlling for possible selection effects.*

The phenomenon summarized in this hypothesis is the so-called dividend of democracy. Its mechanism lies in the democratic process that directly influences worker tendency (e.g., Dal Bó *et al.*, 2010; Dal Bó *et al.*, 2019; Sutter *et al.*, 2010; Tyran and Feld, 2006; Kamei, 2016). In a workplace setting, Kamei and Markussen (forthcoming) model this effect such that workplace democracy lowers workers' marginal effort costs. A model similar to Kamei and Markussen (forthcoming) supports the hypothesis above; a decrease in the marginal effort costs driven by democracy results in hard work among teams (see Appendix B.B for the detail). Part of the dividend of democracy can also be attributed to signaling effects (e.g., Tyran and Feld, 2006; Kamei, 2019a; Jensen and Markussen, 2022).

It should be noted that identifying the dividend of democracy requires care because of the possible selection bias already discussed (Dal Bó *et al.*, 2010; Dal Bó *et al.*, 2019; Tyran and Feld, 2006). By design, pro-reduction teams are overrepresented (underrepresented) in groups where the reduction policy was (was not) endogenously enacted. As voting behaviour is likely related to teams' skills and work behaviour, group behaviours are not comparable between the ENDO and EXO treatments unless the distributions of votes are balanced. The present paper adopts the "weights-based identification strategy" proposed by Dal Bó *et al.* (2019). This estimation method uses weights under the *whole* population when calculating the average behaviour in the ENDO treatment, rather than the realized vote shares in specific institutional outcomes. For instance, suppose that 50% of teams vote for the reduction policy and the policy is imposed in 50% of groups. The % of pro-reduction teams would be much more (less) than 50% in groups where the policy is (is not) endogenously imposed because of majority voting. Instead of the high (low) percentage in such groups, 50% is used as a weight in calculating the average behaviours of pro- and anti-policy units with this method. The detail of the re-weighting method along with the data will be provided in Section 3.3.

### 3.3. Sacrifice, and the Dividend of Democracy

552 students (279 for the ENDO treatment and 273 for the EXO treatment) at the University of York in the United Kingdom participated in the experiment. No subjects participated in more than one session. The experiment followed standard practices in economics, such as neutral framing. Appendix B.A includes the procedure and the instructions.

Table 3.1 of Section 3.2 includes the distribution of team votes in the experiment. Consistent with the literature on voting experiments among individuals (e.g., Aimone *et al.*, 2013; Dal Bó *et al.*, 2010), it reveals that some teams do vote to reduce their private returns from shirking. It indicates that 40.9% of teams (= 38/93×100%) voted for the reduction policy. As a result of majority voting, the policy was enacted in 38.7% (= 36/93×100%) of groups in the ENDO treatment. Table 3.1 also shows a clear pattern of selection bias. In the ENDO treatment, the percentage of pro-reduction teams was 69.4% (= 25/36×100%) in groups where the policy was enacted, while the percentage was only 22.8% (= 13/57×100%) in groups where it was not enacted. Hence, pro-reduction teams were overrepresented (underrepresented) in groups where the reduction policy was (was not) enacted in the ENDO treatment. This is a pattern similar to the selection bias discussed in Dal Bó *et al.* (2010) and Dal Bó *et al.* (2019).

In fact, teams' support for the policy was positively correlated with their performance before voting. In the practice phase, teams performed the task for only three minutes under individual-based remuneration. The data indicate that teams which voted for the reduction policy on average answered 1.001 questions correctly in the practice phase; their performance was significantly better at two-sided $p < 0.01$ ($z = 4.230$) than teams which voted against the policy (the average number of correct answers by anti-reduction teams was 0.414). This pattern holds regardless of the institutional outcome, i.e., whether the policy was enacted or not (Appendix B.C, Figure C.1). This means that pro-reduction teams may have characteristics different from anti-reduction teams. As shown in Appendix B.C, Figure C.1, the performance of teams in the EXO treatment was somewhere in the middle of the pro- and anti-reduction teams (was similar to that of anti-reduction teams) in groups where the policy was enacted (was not enacted).

In sum, selection bias must be controlled for when identifying the dividend of democracy in the data. This paper utilizes the method proposed by Dal Bó *et al.* (2019) to remove selection effects. Section 3.3 first discusses the dividend of democracy on work productivity, after which it discusses workers' effort choices in detail and their welfare consequences.

**Result 1**: *40.9% of teams voted for reducing returns from staying in the Game screen. As a result of majority voting, the reduction policy was enacted in 38.7% of groups in the ENDO treatment.*

### *3.3.1. Dividend of Democracy on Work Productivity*

The first key result of this study is the positive effect of democracy on work productivity. The dividend of democracy is quite strong: around 20% on average. Consider, first, groups where the reduction policy was enacted. Productivity, defined as the number of correct answers per minute of teamwork (i.e., per average time spent in the task screen by a team member), is 0.594 in the ENDO treatment. 0.594 means that if a team, i.e., all three members, worked the entire 35 minutes of the task-solving phase without playing Tetris, they would be able to answer on average 20.79 (= 0.594×35) tasks correctly. This productivity is 28.5% larger than the productivity in the EXO treatment, which is calculated as 0.462.[42] Part of the productivity increase can be attributed to selection bias as already discussed. Thus, such bias must be controlled for to isolate the dividend of democracy by adjusting the "weights," i.e., the distribution of votes. This paper follows Dal Bó *et al.* (2019) calculating the re-weighted productivity with the two steps:

Step 1: Calculate (a) the average number of correct answers and (b) the per member average work time, using as weights the percentage of pro-reduction teams in the population (40.9%) rather than the percentages under the reduction regime in the ENDO treatment (69.4%).

Step 2: Calculate (a)/(b).

The re-weighted work productivity in the ENDO treatment found using these steps is still quite large – i.e., 0.529, 14.5% larger than that in the EXO treatment.

Consider, next, groups where the reduction policy was *not* enacted. There is also a strong effect of democracy for these groups. First, the productivity before reweighting was modestly different between the two conditions: 0.488 in the ENDO and 0.431 in the EXO treatment. However, this mild difference is due to selection, in that pro-reduction teams are underrepresented in the ENDO treatment, i.e., these account for only 22.8% of teams (Table 3.1). Productivity after reweighting was large, 0.539, in the ENDO treatment. This means that the dividend of democracy is 0.108 (= 0.539-0.431) correct answers per min. of teamwork, i.e., a 25.1% increase in productivity. The fact that democracy strongly affects behaviour irrespective of the policy implementation outcome suggests that being involved in the democratic process by itself, i.e., democratic culture, affects their work

[42] The average number of correct answers and average per member working/shirking time by institutional condition can be found in Table 3.3.

motivation directly, which is consistent with the idea that democracy directly enhances intrinsic motivations to work (e.g., Deci and Ryan, 1985, 2000).

In sum, the reweighted dividend of democracy without the reduction policy (i.e., 0.539 versus 0.431) was of almost a similar magnitude to the one in groups with the reduction policy (0.529 versus 0.462). This underscores the strong role of democracy in improving productivity. For this reason, the two institutional outcomes (with or without the policy) are pooled to statistically test the significance of the dividend of democracy (Table 3.2).

Table 3.2 reports test results for the dividend of democracy on work productivity using all of the data. In order to calculate each *p*-value, the estimates for the dividends of democracy were calculated 20,000 times based on session-level bootstrapping.[43] Panel A of Figure 3 reports the distributions of estimated dividends of democracy. These reveal that the size and the significance of the dividend of democracy are only slightly affected by the correction of the selection bias. The overall impact is economically large: democracy boosts productivity by 20.02% (= (0.535 – 0.445)/0.445×100%) and it is significant at the 5% level. Hence, it can be concluded that democracy by itself strongly improves productivity.

Readers may also be interested in knowing how the dividend of democracy persists in the workplace. To answer this question, work productivity measures are calculated by splitting the data into quarters of the experiment. It first shows that experience does help to improve workers' problem-solving skills, and hence their per-minute-of-teamwork performance. Panel B of Figure 3.3 indicates that, whether in the ENDO or EXO treatment, work productivity increased from quarter to quarter. The dynamics also reveal that higher work productivity in the ENDO treatment, relative to EXO treatment, was remarkably stable throughout the experiment. This means that fatigue (whether physical or mental) and/or monotony may not weaken the dividend of democracy in the workplace.[44]

**Result 2**: *(a) There is strong evidence that democracy significantly boosted work productivity, defined as the production per minute spent working. (b) The positive dividend of democracy persisted throughout the task-solving phase.*

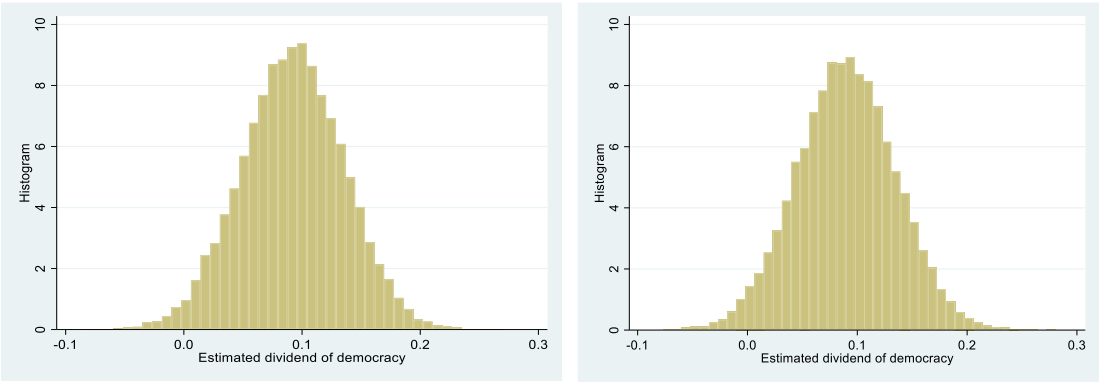
[43] Each estimate was calculated using 62 sessions randomly drawn from the set of the original 62 sessions.
[44] An analysis in Section 3.2 suggests that workers in the ENDO treatment did not accumulate fatigue with a higher work pace, as they instead increased the time spent in the Game screen.

**Table 3.2:** *Dividend of Democracy in Work Productivity*

| | A. Using original weights | B. Using adjusted weights following Dal Bó *et al.* |
|---|---|---|
| **Team production per minute of its three members' working:**[#1] | | |
| (a) ENDO | 0.536 | 0.535 |
| (b) EXO | 0.445 | 0.445 |
| (c) Dividend of Democracy (= (a) – (b)) | 0.091 | 0.090 |
| Two-sided *p* for $H_0$: (a) = (b)[#2] | **0.036**** | **0.046**** |

*Notes*: The overall productivity measures in rows a and b were calculated using the distribution of policy implementation in the EXO treatment (i.e., % of groups with policy: % of groups without policy = 123/273: 150/273). The numbers in column A are productivity measures calculated using the original distributions of voter types under institutional outcomes (pro- or anti-reduction teams) shown in rows i and ii of Table 3.1. The numbers in column B are productivity measures using the distribution of voter types in the population following the weights-based identification strategy proposed by Dal Bó *et al.* (2019). [#1] The number of correct answers per minute of teamwork [#2] The *p*-values were calculated using the bootstrapping procedure described in Dal Bó *et al.* (2019). The number of bootstrap iterations was 20,000 (Figure 3.3).

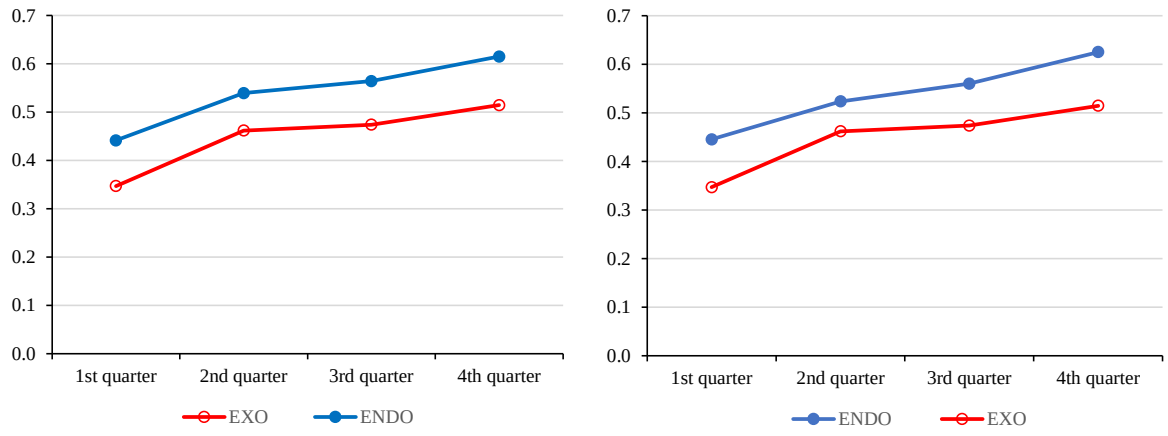
**Figure 3.3**: *Dividends of Democracy for Work Productivity*




(i) When using original weights

(ii) When using adjusted weights according to Dal Bó *et al.* (2019)

A. Distribution of bootstrapped dividends of democracy for productivity based on Dal Bó *et al.* (2019)




(i) When using original weights

(ii) When using adjusted weights according to Dal Bó *et al.* (2019)

B. Dividend of democracy, quarter by quarter

*Notes*: 1. Each distribution in panel A was drawn using 20,000 estimated dividends of democracy based on bootstrap iterations. 2. The productivity measures of each quarter in panel B were calculated by splitting the duration of 35 task-solving phase by four (e.g., the first quarter is the first 35/4 minutes).

While the strong role of democracy is consistent with the findings from prior research on democracy using 'experimental games', such as prisoner's dilemma and public goods games (e.g., Tyran and Feld, 2006; Dal Bó *et al.*, 2010; Sutter *et al.*, 2010; Kamei, 2016), it is at odds with the finding from the 'real-effort' experiment of Kamei and Markussen (forthcoming). In Kamei and Markussen (forthcoming), subjects were assigned to a group of three and then worked individually on either the "counting task" (e.g., Falk *et al.*, 2006; Abeler *et al.*, 2009) or the "addition task" (e.g., Niederle and Vesterlund, 2007; Corgnet *et al.*, 2015) on condition that a revenue-sharing rule is in use and a funny video is available as an alternative activity. Kamei and Markussen (forthcoming) found little evidence of the effects of democratic task selection. The null result was indeed a puzzle which Kamei and Markussen (forthcoming) were not able to explain. A similar null result for the dividend of democracy was also observed and posed as a puzzle in the real effort experiment of Dal Bó, Foster and Kamei (2019) where internet surfing (e.g., Corgnet *et al.*, 2015) was available as an alternative activity. So, why did we get a strong dividend of democracy in the present study? A likely reason is that each team member had stronger shirking opportunities in the present study. Subjects in the present experiment *jointly* solved a *collaborative* counting task as a team in a group, unlike in the prior experiments where subjects *individually* solved an *individual* real effort task in a group. Specifically, while incentives to shirk as a decision-making unit (teams in this study or individuals in the other research) in a group are the same, each team member in the present study has additional

opportunities to shirk by playing Tetris privately, i.e., without notifying their other team members, whose structure features a coordination game inside the team.[45] The difference between the present and the earlier experiments suggests that the dividend of democracy may be more important in an environment where workers have stronger incentives to shirk.

### *3.3.2. Effort Choices and Welfare*

The larger size of work productivity (Result 2) does not mean that democracy improves production in the workplace. Rows I and II of Table 3.3 report the average numbers of attempts and correct answers in the main task-solving phase. The average results are reported by the policy implementation outcome because work behaviours differed substantially by the presence of the reduction policy. It shows that teams attempted more questions and, as a result, answered more questions correctly, in the ENDO than in the EXO treatment (Rows I and II). However, the positive effects of democracy are far from significant (columns 2, 2a and 2b).

This insignificant impact, despite Result 2, was due to the workers' effort choices. As the collaborative counting task was a relatively challenging real effort task, shirking prevailed in the experiment.[46] Workers (although insignificantly) shirked *more* on average in the ENDO treatment than in the EXO treatment – see columns 2, 2a and 2b of Row III. The higher incidence of shirking undermined the positive impact of enhanced work productivity, which resulted in the insignificant effect on the two effort output measures. In sum, this result suggests that a firm needs to have some mechanism to curb workers' effort choices beyond democracy, because their discretion to decide how much to work may partly cancel out the sustained positive dividend of democracy.

One may wonder why democracy worsened shirking. One possible interpretation here is that democracy enlarged workers' motivations to earn a high payoff in the experiment. The subjects may have perceived it to be more payoff-enhancing if they worked harder for a shorter duration and then secured certain gains from staying in the Game screen once exhausted. Although it cannot be verified, this possibility may partly explain the behaviour since, despite Result 2(b), subjects may quickly have feelings of fatigue if their per-minute effort levels rise. Having said that, such a reduction in work time did not work well for the workers, since, while democracy did increase the average payoff, the

[45] A team cannot complete a task while some member is shirking. Such shirking is also interpreted as maliciousness or lack of team spirit towards members who are motivated and are waiting for the shirker's input to find the answer.

[46] The high difficulty in finding answers to the real effort task is a crucial feature of the experiment, which was intentionally designed. Notice that if the tasks were easy, worker subjects would work hard with small output elasticity of incentive changes in this kind of real effort experiment (Corgnet *et al.*, 2015; Erkal *et al.*, 2018). A challenging real effort task and an availability of alternative activities (Tetris) were thus carefully incorporated in the design to make the output elasticity of incentives sufficiently large.

impact is not significant after controlling for selection (Row IV). This implies that their effort choices were not optimal. But, if this conjecture is relevant, why did perceived fatigue play a large part in the behavioural decisions of experiment subjects? A likely possibility is that Result 2 was still not enough to encourage workers to choose putting in a greater effort over shirking. This possibility is quite reasonable as discussed carefully in Section 3.3.

**Result 3**: *Despite Result 2, democracy did not increase team production significantly, because workers under democracy decreased work time to some degree.*

**Table 3.3**: *Work Performance and the Dividend of Democracy*

| | **Un-weighted** | | | **Re-weighted** | | |
|---|---|---|---|---|---|---|
| | All data | With Policy | W/o policy | All data | With Policy | W/o policy |
| | (1) | (1a) | (1b) | (2) | (2a) | (2b) |
| **I. Avg. number of attempts** | | | | | | |
| (a) ENDO | 19.49 | 25.28 | 14.74 | 18.81 | 20.53 | 17.40 |
| (b) EXO | 16.79 | 19.49 | 14.58 | 16.79 | 19.49 | 14.58 |
| $H_0$: (a) = (b)[#1] | 0.151 | **0.043**** | 0.949 | 0.331 | 0.747 | 0.285 |
| **II. Avg. number of correct answers** | | | | | | |
| (a) ENDO | 12.49 | 16.61 | 9.12 | 11.96 | 13.14 | 11.00 |
| (b) EXO | 10.49 | 12.12 | 9.16 | 10.49 | 12.12 | 9.16 |
| $H_0$: (a) = (b)[#1] | 0.170 | **0.060*** | 0.983 | 0.330 | 0.671 | 0.336 |
| **III. Avg per member time spent in the Game screen [min.]**[#1] | | | | | | |
| (a) ENDO | 12.14 | 7.05 | 16.31 | 12.60 | 10.14 | 14.61 |
| (b) EXO | 11.50 | 8.79 | 13.72 | 11.50 | 8.79 | 13.72 |
| $H_0$: (a) = (b)[#1] | 0.664 | 0.345 | 0.236 | 0.534 | 0.594 | 0.711 |
| **IV. Avg. payoff in the main task-solving phase [pound sterling]** | | | | | | |
| (a) ENDO | 9.62 | 11.09 | 8.41 | 9.35 | 9.51 | 9.23 |
| (b) EXO | 8.29 | 8.68 | 7.97 | 8.29 | 8.68 | 7.97 |
| $H_0$: (a) = (b)[#2] | **0.065*** | **0.062*** | 0.555 | 0.138 | 0.498 | 0.150 |

*Notes*: The *p*-values were calculated using the bootstrapping procedure described in Dal Bó *et al.* (2019). The number of bootstrap iterations was 20,000. The numbers in columns 1, 1a and 1b were calculated using the original distributions of voter types under institutional outcomes (pro- or anti-reduction teams) shown in rows i and ii of Table 3.1. The numbers in columns 2, 2a and 2b were calculated using the distribution of voter types in the population following the weights-based identification strategy developed by Dal Bó *et al.* (2019). The overall measures in columns 1 and 2 were calculated using the distribution of policy implementation in the EXO treatment (i.e., % of groups with policy: % of groups without policy = 123/273: 150/273).

### *3.3.3. Privately versus Socially Optimal Behaviours*

This experiment was designed to model a social dilemma problem, i.e., conflicts among teams, in the workplace. Section 3.3 briefly checks the validity of this design setup, finding that its attempt was successful as intended. This section also tries to find an answer as to why democracy was not enough to boost team production in the experiment.

Since staying in the Game screen was remunerated with 16 or 18 pence per minute, it is possible to calculate for what percentage of teams task-solving was a socially or privately optimal strategy (in the sense of material payoff maximization). In order for task-solving to be privately optimal, a team needs to be able to solve at least 0.80 = 16/20 (0.90 = 18/20) tasks correctly per minute when the reduction policy is (is not) in place. A detailed look at the data (Appendix B.C, Table C1) indicates that gaming was a privately optimal choice for almost all teams in the EXO treatment, whether the policy was in place or not. Specifically, it is so for 95.60% of teams (87 out of 91 teams) in the EXO treatment.[47] This implies that the reduction policy was non-deterrent in the experiment. Consistent with the prior experimental evidence on exogenously introduced non-deterrent punishment (e.g., Tyran and Feld, 2006; Kamei, 2016), the effect of the reduction policy was not large in the EXO treatment. Specifically, while the average number of correct answers in the EXO treatment was larger with than without the reduction policy (12.12 versus 9.16), the difference was not significant at two-sided $p = 0.109$ according to the bootstrap method used in the other tests of the paper (the difference is significant but only at the 10% level, i.e., $p = 0.0707$ if a two-sided Mann-Whitney test is used).[48]

However, as intended, the socially optimal strategy was task-solving for many teams. In order for task-solving to be socially optimal, a team needs to be able to solve at least $0.227 \approx 16/60$ [0.30 = 18/60] tasks correctly per minute when the reduction policy is [is not] in place. Overall, the social optimal condition was met for 61.6% of teams (56 out of 91 teams) in the EXO treatment. Notice that task-solving is never privately optimal for teams whose task-solving is not socially optimal.

---

[47] Material incentives did matter for workers' effort choices. In the EXO treatment, the four teams for which task-solving was privately optimal worked on counting on average 31.80 minutes, which is significantly larger at two-sided $p = 0.0015$ than the average work time by the other 87 teams where gaming was privately optimal (which was 23.12 minutes) – see Appendix B.C, Table C1.

[48] The effect of the reduction policy was apparently strong in the Endo treatment (see Table 3.3 for the numbers). The average number of correct answers in the Endo treatment was significantly larger with than without the reduction policy at two-sided $p = 0.001$*** (0.0020***) according to the bootstrap method (a Mann-Whitney test). However, this strong effect is just due to selection. The difference was not significant at two-sided $p = 0.388$ when using the bootstrap method with the distribution of votes in the population being the weights following Dal Bó *et al*. (2019). Recall that democracy enhanced work productivity in the experiment similarly regardless of whether the policy was imposed or not (Result 2), whose aspect makes the effect of the policy in itself small.

Consistent with this incentive pattern, teams whose task-solving was not socially optimal spent significantly less time working on the task than the other teams at two-sided $p < 0.001$ (15.29 versus 28.63 minutes in the EXO treatment). The average number of correct answers per minute of working by the former was only 0.07, but that by the latter was 0.54 in the EXO treatment.

In sum, the present experiment can be thought of as exploring workers' sacrifice and effort choice decisions under social dilemmas in the workplace when the choice was a non-deterrent reduction policy.

Then, one may ask whether democracy might have altered the social dilemma situation to another one (e.g., coordination game), as arguably democracy not only enhances work productivity (Section 3.1), but also reduces effort costs in task-solving. Another look at the data, however, shows that the answer is negative. Specifically, a calculation finds that gaming was a privately optimal choice for almost all teams in the ENDO treatment, i.e., 91.40% of teams (85 out of 93 teams); and task-solving was a socially optimal choice for 61.3% (57 out of 93 teams) in that treatment – see again Appendix B.C, Table C1. These numbers are quite similar to those in the EXO treatment already discussed.

The reason why worker behaviour was characterized by Results 2 and 3 is explained by the theoretical analysis summarized in Appendix B.B. The model there assumes that, following the prior research findings, democracy eases a worker's effort cost, and it also boosts their productivity (its positive effect on work productivity is a parameter $\mu$ in that model). $\mu > 0$ was confirmed by the experiment data as summarized in Result 2. The team's optimal effort provision can then be determined by the relative strength between (a) work productivity [$s + \mu$ in the theoretical model, where $s$ is the marginal return of effort provision by team $i$] and (b) the material incentives to shirk by staying in the Game screen. Theoretically, the positive value of $\mu$ (Result 2) possibly changes the materially beneficial choice from gaming to task-solving – see Appendix B.B, Figure B.2. However, the analysis in the Appendix indicates that if the impact on work productivity is not *economically* large enough, gaming is still the most materially beneficial activity even when teams have a statistically significant dividend of democracy. This is exactly what the above calculations on privately versus socially optimal choices in the experiment data demonstrate. The calculations clearly reveal that democracy did not change the underlying private incentives in the experiment. This means that additional mechanisms on top of democracy would be required to change the incentive structure so that task-solving becomes a privately optimal choice for workers.

## 3.4. Understanding Sacrifice Behaviour: Communication Contents

While the decision data not only uncovered some subjects' preferences to sacrifice but also detected a significant dividend of democracy on work productivity (Section 3.3), it is still unclear what drove such behavioural patterns. Communication contents obtained in the experiment may provide some insight on this question.

Two independent coders were hired to read and classify the communication contents based on their judgment of the subjects' motives. Specifically, a list of codes was designed by the authors, based on the theoretical predictions of the setup and related literature, that could potentially reflect a subject or teams reasoning and/or behaviour. The list was given to the coders to assign whichever codes (including none) they deemed relevant to a given communication log. The coding procedure follows Kamei and Tabero (2022) which utilized the standard coding approach in economics to analyze teams' behavioural reasoning in the context of institutional choices based on intra-team communication logs. The detail of the coding procedure and the full lists of codes used for the present paper can be found in Appendix B.D, sections D.1 and D.2.

The agreement rates and Cohen's Kappa values (Cohen, 1960) can be used to judge the consistency of the coding process between the two coders. Overall, the agreement rates (Kappa values) between the two coders were 96.9% (0.87) and 94.8% (0.78) in the ENDO and EXO treatments, respectively. The Kappa values are at least 0.4 for 92.5% and 78.0% of individual codes in the ENDO and EXO treatments, respectively (Appendix B.D, section D.3). As a Kappa value of 0.4 is usually used as a threshold for a researcher to judge the reliability of coding, we use only the codes that exceed this boundary in this analysis.

Table 3.4 summarizes the list of codes that are found to have impacted the units' voting significantly at least at the 10% level. Their voting is clearly linked to their intention regarding what to do during the main task-solving phase (Code Bs): while units supported the reduction policy if they planned to focus on task-solving, they opposed it if they were considering using the game screen. The coding category linked to pro/anti-policy reasoning (Code Cs) reveals clear motives behind the policy preferences. While the policy is non-deterrent, those who voted in favor of it did so to deter others from shirking (Code C1). On the other hand, those who intended to game or believed that the policy was too weak to alter shirking opposed its enaction. Lastly, unsurprisingly, their views on materially beneficial behaviour and team atmosphere influenced voting. Specifically, units that believed their privately-optimal behaviour was task-solving supported reducing the return from gaming. By contrast, units who experienced discomfort or poor performance from task-solving in the practice

phase opposed such a reduction. While teams with a positive atmosphere (E2) supported the reduction policy, those with poor or lacking communication opposed it (E5).

**Result 4**: *The units that planned to exclusively work on task-solving, believed that the reduction policy would deter others from shirking, or those that had supportive team atmospheres, voted for the reduction policy. However, those who previously experienced discomfort or poor performance from working, considered (even only potentially) using the Game screen, believed that the policy was too weak to alter peers' shirking, or had poor communication with their teammates, voted against the reduction policy.*

**Table 3.4:** *Significant Code Meanings and Its Impact on Voting for the Reduction Policy*

| Code | Meaning | Direction |
|---|---|---|
| B1 | Agree/Imply to count as primary behavior | (+)*** |
| B2 | Agree/Imply to game as primary behavior | (-)* |
| B3 | Agree to hybrid behavior e.g. so many tasks/minutes before switching to the game screen | (-)** |
| B4 | Agree to discuss, decide and/or alter behavior during the counting task later (35-minute phase) based on performance/needs in Phase 2 | (-)** |
| C1 | Pro-policy to deter others from switching to the game screen by reducing the return (monetary deterrence) | (+)*** |
| C8 | Anti-policy as they intend to game for at least some of the task-solving period | (-)*** |
| C11 | Express that the policy is not strong enough to deter others from switching to the game screen (monetary) | (-)** |
| D2 | Believe they as a team make the most money from counting | (+)*** |
| D5 | Discuss their performance or comfort in Phase 1 (weak/negative) | (-)** |
| E2 | Positivity towards teammates e.g. attempts to encourage others or being supportive | (+)* |
| E5 | No communication from just 1 or 2 team members | (-)*** |

*Notes*: +(-) in "Direction" means the reasoning en(dis)courages voting for the policy. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

As summarized in Result 4, units' commitment to task-solving and their intention to affect others' shirking were the drivers behind their votes in favor of the reduction policy. To explore how policy implementation outcomes affected units' behaviours, coding analyses were further performed using the communication logs of the 35-minutes task-solving phases (Table 3.5). Three similar tendencies were observed for both the ENDO and EXO treatments. First, those who reacted negatively to the implementation outcome tended to work less (F1, I1). Such negative reciprocal tendencies were unsurprising considering the large findings of other-regarding preferences– see, e.g., Sobel (2005) and Fehr and Schmidt (2006). Second, a units' plan to work on counting or engage in

gaming affects performance (F3, F4, F5, F6, I4, I5, I6), similar to Result 4. Third, units' positive and negative experiences of task-solving, respectively, improve and hurt performance (G4, G5, D4, D5).

The results reveal *signaling* effects of voting on task-solving, and some nuanced evidence about the teams' dividends of democracy seen in Result 2. First, units that considered the distribution of votes to predict others' task-solving or discussed changing behaviour worked longer (F9, F15). Second, units who believed that other teams would complete tasks following the vote performed strongly (F7), resonating with the idea that voting has a signaling value, thereby encouraging collaboration. Further, even units who thought that others would not respond to the reduction policy improved their performance (F8), which implies democracy directly affects behaviour beyond signaling. Nevertheless, its effects are cancelled out if an anti-policy team is present in a group and units have a negative view on the task-solving behaviour of the anti-policy team (F13).

**Table 3.5:** *Reasoning behind Work Choice and Productivity*

| Code | Meaning | Direction |
|---|---|---|
| | **Codes related to reactions to vote outcome in ENDO (Code Fs) or policy outcome in EXO (Code Is)** | |
| [ENDO treatment:] | | |
| F1 | Express negative emotions (e.g., upset, anger) about the outcome of the vote | (wt-)*** |
| F3 | Agree/Imply to count as primary behavior | (wt+)***, (p+)*** |
| F4 | Agree/Imply to game as primary behavior | (wt-)***, (p-)*** |
| F5 | Agree to hybrid behavior e.g. so many tasks/minutes before switching to the game screen | (wt-)**, (p-)* |
| F6 | Agree to discuss, decide and/or alter behavior during the counting task later (35-minute phase) based on performance/needs in Phase 2 | (p-)* |
| F7 | Express belief/hope that other teams will complete tasks following the vote | (wt+)***, (p+)** |
| F8 | Express belief that teams will not complete tasks following the vote | (wt+)***, (p+)*** |
| F9 | Discuss the distribution of votes and predict how each team may respond to one another | (wt+)*** |
| F13 | Belief on other teams' responses: anti-policy teams will work little | (p-)*** |
| F15 | Discuss whether to change behavior based on the vote outcome | (wt+)*** |
| [EXO treatment:] | | |
| I1 | Express negative emotions (e.g., upset, anger) about the policy outcome | (wt-)***, (p-)*** |
| I4 | Agree/Imply to game as primary behavior | (wt-)***, (p-)*** |
| I5 | Agree to hybrid behavior e.g. so many tasks/minutes before switching to the game screen | (wt-)***, (p-)** |
| I6 | Agree to discuss, decide and/or alter behavior during the counting task later (35-minute phase) based on performance/needs in Phase 2 | (wt-)* |
| | **Other Codes [The same codes were used for the ENDO and EXO treatments]** | |
| G4 | Expression of strong negative emotion e.g. frustration, anger, disappointment | (p-***, Exo) |
| G5 | Expression of strong positive emotion e.g. enjoyment, things are going well | (wt+***, p+** , Exo) |
| D4 | Discuss their performance or comfort in Phase 1 and/or so far in Phase 2 (strong/positive) | (wt+***, p+***, Endo) |

| | | |
|---|---|---|
| D5 | Discuss their performance or comfort in Phase 1 and/or so far in Phase 2 (weak/negative) | (wt-***, p-***, Endo), (wt-**, Exo) |
| D8 | Discuss uncertainty surrounding other teams' work choices or abilities | (wt-***, p-**, Exo) |
| D9 | Suggest distrust of other teams e.g. expect them to take advantage | (wt-***, p-***, Endo) |

*Notes*: wt and p in the "Direction" column indicates two work performance measures: work time (minutes) and productivity defined as the number of correct answers divided by the work time. +(-) means the reasoning in(de)creases the performance measures. All significant codes are listed for Code Fs and Is, while only some key codes are included for the other coding categories to conserve space (Appendix B.D4 includes the full estimation results). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

## 3.5. Conclusion

Teams are popular decision-making and work units in organizations, however they feature a complex coordination problem. Overcoming moral hazard among teams in the workplace plays a crucial role in maintaining productivity in the firm, whether in the traditional work environment or in a remote working setting, such as that triggered for many by the Covid-19 crisis. The present paper investigated how frequently groups reduce the return from shirking by enacting a formal non-deterrent sanction policy, and how such endogenous choices per se improve work productivity. To achieve this, a novel real effort experiment was designed, equipped with (a) a collaborative counting task and (b) gaming (Tetris) as a real leisure activity. The experiment results showed that around 40% of teams voted to reduce the return from staying in the Game screen. A contents analysis using teams' communication logs showed that such sacrificial behaviour was driven by not only their commitment to work on counting but also their belief that the reduction policy would deter others from shirking.

The decision data also uncovered a significant and strong dividend of democracy on work productivity. Strikingly, whether the policy was enacted or not, teams in the ENDO treatment displayed significantly higher per-work-time production than those in the EXO treatment. Thus, democratic culture directly affects behaviour positively. However, the workers under democracy also experienced *higher* levels of shirking, i.e., the time spent on the Game screen was larger in the ENDO than in the EXO treatment, presumably driven by their enhanced fatigue due to the more intensive working in the former. This implies that while additional mechanisms that affect incentives besides democracy may be required to increase production, democracy may improve efficiency. What kinds of mechanisms would work best to instead increase production further remains for future research. Having said that, it should be emphasized here that the average production of the workers under democracy did not decrease (it increased, although insignificantly) thanks to their strong per-work-time production.

The findings on the positive dividend of democracy on work productivity have a policy implication for effective human resource and management practices. While prior research suggests that innovative human resource management involving worker participation (such as that in production sites) lead to better work performance (e.g., Ichniowski *et al.*, 1997), it is unclear how democracy affects behaviour, as earlier real effort experiments failed to detect strong dividends of democracy (e.g., Dal Bó *et al.*, 2019; Kamei and Markussen, forthcoming). Using an environment with strong shirking incentives, the present experiment suggests that organizations with a shared goal can benefit from introducing participatory decision-making with their employees or group members, by potentially improving their work productivity. Even when democracy induces the workers to work less, the improvement to productivity allows for achieving a production goal with fewer working hours.[49] This boost to productivity is achieved through signaling effects; workplace democracy provides the workers the ability to indicate their intentions or desire to cooperate with each other through democratic procedures such as voting, and the recipients can then respond to these signals. Such a social exchange may be fundamental for workers to achieve collaboration through reducing uncertainty surrounding each other's behaviour in a democratic workplace environment. Given this, the firm may create a positive and collaborative atmosphere and improve productivity by designing democratic systems (encouraging signaling effects) in multiple layers and activities across the organization.

[49] There is a trend to transform the traditional workplace to an employee-centered workplace in many countries. For example, in the United Kingdom, some firms recently tested four-day work weeks to make working conditions flexible to meet the different needs of employees. Having higher work productivity in a democratic environment certainly helps the firms achieve the same or even potentially better outcomes with fewer working hours.

# 4. Cooperation for Accountability: Civic Engagement as a Second-Order Public Good

## 4.1 Introduction

Centralised institutions exist to overcome the public good problem faced in our everyday lives, whether gathering and redistributing resources, providing critical public services, or upholding laws and rights. Such institutions are necessary as, without them, contributions for these goods would be far too small, and the incentive to free-ride on others' efforts and contributions much too great. However, it is often taken for granted, in the experimental literature at least, that these institutions operate perfectly when contributions are substantial. Public officials are given power over individuals to extract contributions and hold them accountable if unpaid, yet without being held accountable themselves, they may misuse public funds through corruption, negligence, or inefficiency.

Unsurprisingly, misuse or loss of public funds is detrimental to citizens, with state corruption in particular linked to a negative impact on economic growth (see Campos et al., 2004, and Aidt, 2009 for surveys). A more recent study by Gründler and Potrafke (2019) found that an increase of one standard deviation in the Perception of Corruption Index decreased real per capita GDP by around 17% in the long-run. Specific examples are also readily available, such as the misuse and embezzlement of public funds by numerous governments during the Covid-19 pandemic, especially with regard to public procurement of medical supplies (Teremetskyi et al., 2020). Given that citizens are ordinarily unable to reduce or refuse to pay their taxes if they disapprove of their government's actions or ability, how can they ensure that their contributions are not wasted and that critical services are provided?

One mechanism is democracy, which holds the state accountable through the threat of removal from power, however, it requires active participation to function. For example, a poor government official or party will not be voted out if too few are motivated to learn of or react to their performance. Noting that being democratic in itself is not sufficient to hold the state accountable, Mungiu-Pippidi (2013) explores this relationship and finds that control of corruption is strongly and significantly related to normative constraints including the number of civic society associations, freedom of the press, and internet connection; proxies used to gauge the accessibility and level of civic engagement. Yet, the actions required to hold a government accountable, including keeping up with the news and research, attending committees, protests or rallies, creating and sustaining activist groups, and even voting, are all privately costly and needed on a large scale to have an impact. Given

this, citizens face a dilemma between voluntarily engaging with politics in the hope that others do the same to improve the use of public funds, or spending their time and effort on private interests which may mitigate some of the loss generated by an imperfect government on a personal level. As a result, a second-order public goods dilemma exists by which costly civic engagement is required to counter issues associated with unaccountability, such as corruption or inefficiency, and to maintain the systems used to solve the first-order problem of below-optimal levels of contribution. However, unlike the first-order problem, the second-order problem cannot be solved by state enforcement.[5051]

This paper utilises a two-stage public goods game (PGG) to examine whether individuals will put in the effort to leverage the benefits of civic engagement for constraining official malfeasance and negligence. Specifically, in each period of the experiment subjects participate in two stages, a main stage and a pre-stage. In the main stage, they play a PGG with a minimum contribution to the public sector, reflective of a state with an enforced tax system. However, emulating poor governance, the amount contributed to the public sector may be reduced before it is used to calculate subject payoffs. The amount that the public sector revenue is reduced by is directly related to subject behaviour in the pre-stage, where subjects may choose to complete real effort takes which either add directly to their payoff, or which reduce the inefficiency of the institution in the following stage. To reflect the issues of an unaccountable state, such as corruption or negligence, and the role that civic engagement plays in holding the state accountable, the benefits of contributions to both the public sector and private activity vary with the level of civic tasks completed in the previous stage.

This is not the first study to explore such a two-stage stylised mechanism, with the most comparable being Kamei et al., (2019), from which the present study is based. While both studies explore the leveraging of civic engagement, the present study explores the relationship with state inefficiency, while the prior study focussed on the probabilistic establishment of a fully accountable state. These design choices have different implications for behaviour, which are discussed in more detail in section 4.2 below. The benefits of establishing a centralised minimum contribution, or state, compared to the anticipated complete free-riding without one are large and clear to individuals,

[50] Specific aspects of civic engagement can be institutionalised or mandated by the state, such as compulsory voting or civic service, however engagement and effort by the citizen to make these effective are still provided on a voluntary basis. For example, Jakee and Sun (2006) show theoretically that compulsory voting when participants are not engaged or are ill-informed increases incidences of random voting, which may then undermine rather than strengthen state accountability. Similarly, centralised solutions may be self-enforced by civic groups, however individual voluntary effort and engagement are still required for these to be sustainable.
[51] There are some centralized checks in place to prevent corruption and negligence, however these are prone to the same issues of the state owing to their operators being in positions of power or even direct accountability to the state and its officials. Given this, even these institutions and their officials require discipline through some level of civic accountability.

making civic engagement very attractive in the prior study. This is useful to demonstrate the effectiveness of leveraging relatively fewer resources in one stage to overcome a larger social dilemma in the second, which Kamei et al., (2019) term the *leverage effect*, but does not speak for its presence when the marginal benefits of completing more civic tasks are known and privately suboptimal, as in this study.

Despite the strong incentives for individuals to free-ride, we find the persistence of the leverage effect in this experiment in the form of high levels of civic engagement. While there is a downward trend in the amount of civic engagement in the latter half of the experiment, it is remarkably small, especially when compared to standard social dilemma cooperation trends. We further find that a cost- and ramification-free reputation and feedback mechanism is surprisingly effective in encouraging and maintaining the level of civic engagement, which we believe to operate through social-image concerns and group norms. This paper then distinguishes itself from other PGG literature by specifically modelling the problem of the unaccountable state and the second-order social dilemma citizens need to overcome to correct for it. By doing so, we demonstrate experimentally how society is able to reduce corruption or inefficiency at the state level through voluntary civic engagement.

The paper is structured as follows; Section 4.2 explores the related literature and Section 4.3 introduces the experimental design, treatment variations and predictions. Section 4.4 presents the results of the analysis, and Section 4.5 summarises and concludes.

## 4.2 Related Literature

Public goods games are used to experimentally explore the provision of a public good; here, participants have the incentive to allocate their resources to a private account while free-riding on the contributions of others to a shared account. If participants cooperate sufficiently then the public good may be well-funded, however, a large body of experimental literature finds that when mechanisms to encourage cooperation are unavailable, cooperation towards sustaining a public good expectedly declines in favour of free-riding (see Ledyard, 1995; Chaudhuri, 2011). Several mechanisms have been suggested to counter this declining trend, including centralized sanctioning systems (Falkinger *et al.*, 2000, Alm, 2018), informal peer-to-peer punishment (Fehr and Gächter, 2000, 2002), endogenous group formation (Page, Putterman and Unel, 2005), cheap-talk (Isaac and Walker, 1988), among others.

Centralized sanction mechanisms have clearly demonstrated their usefulness in increasing cooperation in the laboratory. Falkinger *et al.*, (2000) found that when available, a system that fined contributions below the group mean and rewarded contributions above it saw increasing and sustained levels of contribution over time. Similarly, experiments that set a minimum contribution (see Andreoni, 1993; Keser et al., 2017 for examples) unsurprisingly see higher contributions than when no minimum contribution is required. The effectiveness of these systems, both empirically and in the laboratory, does vary with the perception of the chance of audit, the strength of fine for non-compliance, as well as other aspects such as social norms and information (Alm, 2018), but if auditing is certain and the fine is set to a deterrent level, it can be made privately optimal for even self-interested players to fully comply. It is also notable that the centralized system is the most closely related to that seen in the real world for large-scale public good provision, as governments implement (and punish the evasion of) a taxation system as a key source of state revenue.

However, these experimental setups do not address an important concern surrounding real-world centralized systems, their efficiency. As discussed in the introduction, such public or shared accounts are subject to negligence, corruption, and inefficiency, meaning that often much of what is contributed to the system is lost before the public good can be fully provided. Not only is this costly to society's welfare, but it can risk discouraging contributions in the future. Specifically relating to corruption, Cagala *et al.*, (2017) found that when a corrupt official was present and able to extract 10% of the contributions to a shared public good account, contributions decreased significantly. Similarly, a field experiment in Liberia found that exposure to greater corruption levels by chiefs reduced voluntary contributions to local public goods (Beekman *et al.*, 2014). Further, Campos-Vazquez and Mejia (2016) found that exposure to higher corruption in a pre-game led to lower contributions in a PGG, even when the punishment of lower contributions and counter-punishment was possible.

Of course, in the real world, it is not always possible, and often illegal in the case of taxation, to reduce one's contributions towards a system regardless as to whether it is corrupt or poorly managed. It is also not an ideal response, as it reduces the centralized system's ability to overcome the public good dilemma. How then can public goods be provided in a system that is prone to wastage or corruption?

Our experiment utilizes a two-stage design in which participants may expend effort in the pre-stage, as a proxy for civic engagement, to reduce inefficiency in the main stage featuring a PGG with a centralized sanction mechanism. While both stages are social dilemmas, the pre-stage forms a mechanism that allows participants to leverage a smaller amount of resources in order to overcome

the larger second-order dilemma. This 'leverage effect' is similarly proposed by Kamei *et al.,* (2019) to overcome the issue of introducing an accountable state (a sanction to enforce a minimum contribution). In their experiment, subjects completed either private or civic tasks in order to increase the probability that a minimum contribution would be imposed in the main stage, with task completion above a certain amount guaranteeing the mechanism was implemented. They found that the level of civic engagement was surprisingly high and stable compared to the downward trend in cooperation ordinarily observed in PGG without mechanisms.

However, the contrast in payoff between having a minimum contribution or not is stark as, like the literature discussed above, Kamei *et al.,* (2019) model that the system is implemented effectively. The experiment setup is then comparable to a choice between having no government or having a perfect government, which is useful in demonstrating the leverage effect but perhaps unrealistic. The present study builds on elements of their design by exploring its application when civic engagement is instead used to improve the accountability of a government, and so making it more efficient. While civic engagement in Kamei *et al.*, (2019) increased the probability of installing a perfect government, here it reduces the amount of money in the shared account which is lost as a result of poor governance. As a minimum contribution exists regardless of the level of loss, contributing to a negligent government is unavoidable, however, it remains privately optimal to free ride on the civic engagement others provide for improving governance.

Similar to Kamei *et al.*, (2019) we also allow for reputation and feedback in half of our treatments. The availability of reputation-building and feedback is anticipated to increase cooperation. Focusing on reputation, van Vugt and Hardy (2009) found that when contribution decisions were made public to their group, subjects increased their contribution even when the provision of the public good was no longer possible or had already been provided. They attribute this costly signalling to reputation or social image concerns. Similarly, costless disapproval was found to increase contributions in an experiment by Masclet *et al.*, (2003), despite having no monetary ramifications for a subject. They found that costless disapproval was even more effective under partner-matching, demonstrating the impact of repeated and visible interactions. Kamei *et al.,* (2019) also find increased civic engagement when smaller social circles and feedback are available. While they made use of a review system on a scale of one to five, we simplify the feedback mechanism to a binary choice of assigning positive feedback (a smiley face) or not. This more closely replicates how feedback is given for civic engagement or social actions outside of the laboratory. For example, scale rating systems are more commonly used for consumer products, while on social media feedback is given by 'liking' or assigning a visual reaction to a message or action. Further, in our feedback treatments, subjects only

learn the efforts of their fellow social circle members after a round is completed, while in the previous paper they could choose to inform others of their efforts during the task-solving stage. Removing this notification option enables us to isolate the motivational effect of positive feedback from that of in-task motivation from knowing others are completing tasks. By keeping a relatively simple feedback mechanism in half of the treatments while maintaining anonymity in the others, we can separate the impact of intrinsic and extrinsic motivation in civic engagement.

The second variation in the present study is that of framing. The way in which a PGG experiment is posed to subjects is known to influence decision-making. For example, whether a decision is framed as giving or taking from a public good can influence contributions (see Andreoni, 1995; Park, 2000; Cox and Stoddard, 2015; Fosgaard *et al*., 2019), as can whether contributions promote a public good or prevent a public bad (Sonnemans *et al.*, 1998; Iturbe-Ormaetxe *et al.*, 2011), or if they are framed in terms of having positive or negative outcomes and/or externalities (Bohm and Theelan, 2016). While this experiment is framed in terms of preventing a public bad (loss), this is held consistent across treatments. The treatment variation in framing is the terminology used to contextualise the experiment, or a form of *label framing*.

Experimental studies have found that the context of a PGG can influence contributions, although the outcomes are mixed. For example, Regge and Telle (2004) found a slight positive impact on contributions of framing the shared account as a 'community box' rather than just a box. By contrast, Dufwenberg *et al*. (2011) found that this effect was culture-sensitive and improved contributions in a community-spirited University population while reducing contributions in a larger, less community-spirited University, where 'community' may have had other connotations. They posit that the impact on contributions works through the beliefs that a given frame creates. In some populations, 'community' has a positive or cooperative connotation and so subjects believe that other participants are more likely to cooperate, whereas in places where 'community' has individualistic connotations the opposite is true. This is supported in a study using 'teamwork' or 'paying taxes' frames by Eriksson and Strimling (2023), who found those who played under the teamwork-framed treatments, or who spontaneously associated a PGG with teamwork when frames were neutral, believed others would contribute more and did so themselves as a result.

The framing used in the present study differs to that in the above studies as it is not trying to stimulate specific team or community thinking, rather contextualise the experiment to the real-world setting which it emulates. Several studies have done similarly, for example, Milinski *et al.*, (2007) used preventing climate change as a motivation for contributing to a public good, and Kamei *et al.*, (2019) used political terminology to encourage the perception of the experiment as a simulation of

society and government. Nevertheless, as the framing may impact behaviour as discussed above, without including an unframed variation it would be impossible to distinguish any effects on cooperation that resulted from the two-stage design from that of framing effects. Further, having both politically and neutrally framed treatments allows for comparison of cooperation behaviour between other neutrally framed PGG experiments and the politically framed but more structurally comparable Kamei *et al.*, (2019) paper.

## 4.3 Experimental Design

### *4.3.1 Experimental Design and Procedures*

The base of the experiment is a first-order PGG in which subjects must allocate their endowment between a public sector and a private activity. Unlike the standard game, allocations to the public sector not only benefit subjects directly, and equally, but also increase earnings from tokens allocated to the private activity. This design reflects that private activity, whether at work or in life more generally, is supported or enhanced by the public sector through the provision of law and order, welfare, infrastructure, and so on. However, as with the standard PGG, the individual has a private incentive to allocate their resources to their private activity only, resulting in sub-optimal funding of the public sector. In this experiment, we introduce a minimum allocation that sanctions subjects for each token less than the required amount that they allocate to the public sector; this minimum required allocation and sanction mechanism may be interpreted as a tax system implemented by the state or other institutions to ensure critical services are provided. While this element is not novel (Kamei *et al.*, 2019), we do not assume that such an institution is perfect. Specifically, we model that the institution's efficiency is a direct function of the effort of its subjects, which is captured in real effort tasks. Subjects have a choice between two types of tasks, one that improves their earnings directly and a second that decreases the reduction to the public sector in the first-order problem. Whilst it is privately optimal to complete only the first type of task, completing the latter type is socially optimal, forming a second-order social dilemma problem.

To examine this, we use a 2 x 2 between-subjects design. The first treatment variation is framing; treatments feature either neutral or political framing, with the latter designed to encourage individuals to view the experiment as a proxy for a political setting. For example, the politically framed treatments use words such as "government", "public sector", and "civic task", while the neutrally framed treatments use "group", "group account", and "Type A Task". The second variation is reputation and feedback; in the treatments without reputation, subjects' actions are anonymous, but in the treatments with reputation, subjects are given a player identity which is both associated with

their actions and subject to feedback from other group members. The treatments are summarised in Table 4.1, where neutral framing is indicated by the prefix "Neutral", political framing by the prefix "Political", and the ability to use the feedback mechanism is indicated by either "No Feedback", or "With Feedback".

**Table 4.1**: *Summary of Treatments*

| | | Framing | |
|---|---|---|---|
| | | Neutral | Political |
| Feedback | No | Neutral-No Feedback | Political-No Feedback |
| | Yes | Neutral-With Feedback | Political-With Feedback |

The Main Stage

In each of the four treatments, subjects participate in a first-order public goods problem. First, they experience 4 periods without any institution to enforce allocation to the public sector, this is called Part 1, followed by 15 periods with the institution in place. As such, we will first explain Part 1, and then build on it to develop Part 2. In every period, each of the 12 individual subjects, *i,* is given 20 "tokens" to allocate to their private activity or to the public sector.[52] Tokens are allocated to earn "points" which determine a subject's payment at the end of the experiment, where 260 points converts to £1. Each period is independent in that tokens cannot be transferred across periods and must be fully allocated in each period. Tokens allocated to the private activity are called $b_i$, while tokens allocated to the public sector are called $p_i$. The sum of allocations to the public sector by all 12 subjects is $P = \sum_{all\ j} p_j$ (where j includes i). An individual's earnings, in points, can be described by:

$$Y_i\ (p_i\ |p_{-i}\ ) = b_i V(P) + D(P) \qquad (1)$$

where $p_{-i} = \sum_{j \neq i} p_j$ , $V(P)$ is the productivity of allocations to the private activity, $b_i$, and $D(P)$ is the per-person direct benefit from the public sector. As such, both earnings from one's private activity and the direct benefit from the public sector depend on how well-funded the public sector is. This design reflects that private activity is supported by a functioning state, and so underfunding of the state not only impacts social welfare projects but also the profitability of individual entrepreneurship or work. These elements are captured in the V(P) and D(P) functions below. First:

[52] 12 subjects are used so that each subject has a relatively small impact on the group's behaviour as a collective, as in society, and to allow for smaller subgroups within the group in the treatments with feedback.

$$D(P) = \frac{101}{1+(49)\mathrm{Exp}[-0.04*\mathrm{P}]} - 2. \quad (2)$$

The shape of D(P) reflects that many public projects require a considerable level of funding before they can be implemented effectively, while returns start to diminish once the service has become well-funded. V(P) is described by:

$V(\mathrm{P}) = \alpha + \beta\mathrm{P}$ for $\mathrm{P} \leq \mathrm{P}^*$, specifically for $\mathrm{P} \leq 96$:

$$\mathrm{V(P)} = 6 + (1/8)\mathrm{P}$$

$\mathrm{V(P)} = \alpha + \beta\mathrm{P}^*$ for $\mathrm{P} > \mathrm{P}^*$, specifically for $\mathrm{P} > 96$:

$$\mathrm{V(P)} = 6 + (1/8)96 = 18 \quad (3)$$

where 96 tokens is set to be the optimal funding level for the public sector, P*=96, or an average of 8 tokens per subject. The shape of V(P) reflects how funding to the public sector is beneficial to the private sector up to a point, for example, to ensure contract enforcement and infrastructure, but beyond the optimal level it offers no additional return.

Note, subjects did not need to be able to calculate these functions as they were provided with a table showing their payoffs given their own allocations and that of others between the two accounts (the outcome of equation (1)), as well as graphs of each function to demonstrate how they changed with P (see the instructions in Appendix C.B). While these functions are more complex than those of standard PGG, to reflect the context they represent more accurately and to be comparable to those in Kamei *et al.*, (2019), both are designed to preserve the fundamental dilemma of a PGG in that it is always privately optimal to allocate as few tokens as possible (0 tokens in Part 1) to the public sector while free-riding on the contributions of others.

In Part 2, where subjects play a further 15 periods, a minimum allocation to the public sector is introduced of 8 tokens. For each token less than 8 that subjects allocate to the public sector, they are fined 35 points. The fine is set so that it is never privately optimal to contribute less than 8 tokens in Part 2, making 8 the new privately optimal allocation to the public sector, and this is highlighted to subjects in the instructions. In Part 2, along with the minimum allocation, we also introduce a reduction mechanism that impacts the public sector. This mechanism is designed to replicate an imperfect state or institution that, through corruption, negligence, or inefficiency, reduces the benefits of the public sector. In the experiment, the number of tokens allocated to the public sector may be reduced by a given percentage before payoffs from the public sector and private activity are

calculated. Specifically, tokens allocated to the public sector are halved, but this reduction may be mitigated depending on performance in the pre-stage.[53]

The Pre-Stage

In each period of Part 2, before the main allocation decision (main stage), subjects are given 40 seconds to complete real-effort tasks. There are two types of real-effort tasks, those that reduce the percentage reduction in the main stage, and those which benefit the subject directly. The tasks are identical but identified differently depending on the treatment; here we will refer to the former as 'civic tasks' and the latter as 'private tasks' as in the politically framed treatments. In the task, subjects are given a description of a fictional person's preferences in two topics, such as travel destination and transport, and must drag and drop an icon of the person onto the respective area of a 2x2 axis (shown in the instructions).[54] Each task takes approximately 5-10 seconds to complete, and they may complete as many tasks as they wish in the 40 seconds. For each private task they complete correctly, they receive 10 points. By contrast, civic tasks do not offer a direct benefit to the individual; at the end of the pre-stage, the total number of civic tasks is summed (henceforth referred to as TTA) and used to determine the number of tokens removed from the public sector in the following main stage. The percentage reduction, %R, is calculated as:

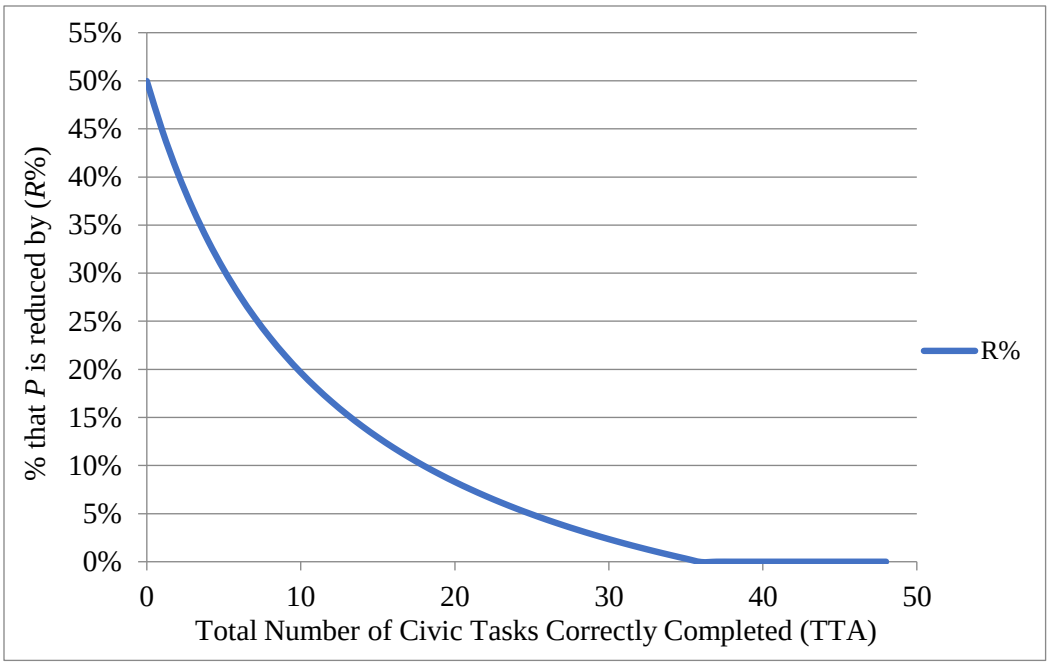
$$\%R = \left(\frac{\frac{800}{(TTA + 12) - 16.7}}{100}\right) \quad (3)$$

The reduction function is designed to start at 50%, decline relatively quickly with a small level of TTA, but reach 0% only when subjects complete an average of 3 civic tasks each (TTA = 36), see Figure 4.1 below.

[53] Note that, while 8 tokens is the socially optimal symmetrical contribution to the public sector when poor governance is sufficiently reduced, as P approaches P*, it is sometimes socially optimal to contribute more than 8 for high levels of poor governance. Subjects had this information available through the table on their screen, and a table to highlight these instances is available in Appendix C.A (Table A.1). 8 tokens remains the privately and socially optimal contribution when both the privately optimal level of civic tasks is completed (0), and when the socially optimal level of civic tasks is completed (35).

[54] This task is adapted from that in Kamei *et al.*, (2019) where it is used to fit the framing of the experiment rather than use a more standard real-effort task, such as a counting task, which may have been jarring for subjects. As we have a neutrally framed treatment, the names of the tasks are changed and we do not vary the questions depending on which task type is selected under either framing.

**Figure 4.1:** *R% as a Function of Total Civic Tasks Correctly Completed (TTA)*




Given the amount earned by completing one private task, it is never privately optimal to complete a civic task in the pre-stage, as completing one civic task does not reduce %R enough in the main stage to offset the potential earnings from a private task (see Table A.1 in Appendix C.A for a payoff schedule). As a result, the pre-stage forms a second-order social dilemma problem, with an expectation (absent behavioural or social preference effects) that subjects dedicate the pre-stage to completing private tasks and suffer a large percentage reduction in allocations to the public sector in the main stage. Civic tasks in the real-world may be thought of as civic engagement, including keeping up with the news, research, protesting, voting, and petitioning, among other activities. This design reflects that, while individual effort can have an impact, collective action is needed to hold the state accountable.

Treatment Variation 1: Framing

The first variation of interest is whether framing makes a difference to subjects' behaviour in the pre-stage. Specifically, whether subjects complete more civic tasks when the experiment is framed to encourage perceiving the experiment as a simulation of society and government, and the percentage reduction as a loss through poor governance. In the politically framed treatment, framing is applied to both the instructions and the experiment and is achieved by replacing several keywords and including two additional segments. The first explicitly links the percentage reduction to imperfect governance:

> "Although having a government to enforce a penalty scheme can increase the amounts citizens allocate to the public sector, potentially increasing earnings, real-world governments sometimes have leaders and officials that don't act fully in the public's interest. Indeed, some government revenue can be lost to lax oversight, negligence, or corruption by government officials."

The second provides examples of civic engagement:

> "In what follows, we assume that governments exhibit less corruption when citizens engage more in public affairs. Examples of civic engagement in the real world include paying attention to the news, voting in elections, participating in a campaign or rally, signing a petition, or other actions that may hold a government to account."

Full sets of both politically and neutrally framed instructions are included in Appendix C.B.

Treatment Variation 2: Reputation and Feedback

The second variation is one of reputation and feedback, compared to anonymity. In the treatments without feedback, subjects are only made aware of their own actions and overall totals or averages of the group's behaviour. By contrast, in the treatments with feedback, subjects are given fixed player identities (letters) and can view and give feedback on the actions of 3 other members in the group (the four forming a permanent subgroup). These subgroups, or 'social circles' in the politically framed treatments, are designed to simulate the smaller social sphere of people within society that may observe your social activities directly and that you may follow the social activities of, such as family, friends, or local community. At the end of the pre-stage, each subject is told how many civic tasks each of their subgroup members completed, as well as the group total. They can then choose to assign positive feedback (a smiley face) to as many or few of their subgroup members as they like (see image 1.a)[55]. Once feedback has been assigned, the subgroup members will be shown the feedback results for each subgroup member, including themselves, next to their player identity as shown below (image 1.b). The feedback does not impact either the sender or receiver's earnings; it is designed to give them a sense of reputation and a mechanism for social approval.

[55] This experiment focuses on positive feedback (a smiley face) only as this most closely represents what is available on most social media. While a 'dislike' button was considered for Facebook, it was intended for conveyance of empathy rather than negativity (Morse, 2015), and the company later moved in the direction of specific emotion emojis such as laughing or angry faces. Similarly, downvoting was trialled on Twitter and is currently available on TikTok, however the information is private to viewers and so does not convey a social cue. Using only positive feedback, compared to using multiple more complex symbols, also has the benefit of simplifying the interpretation of feedback sent and received.

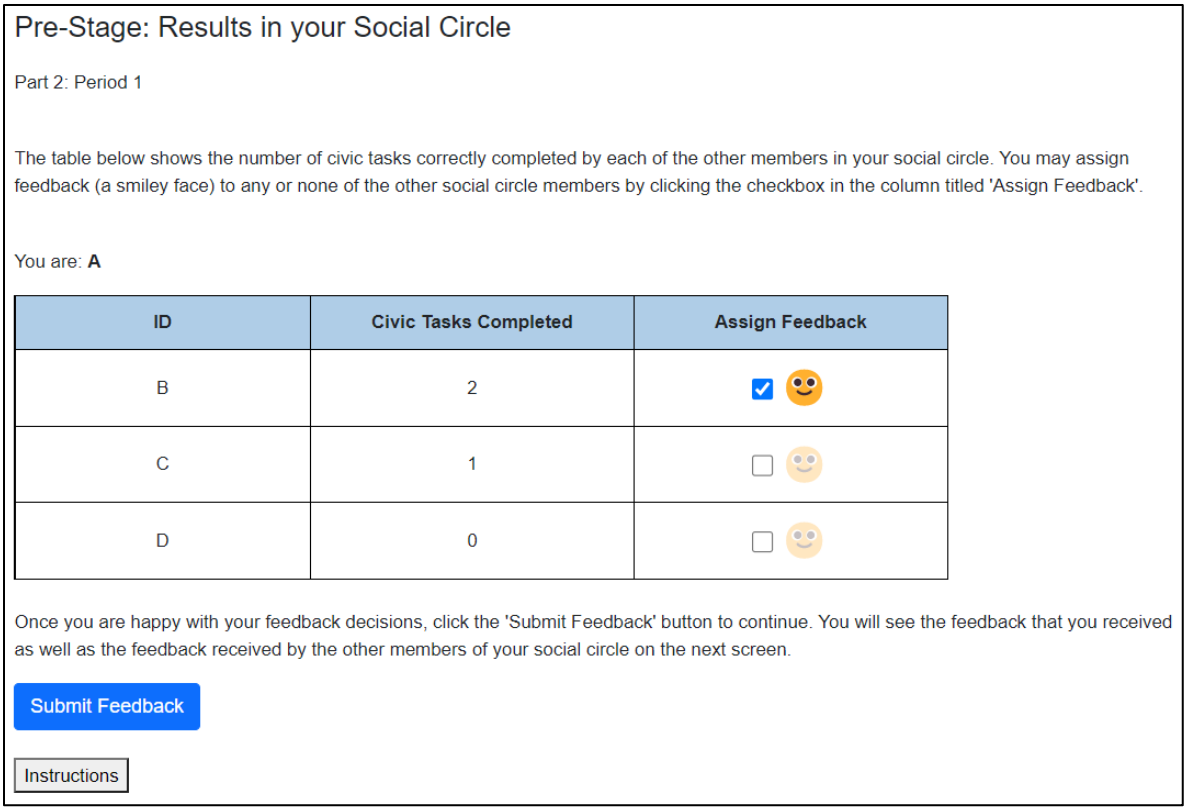



**Image 1.a:** *Assignment of Positive Feedback (Politically Framed Treatments)*

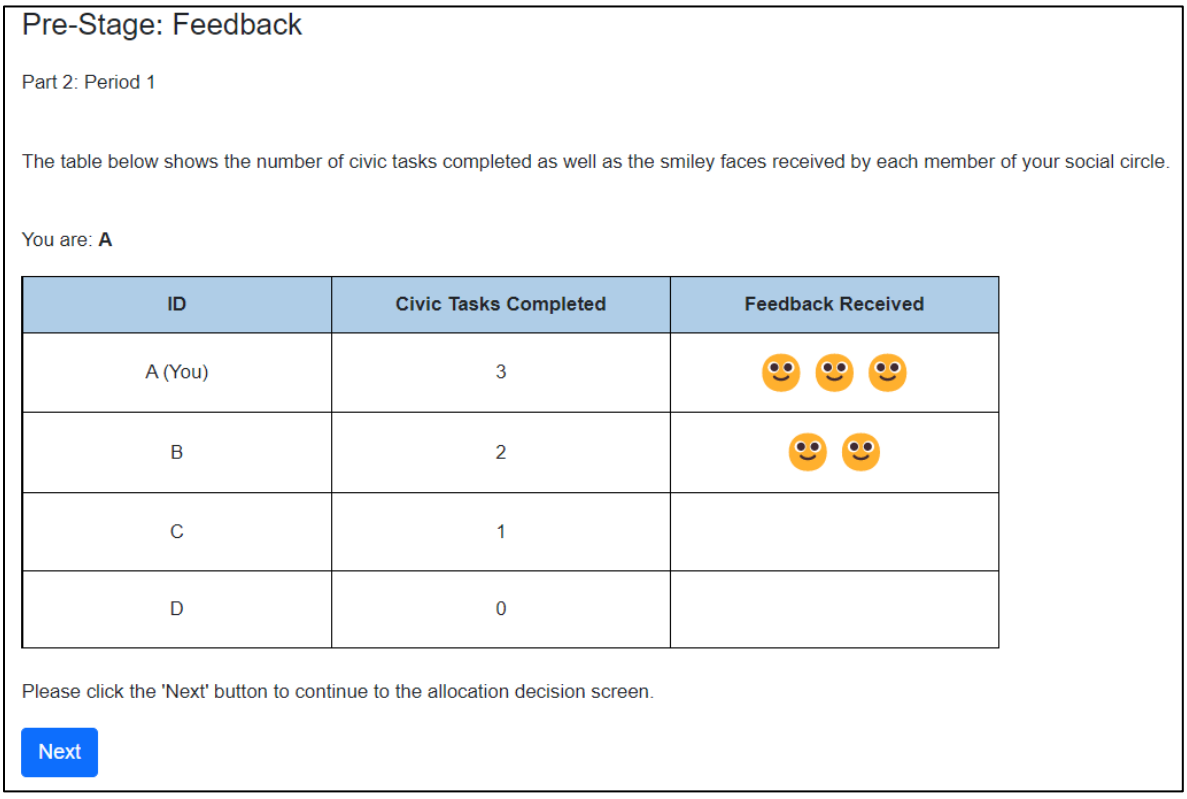



**Image 1.b:** *Receipt of Positive Feedback (Politically Framed Treatments)*

Four sessions per treatment were conducted across the Experimental and Behavioural Economics Lab at Newcastle University and the Exec Lab at York University. The treatments were split so that 2 of each treatment session were conducted at each location, and a control dummy is included in all analysis to account for any differences between the populations. A total of 384 subjects were used, with 12 subjects forming a group and 2 groups per session. Each session took approximately 90 minutes with an average payment of £21.35 including a £3 show-up fee.

*4.3.2 Hypotheses*

In this section, we consider the actions of a self-interested utility maximiser, with the assumption of common knowledge, for this experimental setup. First, in Part 1 of all treatments, subjects play a PGG with the standard incentive to allocate 0 tokens to the public sector, as described in Section 4.3.1. However, the literature suggests that the initial level of cooperation in a standard repeated PGG start at around 40-60% of the social optimum (Ledyard, 1995) before declining toward the privately optimal level of 0, and so we anticipate a similar trend for allocations to the public sector here as there is no reason to believe the population should differ.

**Hypothesis 1:** *In Part 1, allocations to the public sector will decline towards or approximate 0.*
**H1**: $p_i \; declines \; toward \; 0 \; for \; all \; i$

In Part 2, we introduce the minimum required allocation of 8 tokens to the public sector and reductions to the public sector through a percentage loss. As the fine for allocating less than 8 tokens to the public sector is large and deterrent, it is no longer privately optimal to continue reducing allocations beyond this point. It is also never privately optimal to allocate more than 8 tokens to the public sector as the shape of the D(P) and V(P) functions, and so the payoff schedules, are identical to those in Part 1 for allocations to the public sector of 8 and above, irrespective of the level of loss.

**Hypothesis 2:** *In Part 2, allocations to the public sector will remain stable at the minimum required amount of 8 tokens.*
**H2**: $p_i = 8 \; for \; all \; i$

We now turn to predictions for pre-stage behaviour. As discussed in Section 4.3.1, due to the shape of the loss function and the opportunity cost of completing a civic task at the cost of forgoing a private task, it is never privately optimal to complete a civic task. This is because the individual increase to points earned from reducing R% by one civic task never exceeds 10 points, and this only decreases as more civic tasks are completed by others (shown in Appendix C.A, Table A.1). Given this, a self-interested utility maximiser should exclusively complete private tasks in the pre-stage, and

as a result, experience the highest percentage loss (50%) to the public sector in the main-stage. Some subjects may complete civic tasks initially out of curiosity or error, but if this is the case we would expect civic engagement to be at very low levels and decline toward 0 over time.

**Hypothesis 3:** *In Part 2, civic task completion will decline towards or approximate 0.*
$\boldsymbol{H3}$: $t_{c,i}$ $declines\ toward\ 0\ for\ all\ i$

We further consider how the two treatment variations, feedback and framing, may impact a self-interested utility maximiser's decisions. Feedback in this setup is ramification free, meaning that it does not impact the payoffs of an individual and so should not affect incentives. However, much larger payoffs are available should even a small subset of the group manage to cooperate. In line with the literature discussed above, the repeated visibility of the civic task completion of 3 other social circle members and the ability to signal one's intention to cooperate (or approval of others' cooperation) may serve as a tool to enable and sustain cooperation. As sending feedback is also cost free, a self-interested utility maximiser may choose to encourage others to complete civic tasks, which is materially beneficial regardless as to whether one intends to complete tasks themselves. Whether attempts to encourage cooperation are successful or not, they are more likely to occur and succeed in the treatments with feedback than without, owing to this ability to track the actions of and signal to those within their social circle.

**Hypothesis 4:** *Civic task completion will be greater in treatments with feedback than in treatments without feedback.*
$\boldsymbol{H4}$: $0 \leq {T_c}^{NF} < {T_c}^{F}$

Lastly, we do not anticipate any difference in behaviour, either in allocation decisions or civic task completion, due to framing as it does not alter the material incentives of the experiment. While some literature has suggested label framing may increase cooperative behaviour when a community context is highlighted, the results are mixed and heterogenous among different populations.

**Hypothesis 5:** *Allocations to the public sector will not differ between the politically and neutrally framed treatments.*
$\boldsymbol{H5}$: $P^N = P^P\ for\ all\ i$

**Hypothesis 6:** *Civic task completion will not differ between the politically and neutrally framed treatments.*
$\boldsymbol{H6}$: ${T_c}^{N} = {T_c}^{P}\ for\ all\ i$

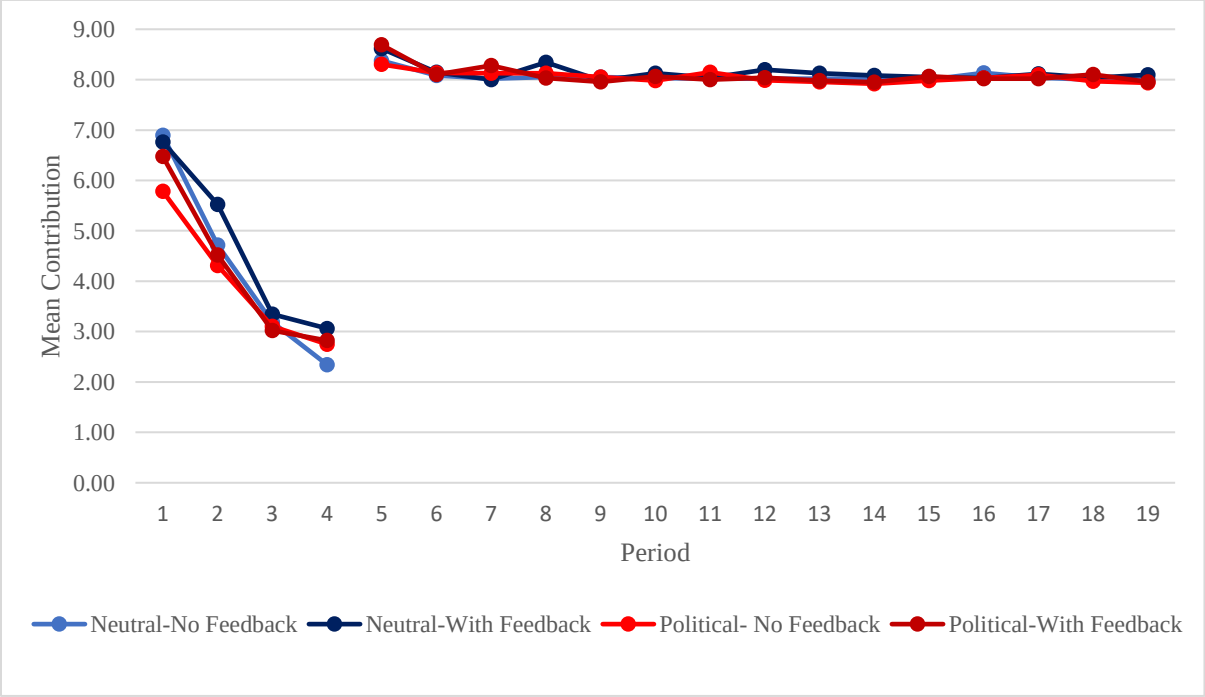
## 4.4 Results

### *4.4.1 Main-Stage Allocations to the Public Sector*

First we provide an overview of allocation decisions in Part 1 and the main stage of Part 2 in the experiment. As shown in Figure 4.2, all four treatments exhibit the expected downward trend of contributions to the public sector in Part 1, when no minimum contribution was required. When pooled, the average allocation to the public sector was 6.48, or 81% of the social optimum in the first period, which declines steeply to 2.74, or 0.34% of the social optimum, by period 4. This quick descent highlights the standard public goods dilemma, as free-riding is quickly realised as the privately optimal strategy as predicted in H1.

The introduction of the minimum required contribution of 8 tokens has a clear effect in Part 2 by increasing allocations to the public sector considerably. While not an unexpected result, given the highly deterrent level of fine, Part 2 allocations to the public sector highlight the efficiency of an effectively implemented sanction system on securing sufficient revenue to provide a public good.

**Figure 4.2:** *Mean Contribution to the Public Sector by Treatment*




These findings are confirmed in Table 4.2 using non-parametric tests. Part 1 allocations are significantly different to Part 2 allocations in each of the four treatments. While feedback should not have an impact on allocation decisions, as it relates to civic task completion, framing has been shown

to increase cooperativeness in some circumstances (see Section 4.2). However, in line with H5, we do not find any differences in allocation behaviour between treatments in either Part 1 or Part 2.[56]

**Table 4.2:** *Average Contributions to the Public Sector*

| | Avg. Allocation to the Public Sector | | | |
|---|---|---|---|---|
| Treatment | (i) Part 1 | (ii) Part 2 | *p*-value for $H_0$: (i) = (ii) | *p*-value for $H_0$: (ii) = 8 |
| [Individual treatments:] | | | | |
| (a) Neutral-No Feedback | 4.29 | 8.06 | 0.0117** | 0.0018*** |
| (b) Neutral-With Feedback | 4.67 | 8.13 | 0.0117** | 0.0289** |
| (c) Political-No Feedback | 3.99 | 8.05 | 0.0116** | 0.6110 |
| (d) Political-With Feedback | 4.21 | 8.08 | 0.0117** | 0.0738* |
| [Across-treatment comparisons:] | | | | |
| *p* for $H_0$: (a) = (b) | 0.4005 | 0.3703 | --- | --- |
| *p* for $H_0$: (a) = (c) | 0.4008 | 0.2926 | --- | --- |
| *p* for $H_0$: (a) = (d) | 1.0000 | 0.8332 | --- | --- |
| *p* for $H_0$: (b) = (c) | 0.1412 | 1.0000 | --- | --- |
| *p* for $H_0$: (b) = (d) | 0.3442 | 0.3710 | --- | --- |
| *p* for $H_0$: (c) = (d) | 0.6744 | 0.3713 | --- | --- |

*Notes*: All *p*-values are based on two-sided tests. Wilcoxon signed rank (Mann-Whitney) tests were conducted for within(across)-treatments comparisons in columns 1 to 3, and single-sample t-tests were used in column 4, using group means of contributions to the public sector. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Having said that, contrary to H2, in three of the four treatments average allocations are significantly different to 8 in Part 2 (at the 10% level or lower). While only 6.51% of subjects contributed less than 8 tokens to the public sector on average, and the majority of subjects (58.33%) do contribute exactly 8 tokens, 35.16% contribute more than 8 tokens to the public sector on average. As discussed in section 4.3, there are some instances in which it is socially optimal to contribute more than 8 tokens due to the level of loss. Those averse to inefficiency or with other-regarding preferences may be willing to allocate more than privately optimal in order to offset the loss caused by insufficient civic engagement in the pre-stage. To check this, we run individual-level regressions to see if Part 2 contributions are affected by the level of loss in a given period, as well as other

[56] We further check for differences in rate of decline of allocations to the public sector in Part 1 and Part 2 by regression (see Appendix C.A, Table A.2). While behaviour differs slightly in the politically framed treatment without feedback (Political-No Feedback) compared to the two neutrally framed treatments, this difference is not consistent across models and does not extend to Part 2.

cooperative indicators such as initial levels of contribution in the experiment and civic task behaviour; results are reported in Table 4.3 below. When a trend variable is included (Period), the level of loss to the public sector is a significant and positive predictor of Part 2 contributions, as is a subjects first allocation decision in the experiment ($p_{i,1}$). While the effects are small, these suggest that some subjects do attempt to offset the level of loss by contributing more than the privately optimal amount to the public sector. Lastly, while allocations to the public sector are not equal to 8, Table 4.3 does confirm a downward sloping trend as predicted in H2 (this is also confirmed in Appendix C.A, Table A.2). Indeed, when comparing the first and second half of the experiment (periods 1-7 and periods 8-15, separately), contributions to the public sector are not significantly different from 8 in the latter (see Appendix C.A, Table A.3.)

**Table 4.3:** *Contributions to the Public Sector in Part 2*

Dependent variable: The tokens allocated to the public sector per person in period t

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (a) Neutral-With Feedback dummy | 0.11** | 0.11** | 0.15*** | 0.16** |
| | (0.05) | (0.05) | (0.05) | (0.06) |
| (b) Political-No Feedback dummy | 0.00 | 0.02 | 0.01 | 0.04 |
| | (0.06) | (0.08) | (0.08) | (0.08) |
| (c) Political-With Feedback dummy | 0.05 | 0.06 | 0.09** | 0.10* |
| | (0.05) | (0.05) | (0.04) | (0.06) |
| (d) $p_{i,1}$ {own contribution in pd. 1} | --- | 0.02** | --- | 0.02* |
| | | (0.01) | | (0.01) |
| (e) $t_c$ {own # of civic tasks in pd. t} | --- | -0.01 | --- | -0.01 |
| | | (0.01) | | (0.02) |
| (f) % Loss {% lost from the Public Sector in Period t } | 0.00 | 0.00 | 0.01** | 0.01** |
| | (0.00) | (0.00) | (0.00) | (0.00) |
| (g) Period {=1-15} | --- | --- | -0.02*** | -0.02*** |
| | | | (0.00) | (0.00) |
| Constant | 7.95*** | 7.83*** | 8.02*** | 7.90*** |
| | (0.06) | (0.07) | (0.07) | (0.08) |
| # of observations | 5,760 | 5,760 | 5,760 | 5,760 |
| # of left-censored observations | 18 | 18 | 18 | 18 |
| # of right-censored observations | 7 | 7 | 7 | 7 |

| | | | | |
|---|---|---|---|---|
| Control | Yes | Yes | Yes | Yes |
| Wald $\chi 2$ | 12.27 | 19.50 | 48.86 | 48.97 |
| Prob > Wald $\chi 2$ | 0.031 | 0.007 | 0.000 | 0.000 |
| [P-Value for Wald $\chi 2$ tests of coefficient differences]: | | | | |
| $H_0$: (a) = (b) | 0.195 | 0.380 | 0.179 | 0.274 |
| $H_0$: (a) = (c) | 0.313 | 0.391 | 0.199 | 0.284 |
| $H_0$: (b) = (c) | 0.461 | 0.683 | 0.391 | 0.541 |

Note: Individual-level random effect tobit regressions. Number in parenthesis are bootstrapped standard errors. Individual level data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Result 1:** *(a) Individuals reduced their contribution to the public sector towards 0 when there was no minimum contribution in Part 1, in line with H1. (b) Individuals reduced their contributions to the public sector towards 8 in Part 2, supporting H2. (c) There is no difference in contribution behaviour between treatments in Part 1 or Part 2, as predicted in H5.*

*4.4.2 Civic and Private Task Completion*

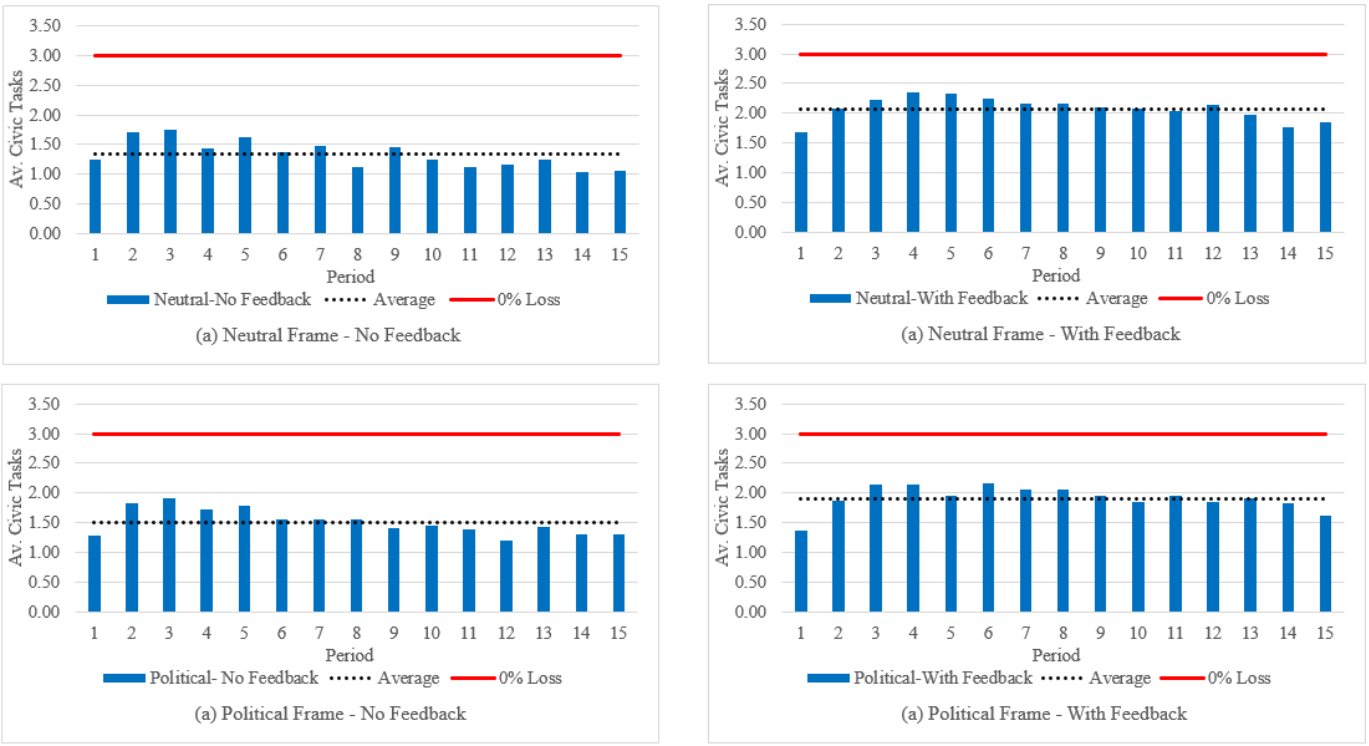
We first consider the average number of civic tasks completed correctly in Part 2. As shown in Figure 4.3, and contrary to hypothesis H3, in each treatment subjects do dedicate effort to completing at least some civic tasks. The average total number of civic tasks completed by a group across all treatments was 20.52 tasks, or an average of 1.71 tasks per subject, resulting in the reduction of the average percentage loss, R%, in the main stage from 50% to 8.99%, less than a fifth of the initial level. Indeed, the level of civic engagement across treatments is relatively high. While no treatment eradicated the percentage loss entirely, in only the Neutral-No Feedback treatment did R% ever exceed 15%.[57]

We also find a downward trend to civic task completion over the course of the experiment (see Table 4.4), in line with H3. However, while significant, the size of the downward trend is surprisingly small given the strong incentive to complete fewer or no civic tasks in favour of private tasks. In all but the Neutral-No Feedback treatment, civic task completion is higher in the last period than it is in the first. Table A.5 in Appendix C.A confirms this by reporting the same analysis as in Table 4.4 but separating the experiment into halves. Without interactions, the coefficient for the trend

[57] Reflecting the relatively high level of civic engagement across all treatments, the payments received by subjects are similar. Table A.4 in Appendix C.A reports Mann-Whitney comparisons for payments by treatment and finds significant differences between Neutral-No Feedback and Neutral-With Feedback, and between Neutral-With Feedback and Political-No Feedback.

variable is positive and significant for the first half of the experiment (periods 1 to 7 of Part 2, see Table A.5, column 2a), while it is significant and negative in the second half of the experiment (periods 8-15 of Part 2, see Table A.5, column 2b)[58].

**Figure 4.3:** *Average Number of Civic Tasks by Treatment and Period*




Note: The red line shows the average number of civic tasks per person per period required to reduce the percentage loss to 0%. The dotted black line shows the average number of civic tasks per person completed across all 15 periods in Part 2.

Figure 4.3 further reveals significant treatment effects. Firstly, the two treatments with feedback mechanisms available see significantly higher levels of civic task completion, as shown by the black dotted lines. Groups in the Neutral-With Feedback and Political-With Feedback treatments answered on average 24.95 and 22.95 civic tasks correctly, respectively, compared to 16.03 and 18.16 in the Neutral-No Feedback and Political-No Feedback treatments, respectively. These differences are confirmed in Table 4.4 using group-level linear regressions. Note that the coefficients are significant

[58] Kamei *et al.*, (2019) find a similar upwards trend in the initial number of civic tasks completed, which they attribute to learning the difference in earnings between having the minimum contribution implemented or not. While this may be the case for learning the impact of loss also, here it may also be explained by becoming more effective at task completion overall. Table A.6 in Appendix C.A shows that task completion overall (private and civic tasks combined) increased over time, suggesting an improvement in task ability rather than substitution between civic and private tasks in these early periods. Table A.6 also confirms that there is no difference in task-completing ability between treatments.

and stable across specifications without interaction terms. When interaction terms are introduced in model (3), both interactions with respect to the treatments with feedback are positive and significant.

**Table 4.4:** *Group Average Civic Task Completion per Person*

Dependent variable: The group average number of civic tasks correctly completed per person

| | (1) | (2) | (3) |
|---|---|---|---|
| (a) Neutral-With Feedback dummy | 0.74*** (0.16) | 0.74*** (0.16) | 0.55*** (0.19) |
| (b) Political-No Feedback dummy | 0.18 (0.14) | 0.18 (0.14) | 0.12 (0.22) |
| (c) Political-With Feedback dummy | 0.58*** (0.16) | 0.58*** (0.16) | 0.31 (0.21) |
| (d) Period {=1 to 15} | --- | -0.02*** (0.01) | -0.04*** (0.01) |
| (e) Interaction: (a) × Period | --- | --- | 0.02** (0.01) |
| (f) Interaction: (b) × Period | --- | --- | 0.01 (0.01) |
| (g) Interaction: (c) × Period | --- | --- | 0.03*** (0.01) |
| Constant | 1.20*** (0.13) | 1.38*** (0.15) | 1.51*** (0.18) |
| # of observations | 480 | 480 | 480 |
| Control | Yes | Yes | Yes |
| Wald $\chi 2$ | 39.82 | 83.78 | 153.46 |
| Prob > Wald $\chi 2$ | 0.000 | 0.000 | 0.000 |
| [P-Value for Wald $\chi 2$ tests of coefficient differences]: | | | |
| $H_0$: (a) = (b) | 0.000*** | 0.000*** | 0.015** |
| $H_0$: (a) = (c) | 0.297 | 0.297 | 0.133 |
| $H_0$: (b) = (c) | 0.007*** | 0.007*** | 0.357 |
| $H_0$: (d) + (e) = 0 | --- | --- | 0.073* |
| $H_0$: (d) + (f) = 0 | --- | --- | 0.011** |
| $H_0$: (d) + (g) = 0 | --- | --- | 0.605 |

Note: Group-level random effect linear regressions. Number in parenthesis are robust standard errors. Group average data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

The difference between treatments with and without feedback also persists across the first and second half of Part 2 (periods 1 to 7 and 8 to 15, respectively). Reporting Wald tests for differences between treatment dummy coefficients, the bottom of Table A.5 in Appendix C.A confirms significant differences across specifications in average civic task completion between the Political-No Feedback and the Political-With Feedback treatments ($H_0$: (b) = (c)) and the Neutral-With Feedback and Political-No Feedback treatments ($H_0$: (a) = (b)), across both halves of the experiment.

Lastly, as with contribution decisions, we do not find significant differences in behaviour between the neutrally and politically framed treatments. Whilst positive, the coefficients for the Political-No Feedback dummy (or respective interaction terms) are not significant across the models specified in Table 4.4 or Table A.5 in Appendix C.A. Similarly, Wald tests do not find significant differences between coefficients (a) and (c)[59], which compare the neutrally and politically framed treatments with feedback.

**Result 2:** *(a) In all treatments, subjects dedicate effort to completing civic tasks, contrary to H3, although a slight downward trend is observed in the second half of the experiment. (b) Civic task completion varies considerably by treatment, with treatments allowing for feedback seeing significantly higher civic task completion, in line with H4. (c) There are no significant framing effects present for civic task completion at the group-level, in support of H6.*

### *4.4.2. Treatment Effects*

We now examine the differences between treatments more closely and consider potential motivations for the observed level of civic engagement. Table 4.5 reports individual-level regressions using pooled data and treatment dummies, as in Table 4.4, but replacing the trend variable with other lagged variables, such as a subject's own behaviour and that of their group in the previous round, as well as their first contribution to the public sector in Part 1 as an indicator of their initial cooperativeness. We also introduce variables specific to the treatments with feedback in specifications (3) and (4) to better understand the dynamics between group and social circle behaviour.

[59] In model (3a) of Table A.5 in Appendix C.A, (a)=(c) is significant at the 10% level, however as neither (a) nor (c) are individually significant we do not assign meaning to this result.

Columns (1) and (2) of Table 4.5 show estimation results using pooled observations from all four treatments, while columns (3) and (4) use only data from the treatments with feedback (Neutral-With Feedback and Political-With Feedback). As in Table 4.4, treatments with feedback see significantly higher levels of civic engagement than those without feedback, even when own and group behaviour in previous rounds is taken into account, showing the treatment effects robustness to both trend (Table 4.4) and norm-based variables.

**Table 4.5:** *Dynamics of Civic Task Completion*

Dependent variable: The number of civic tasks correctly completed per person in period t

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (a) Neutral-With Feedback dummy | 1.08*** | 0.86*** | --- | --- |
| | (0.20) | (016) | | |
| (b) Political-No Feedback dummy | 0.45** | 0.36** | --- | --- |
| | (0.20) | (0.18) | | |
| (c) Political-With Feedback dummy | 0.81*** | 0.65*** | -0.27 | -0.21 |
| | (0.26) | (0.20) | (0.21) | (0.16) |
| (d) $p_{i,1}$ {own contribution in pd. 1}$^{2}$* | 0.10*** | 0.08*** | 0.08** | 0.06** |
| | (0.02) | (0.02) | (0.04) | (0.03) |
| (e) $t_{c,-1}$ {own # of civic tasks in pd. t-1} | --- | 0.36*** | --- | 0.35*** |
| | | (0.04) | | (0.05) |
| (f) $T_{c,-i,-1}$ {avg. # of 11 other members' civic tasks in pd. t-1} | --- | -0.06 | --- | --- |
| | | (0.06) | | |
| (g) $T_{c,sc-i,-1}$ {avg. # of 3 other social circle members' civic tasks in pd. t-1} | --- | --- | --- | 0.08** |
| | | | | (0.04) |
| (h) $T_{c,-sc,-1}$ {avg. # of 8 other non-social circle members' civic tasks in pd. t-1}$^{1}$* | --- | --- | --- | -0.16** |
| | | | | (0.07) |
| Constant | 0.06 | -0.14 | 1.15*** | 0.82*** |
| | (0.25) | (0.21) | (0.30) | (0.27) |
| # of observations | 5,760 | 5,376 | 2,880 | 2,688 |
| # of left-censored observations | 1,528 | 1,412 | 568 | 511 |
| Control | Yes | Yes | Yes | Yes |
| Wald $\chi 2$ | 58.67 | 83.78 | 18.98 | 153.46 |
| Prob > Wald $\chi 2$ | 0.000 | 0.000 | 0.000 | 0.000 |

| | | | | |
|---|---|---|---|---|
| [P-Value for Wald $\chi 2$ tests of coefficient differences]: | | | | |
| $H_0$: (a) = (b) | 0.001*** | 0.001*** | --- | --- |
| $H_0$: (a) = (c) | 0.146 | 0.161 | --- | --- |
| $H_0$: (b) = (c) | 0.068* | 0.064* | --- | --- |

Note: [1]* While subjects did not have access to this information, civic tasks by those outside of their social circle can be calculated by deducting the number of civic tasks completed within their social circle from the total amount completed by the group. [2]* This is a subject's allocation to the public sector in the first period of Part 1, used as a moderator for a subject's individual level of cooperativeness. Individual-level random effect tobit regressions; tobit regressions are used due to the number of subjects who completed 0 civic tasks in a given period. Number in parenthesis are bootstrapped standard errors. Individual level data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Cooperativeness also plays a role in motivating civic engagement. The variable capturing one's own contribution to the public sector in the first period of the experiment, indicating a subject's initial willingness to cooperate, remains positive and significant across model specifications, albeit the effect is small. Further, when including a moderator for cooperativeness, we do see a positive and significant framing effect on the individual-level between the treatments without reputation and feedback mechanisms, Neutral-No Feedback and Political-No Feedback, but not between the treatments with such mechanisms.

**Result 3:** *(a) Subjects who contribute more to the public sector in the first period of the experiment complete more civic tasks in the second-order social dilemma. (b) When moderating for initial levels of cooperation, those in the politically framed treatment complete more civic tasks than those in the neutrally framed treatment when the reputation and feedback mechanism is unavailable.*

Specifications (2) and (4) introduce controls for group and/or social circle dynamics. The number of civic tasks a subject completed in the previous period (variable e) is a significant and strong predictor of current period task completion across both models. By contrast, the amount completed by the other eleven members in one's group is not significant.

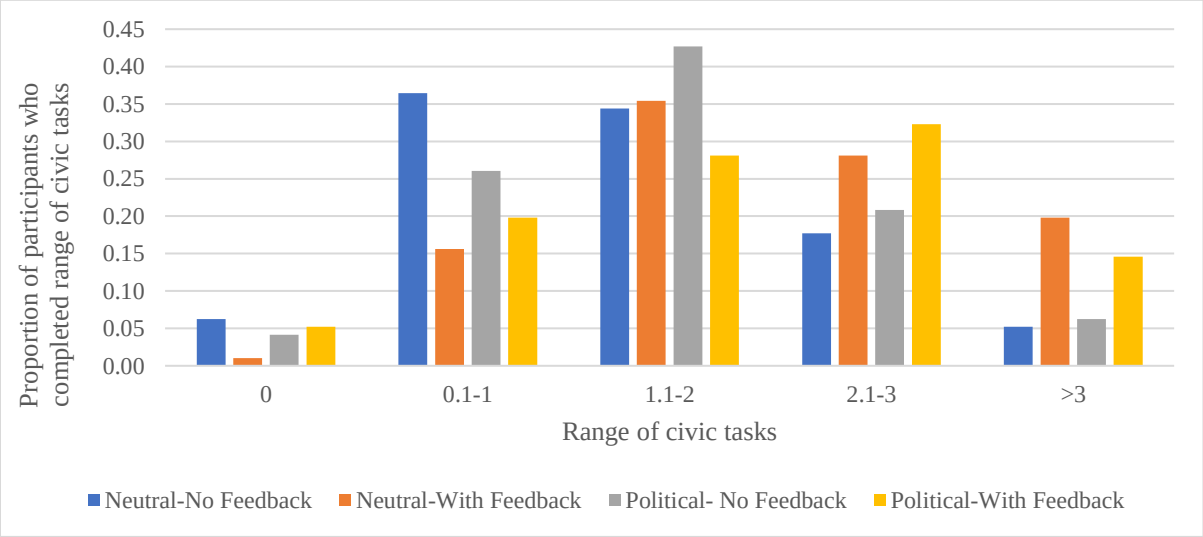
While subjects do not adjust their civic task completion based on the group average, they do react to that of their social circle. Column (4) shows civic task completion for just the two treatments with feedback systems. Here there is a small but positive effect from the civic task completion of the other 3 members in one's social circle in the previous period. This may be a result of social circle norms or encouragement from positive feedback, as discussed in Section 4.4. By contrast, there is a significant and negative effect from the civic task completion of those outside of one's social circle in the previous period. These opposing effects, remarkably similar to those in a parallel specification of

Kamei *et al.*'s (2019) Table 1, help to explain the relative stability of civic task completion in the treatments with feedback compared to those without (Table 4.4). As with the pooled analysis in columns (1) and (2), and the analysis of Table 4.4, there is no framing effect between the politically and neutrally framed treatments when the reputation and feedback mechanism is available.

**Result 4:** *(a) Overall, subjects' civic task completion is sensitive to their own previous civic task completion, but not to that of their fellow group members. (b) When reputation building is available, subjects align their civic task completion to that of their social circle, however, they become negatively responsive to the tasks completed by those outside of their social circle.*

Lastly, we also consider that subject behaviour may be heterogenous as it is possible for coalitions smaller than the full group to entirely eradicate the percentage lost in the main stage by completing more than the symmetric socially optimal level of civic tasks (2.92 per period on average). Figure 4.4 shows the distribution of subjects by the range of civic tasks they completed in a period on average. Notably, across all treatments there are fewer complete free riders than one would expect should there be a significant presence of strict utility-maximisers, with less than 5% of subjects completing no civic tasks correctly.[60]

**Figure 4.4**: *Distribution of subjects by correctly answered civic tasks on average per period*




Further, while the modal group of subjects (35%) correctly completed between 1.1-2 civic tasks, there are once again treatment differences between the treatments with and without feedback.

[60] As we are concerned with correctly answered tasks, this figure includes those who started but did not complete, or incorrectly completed, civic tasks. If we instead look at those who did not start a civic task, this figure falls to 2%.

Strikingly, 20% of subjects in the neutrally framed treatment with feedback, and 15% of subjects in the politically framed treatment with feedback, completed more than 3 civic tasks per period on average, compared to 5% and 6% in their respective no-feedback treatments. These distribution differences are confirmed as significant in Table A.7 in Appendix C.A using Kolmogorov-Smirnov tests. As 2.92 civic tasks is the symmetric socially optimal level of civic task completion per subject, should all group members behave the same on average, these subjects represent high levels of cooperativeness.[61]

**Result 5:** *(a) Complete free-riding with respect to civic engagement only characterises a very small number of participants in the experiment. (b) The majority of participants correctly complete between 1 and 2 civic tasks per period, which while socially suboptimal is enough to eliminate most of the potential loss of public revenue to inefficiency and corruption. (c) The treatments with feedback see considerably higher presence of extremely cooperative individuals, in line with H4.*

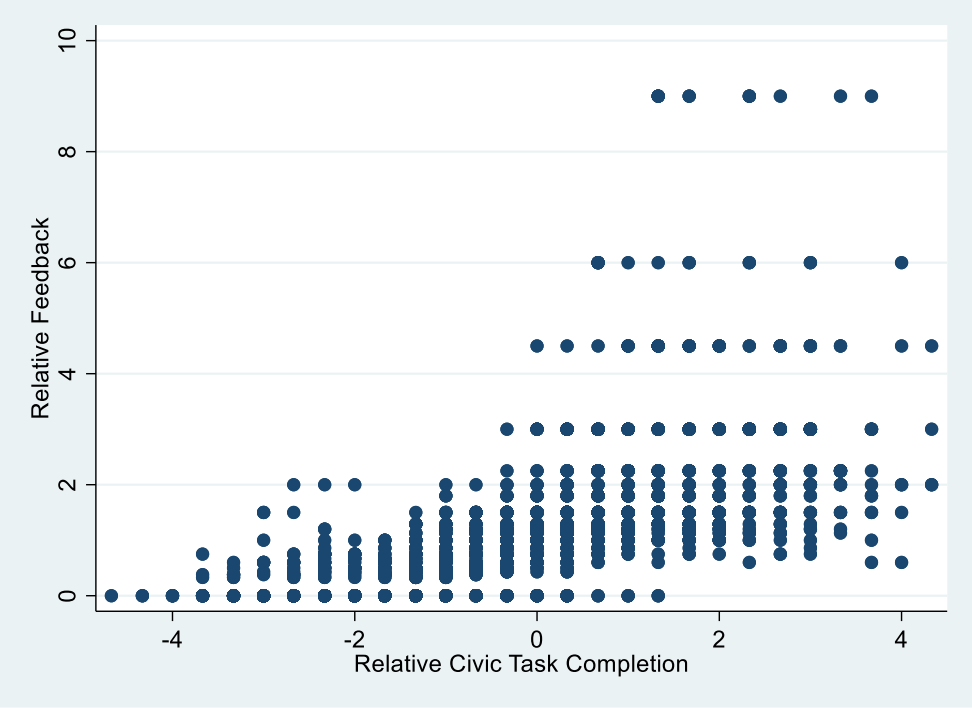
#### *4.4.4.Feedback and Civic Task Completion*

Sections 4.4.2 and 4.4.3 find consistently that having a reputation and feedback mechanism available significantly increases effort towards completing civic tasks. In this section we look at the use of feedback, specifically, who received feedback and how they responded to it.

First, we confirm that feedback is being used in a pro-social manner. Specifically, Figure 4.5 below shows the relative feedback subject i received ($=fb_{i,t}/fb_{sc\text{-}i,t}$) and the number of civic tasks subject i completed relative to their social circle ($=t_{c,i,t} - (\sum t_{c,sc\text{-}i}/3)$). The scatter plot shows that social circle members generally rewarded those who completed more civic tasks with higher levels of positive feedback. This result is confirmed as significant using regression analysis in Appendix C.A, Table A.8. There are also some instances of positive feedback being given to those with low civic task completion, which may indicate heterogeneity in how positive feedback was utilised by social circle members.

[61] As before, this figure underestimates the level of cooperativeness somewhat as it excludes those who incorrectly answered or did not finish attempted civic tasks. If we include these attempts in the distributions, 25% of subjects started 3 or more civic tasks, with a striking 36% and 35% of subjects starting more than 3 civic tasks in the Neutral-With Feedback and Political-With Feedback treatments respectively. Further, the largest distribution bin would instead be 2.1-3 attempted civic tasks at 28% of subjects.

**Figure 4.5**: *Relative Feedback Received to Relative Task Completion in One's Social Circle*




Note: Observations are per individual in Part 2 in treatments with feedback.

**Result 6:** *Positive feedback is given to those who complete more civic tasks relative to the number completed by those in their social circle.*

We can now consider how the feedback received impacted a subject's civic engagement. Table 4.6 reports similar regressions to Table 4.5, but introducing variables that capture one's reaction to the absolute amount of positive feedback received (on a scale of 0 to 3 smiley faces), and the relative feedback received in the form of negative and positive deviation from the feedback received by the other members of their social circle. Specifications (1) and (4) show that receiving more positive feedback, or more positive feedback relative to one's social circle, encourages greater civic task completion. By contrast, receiving less positive feedback relative to the social circle in the previous period discourages civic engagement. The latter result, similar to that found in Kamei *et al.* (2019), may be as receiving less positive feedback is perceived in a similar way to negative feedback, and so may be discouraging rather than motivational, or as those receiving less positive feedback are less (or not) sensitive to it and so reducing their cooperation regardless. These results are robust to specifications (2) and (5) where group dynamics are introduced. While feedback received significantly predicts task completion, we also find (as in Table 4.5) that a subject's civic task completion increases with that within their social circle, but decreases with that outside of their social

circle in the previous period. This suggests that group norms and visibility are also important in a subject's decision whether to civically engage or not. To control for a given subject's level of civic task completion, specifications (3) and (6) introduce a lagged variable for civic task completion in the previous period (f). When doing so, the feedback variables are no longer significant, however, this is most likely the result of strong correlation (as approval is well-targeted) between the feedback variables and variable (f).[62]

**Table 4.6:** *Civic Task Completion and Feedback*

Dependent variable: The number of civic tasks correctly completed per person in period t

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| (a) Political-With Feedback dummy | -0.22 | -0.22 | -0.21 | -0.26 | -0.23 | -0.21 |
| | (0.20) | (0.17) | (0.17) | (0.23) | (0.20) | (0.16) |
| (b) $p_{i,1}$ {own contribution in pd. 1} | 0.08** | 0.08** | 0.06** | 0.08** | 0.08** | 0.06** |
| | (0.03) | (0.04) | (0.03) | (0.04) | (0.04) | (0.03) |
| (c) Feedback received in pd t-1 {=0-3} | 0.23*** | 0.24*** | 0.04 | --- | --- | --- |
| | (0.04) | (0.04) | (0.03) | | | |
| (d) Positive deviation of feedback in pd t-1 {=max{feedback received by i - av. feedback received by others in social circle,0} | --- | --- | --- | 0.13*** | 0.21*** | 0.02 |
| | | | | (0.04) | (0.04) | (0.06) |
| (e) Negative deviation of feedback in pd t-1 {=max{av. feedback received by others in social circle - feedback received by i,0} | --- | --- | --- | -0.17*** | -0.20*** | 0.01 |
| | | | | (0.06) | (0.06) | (0.05) |
| (f) $t_{c,-1}$ {own # of civic tasks in pd. t-1} | --- | --- | 0.32*** | --- | --- | 0.34*** |
| | | | (0.05) | | | (0.06) |
| (g) $T_{c,sc-i,-1}$ {avg. # of 3 other social circle members' civic tasks in pd. t-1} | --- | 0.15*** | 0.09** | --- | 0.26*** | 0.09 |
| | | (0.04) | (0.04) | | (0.05) | (006) |
| (h) $T_{c,-sc,-1}$ {avg. # of 8 other non-social circle members' civic tasks in pd. t-1} | --- | -0.16** | -0.16** | --- | -0.12** | -0.16*** |
| | | (0.06) | (0.07) | | (0.06) | (0.05) |

[62] Pearson's correlation tests reveal strong correlations between the tasks completed in the previous period (f) and feedback received in the previous period (c), $r(686)=0.70$, $p<0.001$, positive deviation of feedback in the previous period (d), $r(686)=0.53$, $p<0.001$, and negative deviation of feedback in the previous period (e), $r(686)=-0.58$, $p<0.001$. This is not unexpected as feedback is well targeted and so the correlation between tasks completed and feedback received should be relatively high.

| | | | | | | |
|---|---|---|---|---|---|---|
| Constant | 0.84*** | 0.84** | 0.80*** | 1.23*** | 0.95*** | 0.80*** |
| | (0.29) | (0.35) | (0.26) | (0.31) | (0.33) | (0.23) |
| # of observations | 2,688 | 2,688 | 2,688 | 2,688 | 2,688 | 2,688 |
| # of left-censored observations | 511 | 511 | 511 | 511 | 511 | 511 |
| Control | Yes | Yes | Yes | Yes | Yes | Yes |
| Wald $\chi 2$ | 63.87 | 67.95 | 129.38 | 49.57 | 98.1 | 144.72 |
| Prob > Wald $\chi 2$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Note: Individual-level random effect tobit regressions; tobit regressions are used due to the number of subjects who completed 0 civic tasks in a given period. Number in parenthesis are bootstrapped standard errors. Individual level data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Result 7:** *(a) Civic task completion in the treatments with feedback increases with the positive feedback received in the previous period, whether absolute or relative to the feedback received in one's social circle. By contrast, receiving less positive feedback relative to one's social circle decreases civic task completion. (b) Civic task completion also varies with group dynamics; subjects' increase task completion with that of the other three members in their social circle, but decrease it with respect to task completion outside of their social circle.*

### *4.4.5. Experiment Behaviour, Demographic Data, and Survey Responses*

Demographic data and survey responses were collected as part of the experiment including gender and academic scores, as well as proxies for civic engagement outside of the laboratory and political views. Table 4.7 reports the partial correlation coefficients and p-values for these questions against several experiment variables.

**Table 4.7:** *Partial Correlations between Subject Behaviours and Survey Data*

| Survey question: | Obs. | Av. # of Civic Tasks Correctly Completed [#1] | Likelihood of Completing < 1 Civic Task Per Period on Average [#1] | Likelihood of Completing > 2.92 Civic Tasks Per Period on Average [#1] | Contribution to the Public Sector in Period 1 of Part 1 |
|---|---|---|---|---|---|
| Female dummy {=1 for female, =0 for not female} | 372 | 0.13* | -0.03 | 0.05 | 0.05 |
| | | (0.063) | (0.447) | (0.701) | (0.587) |

| | | | | | |
|---|---|---|---|---|---|
| British national {=1 for British or joint British nationality, =0 otherwise} | 362 | 0.05*** (0.001) | 0.01 (0.229) | 0.08*** (0.006) | -0.08 (0.414) |
| Economics degree {=1 for studying economics, =0 otherwise} | 372 | 0.06 (0.912) | -0.06 (0.544) | 0.04 (0.628) | -0.06 (0.528) |
| # of economics modules taken | 292 | -0.03** (0.014) | -0.07 (0.399) | -0.07 (0.243) | -0.01 (0.952) |
| # of politics modules taken | 294 | 0.04 (0.722) | -0.06 (0.119) | 0.05 (0.784) | 0.02 (0.802) |
| Higher-level maths dummy {=1 for A-level or equivalent, 0=otherwise} | 320 | 0.21* (0.064) | -0.21 (0.747) | 0.10 (0.163) | 0.21** (0.029) |
| GCSE English grade | 254 | -0.09 (0.918) | 0.05 (0.781) | -0.08 (0.930) | -0.03 (0.766) |
| GCSE maths grade | 247 | -0.09 (0.377) | 0.20 (0.642) | 0.05 (0.398) | -0.13 (0.170) |
| Employment dummy {=1 for employed, 0=otherwise} | 346 | 0.03 0.371) | 0.08 (0.827) | 0.04 (0.661) | -0.09 (0.332) |
| Interest in politics[#2] | 383 | 0.11 (0.660) | -0.10 (0.757) | 0.13 (0.557) | 0.06 (0.524) |
| Engagement with media[#3] | 383 | -0.04 (0.455) | 0.04 (0.256) | -0.06 (0.881) | -0.01 (0.946) |
| Strength of government view[#4] | 374 | 0.00* (0.057) | 0.03 (0.197) | -0.04 (0.357) | 0.02 (0.855) |
| Political views (Left to Right wing)[#5] | 358 | -0.22*** (0.000) | 0.20*** (0.000) | -0.20*** (0.001) | 0.00 (0.978) |
| Likelihood to vote[#6] | 259 | -0.05 (0.686) | 0.18* (0.062) | 0.03 (0.499) | -0.17* (0.080) |
| Civic norm strength[#7] | 376 | -0.02 (0.134) | -0.13 (0.160) | -0.03 (0.799) | 0.08 (0.423) |
| Civic engagement[#8] | 370 | -0.09** (0.013) | 0.10** (0.046) | -0.04 (0.182) | -0.21** (0.032) |
| Trust in others[#9] | 377 | 0.00 | 0.08 | 0.04 | 0.04 |

| | | | | | |
|---|---|---|---|---|---|
| | | (0.947) | (0.610) | (0.779) | (0.648) |
| Religious dummy {=1 for | | -0.15*** | 0.03 | -0.10** | 0.01 |
| religious, 0=otherwise} | 331 | (0.001) | (0.570) | (0.015) | (0.916) |

Notes: Values are partial correlation coefficients of a given decision with survey responses. Numbers in parenthesis are p-values. [#1]P-values are based on individual linear (logit for columns 2 and 3) regressions with standard errors clustered by group due to potential correlation within groups for these decisions. [#2] "How interested would you say you are in politics? Please answer on a 4-point scale." (1=very interested to 4=not very interested). [#3] "How often do you watch or listen to broadcasts or read media (including online) about world, national, or local news, including coverage of the positions of political candidates?" (1= 'Multiple times per day' to 6= 'Never or almost never'). [#4] "Some people believe that a substantial government role is required to achieve a healthy economy for a country's people, while others feel that the smaller the government role, the greater is overall prosperity. Please place your view along or at the appropriate end of the spectrum between these. government role, the greater is overall prosperity. Please place your view along or at the appropriate end of the spectrum between these." (1 = A lesser government role is beneficial to 7 = A substantial government role is beneficial'. [#5] "In politics, people talk of a spectrum of views from 'left' to 'right'. Please characterize where your own views fall by selecting a point among those below, where further towards the Left or Right end indicates the strongest leaning in one or the other direction, and nearer the center means less or no strong leaning." (1 = Left to 7 = Right). [#6] "Please indicate how likely you are to vote in the next UK General Election?" (1, 'Very Likely' to 5, 'Very Unlikely'). [#7] How justified is cheating in one's favor in dealings with public sector? This has 3 components: claiming government benefits, avoiding paying fare on public transit, cheating on taxes. Each part is coded from 1 = always justifiable to 10 = never justifiable. The Civic norm strength variable averages the three scores if more than 1 is answered. [#8]Average of 6 civic activities a person 1='has done', 2='might do', 3='Would never do'. [#9] "Do you think that most people would try to take advantage of you if they got a chance, or would they try to be fair? Please show your response on this scale, where 1 means that "people would try to take advantage of you" and 10 means that "people would try to be fair"". *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Of primary interest is how civic engagement and political views outside of the laboratory impact civic engagement within the experiment, as this speaks to external validity and helps to explain some of the heterogenous behaviour of the subjects discussed in Section 4.4.3. Notably, there is a significant and positive relationship between being more civically engaged outside of the laboratory (smaller values for the 'civic engagement' variable indicate higher levels of engagement) and completing more civic tasks in the experiment. Further, we find a negative relationship between the average level for civic tasks completed and politically right views; put differently, those who subscribe to more left-wing political ideology complete more civic tasks. This is perhaps unsurprising, as left-wing ideology is associated with preferring greater economic or social equality, which is achieved in the experiment through putting effort into reducing the amount lost from the public sector.

We also look more closely at those who completed less than 1 civic task on average (low cooperators) and more than the socially optimal amount on average (super cooperators) in isolation.

As before, those who subscribe more to left-wing ideology are significantly more likely to be super-cooperators, while those who lean more to the right are significantly more likely to be low-cooperators. We also find that those who are less civically engaged outside of the laboratory are significantly more likely to be low cooperators.

**Result 8:** *(a) Those who associate themselves with left-wing ideology complete more civic tasks on average and are more likely to be 'super-cooperators'. Those who lean more toward right-wing ideology are more likely to be 'low-cooperators'. (b) Involvement in more civic activities outside of the laboratory is positively correlated with completing more civic tasks within the experiment. Subjects that are less civically engaged outside of the laboratory are more likely to be 'low-cooperators' in the experiment.*

### 4.5. Conclusion

Centralised sanctioning mechanisms are known to increase contributions to a public good in laboratory experiments and resemble the main method of public good provision by governments of taxation. However, it is often taken for granted that these mechanisms are efficient once the required contributions are collected, whereas in actuality, government revenue is often lost to negligence, corruption, and inefficiency. By leveraging a relatively small amount of effort to be civically engaged, citizens can hold the government accountable and reduce the revenue lost. However, unlike the first-order problem of collecting tax revenue, civic engagement cannot be mandated and requires voluntary individual effort to sustain. In this study, we explicitly model the problem of government accountability as a second-order public good problem using a two-stage experiment. Subjects complete real effort tasks in the first stage, resembling the effort they could divert into private activities which boost their personal income, or into civic engagement to help hold the state accountable. In the second stage, the level of collective civic engagement by the group dictates how much the amount of government revenue lost to poor governance is reduced by before individual earnings are calculated.

The results show a surprisingly high level of civic engagement across treatments, despite the strong private incentive to free-ride. While no treatment reaches the socially optimal level of civic engagement, the vast majority of subjects (71%) correctly completed more than 1 civic task per period and only 4% are classed as complete free-riders. Further, the strong downward-sloping trend in cooperation expected of a social dilemma without corrective mechanisms does not materialise for civic engagement; while a downward trend in civic tasks completed is confirmed in the experiment, it is only very slight, with only one treatment ending with lower levels of civic engagement than it

started with. We also find strong treatment effects for the availability of a reputation and feedback mechanism. When available, civic task completion is significantly greater than when unavailable, and this result is robust across specifications. The impact of reputation and feedback is two-fold. Firstly, feedback is well targeted towards those who complete more civic tasks, and those who receive more positive feedback (absolutely and relatively) increase their civic task completion in the following period. Secondly, subjects increase their civic engagement relative to the level of civic tasks correctly completed by the others in their social circle in the previous period, showing a keenness to be perceived as a high contributor or to maintain group norms. By contrast, they decrease their civic engagement relative to those outside of their social circle, similarly to those who have no social circle in the no-feedback treatments, showing that decision visibility and the opportunity to signal intent or approval is important in encouraging and sustaining cooperation. The impact of the feedback mechanism on cooperation is striking given that there is no material ramification to the positive feedback and subjects were only identifiable by their assigned letter, rather than their name or face.

A politically framed treatment was introduced to contextualise the experiment and to help subjects associate civic tasks with real civic activity. While we do not find strong evidence for framing effects in this experiment, we do find other indications of external validity. Notably, those who are more civically engaged outside of the laboratory complete more civic tasks on average, and those who are less civically engaged are more likely to be low-cooperators. Further, those who hold more left-wing beliefs complete more tasks on average and are more likely to be super-cooperators, while the opposite is true of those leaning toward the right.

The main finding of this study, that subjects will voluntarily put in considerable real effort in order to resolve a second-order social dilemma, explains how governments are held somewhat accountable by their citizens despite the private costs to do so. Completing civic tasks, or civic engagement in the real world, is a relatively small cost compared to the ramifications of a corrupt or ineffective government, and so there is a clear leverage effect for citizens to take advantage of. While civic engagement remains privately costly, despite this effect, mechanisms that increase the visibility of civic behaviour or allow costless signalling can be enough to encourage cooperation. The level of cooperation achieved, and whether it is sufficient to meet the needs of public good provision in the long-run, remain uncertain. In this experiment, all treatments managed to sustain the level of loss consistently below 18%, with an overall average of 8.99%. While this is above the optimal level of loss, it may be that the diminishing marginal returns (although still positive) were not a great enough incentive at this level, or that subjects were willing to accept a relatively small level of inefficiency. Given this, as an area of future research, it would be worth considering how parameter changes, such

as group size or the steepness of the loss function, impact the level of civic engagement or attitudes towards loss.

Nevertheless, a clear finding remains that the impact of reputation and feedback, despite being materially neutral, has a potent effect on civic engagement and could be used as a mechanism to encourage higher levels where needed. This is important as, given the nature of civic engagement and the problem it is trying to overcome, the standard toolbox of formal or centralised mechanisms for encouraging cooperation are unavailable (or face the same second-order problem of requiring individual effort to maintain).

# References


Abeler, J., Falk, A., Goette, L., & Huffman, D. (2009) Reference Points and Effort Provision. *American Economic Review*, *101*(2), 470-492.

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

Adams, R., & Ferreira, D. (2010) Moderation in Groups: Evidence from Betting on Ice Break-ups in Alaska. *Rev. Econ. Stud., 77*(3), 882-913.

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

Aidt, T. (2009) Corruption, institutions, and economic development. *Oxford Review of Economic Policy, 25*(2), 271–29.

Aimone, J., Iannaccone, L., & Makowsky, M. (2013) Endogenous Group Formation via Unproductive Costs. *Review of Economic Studies, 80*(4), 1215-1236.

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

Alm, J. (2018) What Motivates Tax Compliance. *Journal of Economic Surveys, 33*(2), 353-388.

Anderson, C., & Putterman, L. (2006) Do Non-Strategic Sanctions Obey the Law of Demand? The demand for punishment in the voluntary contribution mechanism. *Games Econ. Behav., 54*(1), 1-24.

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

Andreoni, J. (1993) An Experimental Test of the Public-Goods Crowding-Out Hypothesis. *The American Economic Review, 83*(5), 1317-1327.

Andreoni, J. (1995) Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics, 110*(1), 1–21.

Araujo, F., Carbone, E., Conell-Price, L., Dunietz, M., Jaroszewicz, A., Landsman, R., Lamé, D., Wang, S., & Wilson, A. (2016) The Slider Task: an Example of Restricted Inference on Incentive Effects. *Journal of Economic Science Association, 2*, 1-12.

Auerswald, H., Schmidt, C., Thum, M., & Torsvik, G. (2018) Teams in a Public Goods Experiment with Punishment. *J. Behav. Exp. Econ., 72*, 28-39.

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

Baker, R., Laury, S., & Williams, A. (2008) Comparing Group and Individual Behavior in Lottery-Choice Experiments. *South. Econ. J., 75*, 367-382.

Barron, J., & Gjerde, K. (1997) Peer Pressure in an Agency. *J. Labor Econ., 15*(2), 234-254.

Bateman, I., & Munro, A. (2005) An Experiment on Risky Choice Amongst Households. *Econ. J., 115*, 176-189.

Battaglini, M., Morton, R., & Palfrey, T. (2010) The Swing Voter's Curse in the Laboratory. *Review of Economic Studies, 77*(1), 61-89.

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

Beekman, G., Bulte, E., & Nillesen, E. (2014) Corruption, investments and contributions to public goods: Experimental evidence from rural Liberia. *Journal of Public Economics, 115,* 37-47.

Bergh, D., Connelly, B., Ketchen Jr, D., & Shannon, L. M. (2014) Signalling Theory and Equilibrium in Strategic Management Research: An Assessment and a Research Agenda. *Journal of Management Studies, 51*(8), 1334-1360.

Bergstrom, T., Blume, L., & Varian, H. (1986) On the Private Provision of Public Goods. *Journal of Public Economics*, 29, 25-49.

Bisetti, E., Tengelsen, B., & Zetlin-Jones, A. (2022) Moral Hazard in Remote Teams. *International Economic Review, 63*(4), 1595-1623.

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

Böhm, R., & Theelen, M. (2016) Outcome valence and externality valence framing in public good dilemmas. *Journal of Economic Psychology*, 54, 151-163.

Bolton, P., & Dewatripont, M. (2004) *Contract Theory*. MIT Press.

Bone, J. (1998) Risk-sharing CARA individuals are collectively EU. *Econ. Lett., 58,* 311-317.

Bone, J., Hey, J., & Suckling, J. (1999) Are Groups More (or Less) Consistent than Individuals? *J. Risk Uncertain., 8*, 63-81.

Bone, J., Hey, J., & Suckling, J. (2004) A Simple Risk-Sharing Experiment. *J. Risk Uncertain., 28*, 23-38.

Bonner, B., Baumann, M., Dalal, R. (2002) The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes, 88*(2), 719-736.

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

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

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

Brekke, K., Hauge, K., Lind, J., & Nyborg, K. (2011) Playing with the Good Guys. A Public Good Game with Endogenous Group Formation. *Journal of Public Economics*, 95(9-10), 1111-1118.

Brekke, K., Kverndokk, S., & Nyborg, K. (2003) An economic model of moral motivation. *Journal of Public Economics, 87*(9-10), 1967-1983.

Brosig, J., Weimann, J., & Ockenfels, A. (2003) The Effect of Communication Media on Cooperation. *Ger. Econ. Rev., 4*(2):217-241.

Cagala, T., Glogowsky, U., Grimm, V., Rincke, J., & Cueva, A. T. (2017) Does Corruption Affect the Private Provision of Public Goods? *SSRN Electronic Journal*.

Campos, N. F., Dimova, R., & Saleh, A. (2010) Whither corruption? A quantitative survey of the literature on corruption and growth. *IZA Discussion Papers, No. 5334, Institute for the Study of Labor (IZA), Bonn*.

Campos-Vazquez, R., & Mejia, L. (2016) Does corruption affect cooperation? A laboratory experiment. *Latin American Economic Review, 25*(5), 1-19.

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

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

Cason, T., & Mui, V. (1997) A Laboratory Study of Group Polarisation in the Team Dictator Game. *Econ. J., 107*, 1465-1483.

Cason, T., & Mui, V. (2015) Rich Communication, Social Motivations, and Coordinated Resistance against Divide-and-Conquer: A Laboratory Investigation. *Euro. J. Polit. Econ., 37*, 146-159.

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

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

Charness, G., Cobo-Reyes, R., Jiménez, N., Lacomba, J., & Lagos, F. (2012) The Hidden Advantage of Delegation: Pareto Improvements in a Gift Exchange Game. *Am. Econ. Rev.,* 102(5), 2358-2379.

Charness, G., & Sutter, M. (2012) Groups Make Better Self-Interested Decisions. *J. Econ. Perspect., 26*, 157-176.

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

Cinyabuguma, M., Page, T., & Putterman, L. (2006) Can Second-order Punishment Deter Perverse Punishment? *Exp. Econ., 9*, 265-279.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*, 37-46.

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

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

Connelly, B., Certo, T., Ireland, D., & Reutzel, C. (2011) Signaling Theory: A Review and Assessment. *Journal of Management, 37*(1), 39-67.

Cooper, D., & Kagel, J. (2005) Are Two Heads Better Than One? Team versus Individual Play in Signaling Games. *Am. Econ. Rev.,* 95(3), 477-509.

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

Corgnet, B., Hernan-Gonzalez, R., & Schniter, E. (2015) Why Real Leisure Really Matters: Incentive Effects on Real Effort in the Laboratory. *Experimental Economics, 18*, 284-301.

Cox, J., & Hayne, S. (2006) Barking up the Right Tree: Are Small Groups Rational Agents? *Exp. Econ., 9*, 209-222.

Cox, C., & Stoddard, B. (2015) Framing and Feedback in Social Dilemmas with Partners and Strangers. *Games, 6*(4), 394-412.

Cox, C., & Stoddard, B. (2018) Strategic Thinking in Public Goods Games with Teams. *J. Pub. Econ., 161*, 31-43.

Dal Bó, P., Foster, A., & Putterman, L. (2010) Institutions and Behavior: Experimental Evidence on the Effects of Democracy. *Am. Econ. Rev., 100*(5), 2205-2229.

Dal Bó, P., Foster, A., & Kamei, K. (2019) The Democracy Effect: a Weights-Based Identification Strategy. NEBR Working Paper 25724.

Deci, E., & Ryan, R. (1985) *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum.

Deci, E., & Ryan, R. (2000) The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268.

DeGroot, M. (1974) Reaching a Consensus. *Journal of the American Statistical Association, 69*(345), 118-21.

Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011) The framing of games and the psychology of play. *Games and Economic Behavior, 7*, 459-478.

Denant-Boemont, L., Masclet, D., & Noussair, C. (2007) Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment. *Econ. Theory, 33*, 154-167.

Delarue, A., Van Hootegem, G., Procter, S., & Burridge, M. (2008) Teamworking and Organizational Performance: A Review of Survey-based Research. *Int. J. Manag. Rev., 10*(2), 127-148.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004) The Neural Basis of Altruistic Punishment. *Science, 305*(5688), 1254-1258.

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

Elfenbein, D., Fisman, R., & Mcmanus, B. (2012) Charity as a Substitute for Reputation: Evidence from an Online Marketplace. *Review of Economic Studies*, 79(4), 1441-68.

Eriksson, K., & Strimling, P. (2023) Spontaneous associations and label framing have similar effects in the public goods game. *Judgment and Decision Making, 9*(5), 360-372.

Erkal, N., Gangadharan, L., & Koh, B. (2018) Monetary and non-monetary incentives in real-effort tournaments. *European Economic Review, 101*, 528-545.

Ertan, A., Page, T., & Putterman, L. (2009) Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free Rider Problem. *Eur. Econ. Rev., 53*, 495-511.

Eurofound, Cedefop (2020) European Company Survey 2019: Workplace Practices Unlocking Employee Potential. *European Company Survey 2019*, Publications Office of the European Union, Luxembourg.

Falk, A., Huffman, D., & Mierendorff, K. (2006) Incentive Properties and Political Acceptability of Workfare: Evidence from Real Effort Experiments. Working paper.

Falk A., Fischbacher, U., & Gächter, S. (2013) Living in two neighborhoods - Social interaction effects in the laboratory. *Econ. Inq.*, 51(1), 563-578.

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

Farazmand, A., & De Simone, E. (2022) Corruption, lack of Transparency and the Misuse of Public Funds in Times of Crisis: An introduction. *Public Organization Review, 22*, 497-503.

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

Fehr, E., & Gächter, S. (2002) Altruistic Punishment in Humans. *Nature, 415*, 137-140.

Fehr, E., & Schmidt, K. M. (1999) A Theory of Fairness, Competition, and Cooperation. *Q. J. Econ., 114*(3), 817-868.

Fehr, E., & Schmidt, K. M. (2006) The Economics of Fairness, Reciprocity and Altruism-Experimental Evidence and New Theories. In Kolm, S. C., & Ythier, J. M. (Eds.), *Handbook of the economics of giving, altruism and reciprocity* (Vol. 1.) Foundations (615–691). Elsevier Science.

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

Feri, F., Irlenbusch, B., & Sutter, M. (2010) Efficiency Gains from Team-based Coordination—Large-Scale Experimental Evidence. *Am. Econ. Rev., 100*, 1892-1912.

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

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

Fosgaard, T., Hansen, L., & Wengström, E. (2015) Framing and Misperception in Public Good Experiments. *The Scandinavian Journal of Economics, 199*(2), 435-456.

Fosgaard, T., Hansen, L., & Wengström, E. (2019) Cooperation, framing, and political attitudes. *Journal of Economic Behavior & Organization, 158*, 416-427.

Friedkin, N., & Bullo, F. (2017) How truth wins in opinion dynamics along issue sequences. *PNAS. 114*(43), 11380-11385.

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

Gibbons, R., & Henderson, R. (2013) What do Managers Do? (Ch. 17) ) in *The Handbook of Organizational Economics* (ed. by Gibbons, R. and Roberts, J.), 680-731. Princeton University Press.

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

Gibbons, R., & Roberts, J. (2013) *The Handbook of Organizational Economics*. Princeton University Press.

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

Gintis, H., Smith, E., & Bowles, S. (2001) Costly Signaling and Cooperation. *Journal of Theoretical Biology, 213*(1), 103-119.

Glätzle-Rützler, D., Lergetporer, P., & Sutter, M. (2021) Collective Intertemporal Decisions and Heterogeneity in Groups. *Games Econ. Behav., 130*, 131-147.

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

Grimm, V., & Mengel, F. (2009) Cooperation in viscous populations—Experimental evidence. *Games and Economic Behavior, 66*(1), 202-220.

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

Gründler, K., & Potrafke, N. (2019) Corruption and economic growth: New empirical evidence. *European Journal of Political Economy, 60*, 101810.

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

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

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

Hagen, E., & Bryant, G. (2003) Music and Dance as a Coalition Signaling System. *Human Nature, 14*(1), 21-51.

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

Harrison, G., Lau, M., Rutström, E., & Tarazona-Gomez, M. (2013) Preferences over Social Risk. *Oxford Econ. Pap., 65*(1), 25-46.

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

Hawkes, K., & Bliege, R. (2002) Showing off, Handicap Signaling, and the Evolution of Men's Work. *Evolutionary Anthropology, 11*(2), 58-67.

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

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

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

Iannaccone, L. (1992) Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives. *Journal of Political Philosophy, 100*(2), 271-291.

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

Isaac, M., & Walker, J. (1988) Communication and Free-riding Behavior: the Voluntary Contributions Mechanism. *Economic Inquiry, 26*, 586 – 608.

Iturbe-Ormaetxe, I., Ponti, G., Tomás, J., & Ubeda, L. (2011) Framing effects in public goods: Prospect Theory and experimental evidence. *Games and Economic Behavior, 72*, 49-447.

Jakee, K., & Sun, G-Z. (2006) Is Compulsory Voting More Democratic? *Public Choice, 129*, 61-75.

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

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

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

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

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

Kamei, K. (2021) Teams do Inflict Costly Third Party Punishment as Individuals do: Experimental Evidence. *Games, 12*(1), 1-11.

Kamei, K., & Markussen, T. (forthcoming) Free Riding and Workplace Democracy – Heterogeneous Task Preferences and Sorting. *Management Science*.

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

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

Kamei, K., Putterman, L., & Tyran, J-R. (2019) Civic Engagement as a Second-Order Public Good. Working Papers 2019_05, Department of Economics, University of Durham.

Kamei, K., & Tabero, K. (2022) The Individual-Team Discontinuity Effect on Institutional Choices: Experimental Evidence in Voluntary Public Goods Provision, Keio-IES Discussion Paper Series 2022-015, Institute for Economics Studies, Keio University.

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

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

Keser, C., Markstädter, A., & Schmidt, M. (2017) Mandatory minimum contributions, heterogeneous endowments and voluntary public-good provision. *Games and Economic Behavior, 101*, 291-310.

Kocher, M., Cherry, T., Kroll, S., Netzer, R., & Sutter, M. (2008) Conditional Cooperation on Three Continents. *Econ. Lett., 101*(3), 175-178.

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

Kocher, M., Strauß, S., & Sutter, M. (2006) Individual or Team Decision-Making—Causes and Consequences of Self-selection. *Games Econ. Behav.*, 56(2), 259-270.

Kosfeld, M., Okada, A., & Riedl, A. (2009) Institution Formation in Public Goods Games. *Am. Econ. Rev., 99*(4), 1335-1355.

Kreps, D., Milgrom, P., Roberts, J., & Wilson, R. (1982) Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. *J. Econ. Theory, 27*(2), 245-252.

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

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

Landis, R., & Koch, G. (1977) An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics, 33*(2), 363-374.

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

Laurence, I. (1992) Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives. *Journal of Political Economy*, 100(2), 271-291.

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

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

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

Leibbrandt, A., & Sääksvuori, L. (2012) Communication in Intergroup Conflicts. *Eur. Econ. Rev., 56*(6), 1136-1147.

Lugovskyy, V., Puzzello, D., Sorensen, A., Walker, J., & Williams, A. (2017) An Experimental Study of Finitely and Infinitely Repeated Linear Public Goods Games. *Games Econ. Behav., 102,* 286-302.

Luhan, W., Kocher, M., & Sutter, M. (2009) Group polarization in the team dictator game reconsidered. *Experimental Economics, 12*, 26-41.

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

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. (2003) Monetary and non-monetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review, 93*(1), 366 – 380.

McLachlan, G., & Peel, D. (2000) *Finite Mixture Models*. Wiley.

Milinski, M., Sommerfeld, R., Krambeck, H-J., Reed, F., & Marotzke, J. (2007) The collective-risk social dilemma and the prevention of simulated dangerous climate change. *PNAS, 105*(7), 2291-2294.

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

Morse, F. (2015, September) Facebook dislike button: a short history. *BBC Newsbeat*. https://www.bbc.co.uk/news/newsbeat-34269663

Müller, W., & Tan, F. (2013) Who Acts More Like a Game Theorist? Group and Individual Play in a Sequential Market Game and the Effect of the Time Horizon. *Games Econ. Behav., 82*, 658-674.

Mungiu-Pippidi, A. (2013) Controlling Corruption through Collective Action. *Journal of Democracy, 24*(1), 101-115.

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

Niederle, M., & Vesterlund, L. (2007) Do Women Shy Away from Competition? Do Men Compete Too Much? *Quarterly Journal of Economics, 122*(3), 1067-1101.

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

Page, T., Putterman, L., & Unel, B. (2005) Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency. *Economic Journal, 115*, 1032 – 1053.

Park, E. (2000) Warm-glow versus cold-prickle: a further experimental study of framing effects on free-riding. *Journal of Economic Behavior & Organization, 43*(4), pp. 405-421.

Pencavel, J. (2001) *Worker Participation: Lessons from the Worker Co-Ops of the Pacific Northwest*. Russell Sage Foundation.

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

Prather, L., & Middleton, K. (2002) Are N+1 Heads Better than One?: The Case of Mutual Fund Managers. *J Econ Behav. Organ., 47*(1), 103-120.

Prendergast, C. (1999) The Provision of Incentives in Firms. *Journal of Economic Literature, 37*(1), 7-63.

Rege, M., & Telle, K. (2004) The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics, 88*(7-8), 1626-1644.

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

Rockenbach, B., Sadrieh, A., & Mathauschek, B. (2007) Teams Take the Better Risks. *J. Econ. Behav. Organ., 63*, 412-422.

Samuelson, P. (1954) The Pure Theory of Public Expenditure. *Review of Economics and Statistics, 36*(4), 387-389.

Schopler, J., Insko, C. A., Graetz, K. A., Drigotas, S. M., & Smith, V. (1991) The generality of the individual-group discontinuity effect: Variations in positivity-negativity of outcomes, players' relative power, and magnitude of outcomes. *Pers. Soc. Psychol. Bull*., 17(6), 612-624.

Schulze, C., & Newell, B. (2016) More heads choose better than one: Group decision making can eliminate probability matching. *Psychonomic Bulletin & Review, 23*, 907-914.

Sobel, J. (2005) Interdependent Preferences and Reciprocity. *J. Econ. Lit, 43*(2), 392-436.

Sonnemans, J., Schram, A., & Offerman, T. (1998) Public good provision and public bad prevention: The effect of framing. *Journal of Economic Behavior & Organization, 34*(1), 143-161.

Shupp, R., & Williams, A. (2008) Risk Preference Differentials of Small Groups and Individuals. *Econ. J., 118*, 258-283.

Smith, E., & Bliege, R. (2000) Turtle Hunting and Tombstone Opening: Public Generosity as Costly Signaling. *Evolution and Human Behavior, 21*(4), 245-261.

Smith, E., & Bliege, R. (2005) Costly Signaling and Prosocial Behavior. In: *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (eds. Gintis, H., Bowles, S., Boyd, R., & Fehr, E.), 115-148. MIT Press.

Sosis, R., & Alcorta, C. (2003) Signaling, Solidarity, and the Sacred: The Evolution of Religious Behavior. *Evolutionary Anthropology, 12*(6), 264-274.

Sosis, R., & Bressler, E. (2003) Cooperation and Commune Longevity: A Test of the Costly Signaling Theory of Religion. *Cross-Cultural Research, 37*(2), 211-239.

Spence, M. (1973) Job Market Signaling. *Quarterly Journal of Economics, 87*(3), 355-374.

Sutter, M. (2005) Are Four Heads Better than Two? An Experimental Beauty-Contest Game with Teams of Different Size. *Econ. Lett., 88*, 41-46.

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

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

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

Sutter, M., Kocher, M., Strauß, S. (2009) Individuals and Teams in Auctions. *Oxford Econ. Pap., 61*(2), 380-94.

Teremetskyi, V., Duliba, Y., Kroitor, V., Korchak, N., & Makarenko, O. (2020) Corruption and strengthening anti-corruption efforts in healthcare during the pandemic of Covid-19. *Medco-Legal Journal, 89*(1), 25-28.

Thöni, C., & Volk, S. (2018) Conditional cooperation: Review and Refinement. *Economics Letters, 171*, 37-40.

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

Tyran, J-R., & Feld, L. (2006) Achieving Compliance when Legal Sanctions are Non-deterrent. *Scandinavian Journal of Economics*, *108*(1), 135-156.

van Elten, J., & Penczynski, S. (2020) Coordination games with asymmetric payoffs: An experimental study with intra-group communication. *Journal of Economic Behavior & Organization*, *169*, 158-188.

van Vugt, M., & Hardy, C. (2009) Cooperation for reputation: Wasteful contributions as costly signals in public goods. *Group Processes & Intergroup Relations*, *33*(1), 101-111.

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003) Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychol. Bull., 129*(5), 698-722.

Zelmer, J. (2003) Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics*, *6*, 299-310.

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

# Appendix A: Appendix for Chapter 2

## Appendix A.A: Non-parametric test results

[Within-groups comparison:]

### A. Effects of voting

#### (1) Contribution

| | Avg. contribution based on all data | | | Avg. contribution under a given sanction scheme | | | | |
|---|---|---|---|---|---|---|---|---|
| | (i) Phase 1 | (ii) Phases 2-6 | *p* (two-sided) for $H_0$: (i) = (ii) | (iii) FS, Phases 2-6 | *p* (two-sided) for $H_0$: (i) = (iii) | (iv) IS, Phases 2-6 | *p* (two-sided) for $H_0$: (i) = (iv) | *p* (two-sided) for $H_0$: (iii) = (iv)[#1] |
| **[Individual treatments:]** | | | | | | | | |
| I-No | 12.92 | 9.64 | 0.0414** | --- | --- | --- | --- | --- |
| | (1.45) | (1.83) | | | | | | |
| Individual Voting (I-Voting-M, I-Voting-ST) | 10.60 | 12.68 | 0.2914 | 10.24 | 0.8313 | 15.04 | 0.2790 | 0.2330 |
| | (0.96) | (1.26) | | (1.32) | | (1.43) | | |
| I-Voting-M | 11.92 | 11.57 | 0.9292 | 9.69 | 0.4838 | 13.66 | 0.9594 | 0.7353 |
| | (1.28) | (1.79) | | (1.98) | | (1.98) | | |
| I-Voting-ST | 9.27 | 13.80 | 0.1549 | 10.88 | 0.8590 | 16.23 | 0.2026 | 0.1614 |
| | (1.37) | (1.80) | | (1.88) | | (2.12) | | |
| **[Team treatments:]** | | | | | | | | |
| T-No | 13.26 | 10.04 | 0.0096*** | --- | --- | --- | --- | --- |
| | (0.92) | (1.13) | | | | | | |
| Team Voting (T-Voting-M, T-Voting-ST) | 12.53 | 17.67 | 0.0001*** | 18.02 | 0.0002*** | 17.30 | 0.0166** | 0.1054 |
| | (1.04) | (0.63) | | (0.58) | | (1.10) | | |
| T-Voting-M | 12.81 | 16.53 | 0.0128** | 16.87 | 0.0209** | 16.28 | 0.0827* | 0.0966* |
| | (1.62) | (1.15) | | (1.14) | | (1.53) | | |
| T-Voting-ST | 12.24 | 18.80 | 0.0033*** | 18.81 | 0.0051*** | 18.78 | 0.1282 | 0.7532 |
| | (1.37) | (0.33) | | (0.37) | | (1.54) | | |

#### (2) Payoff

| | Avg. payoff based on all data | | | Avg. payoff under a given sanction scheme | | | | |
|---|---|---|---|---|---|---|---|---|
| | (i) Phase 1 | (ii) Phases 2-6 | *p* (two-sided) for $H_0$: (i) = (ii) | (iii) FS, Phases 2-6 | *p* (two-sided) for $H_0$: (i) = (iii) | (iv) IS, Phases 2-6 | *p* (two-sided) for $H_0$: (i) = (iv) | *p* (two-sided) for $H_0$: (iii) = (iv)[#1] |
| **[Individual treatments:]** | | | | | | | | |
| I-No | 30.34 | 27.71 | 0.0414** | --- | --- | --- | --- | --- |
| | (1.16) | (1.47) | | | | | | |
| Individual Voting (I-Voting-M, I-Voting-ST) | 28.479 | 25.27 | 0.0575* | 23.38 | 0.0086*** | 27.09 | 0.0304** | 0.1252 |
| | (0.77) | (1.48) | | (1.12) | | (2.60) | | |
| I-Voting | 29.54 | 23.79 | 0.0208** | 22.88 | 0.0357** | 24.81 | 0.0218** | 0.0280** |
| | (1.03) | (2.09) | | (1.87) | | (3.49) | | |
| I-Voting-ST | 27.42 | 26.75 | 0.7897 | 23.97 | 0.1731 | 29.07 | 0.5076 | 0.8886 |
| | (1.10) | (2.08) | | (1.43) | | (3.85) | | |
| **[Team treatments:]** | | | | | | | | |
| T-No | 30.61 | 28.03 | 0.0096*** | --- | --- | --- | --- | --- |
| | (0.73) | (0.91) | | | | | | |
| Team Voting (T-Voting-M, T-Voting-ST) | 30.02 | 29.90 | 0.9353 | 29.71 | 0.8092 | 30.10 | 0.1701 | 0.1252 |
| | (0.83) | (0.96) | | (0.60) | | (2.82) | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| T-Voting-M | 30.25 | 29.07 | 0.5337 | 28.38 | 0.3743 | 29.56 | 0.1823 | 0.1386 |
| | (1.29) | (1.68) | | (1.19) | | (3.50) | | |
| T-Voting-ST | 29.79 | 30.73 | 0.4236 | 30.63 | 0.5076 | 30.89 | 0.7353 | 0.9165 |
| | (1.10) | (0.96) | | (0.32) | | (5.07) | | |

*Notes*: two-sided *p*-values. The numbers in the parentheses are standard errors. Wilcoxon signed rank tests based on observations of group means. Individual Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. [#1] Only groups that had played under both the FS and IS schemes were used. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: (a) Decision-making units, whether individuals or teams, experienced significantly weaker contributions and accordingly lower payoffs in Phases 2 to 6 than in Phase 1 when sanctions were not available (the I-No and T-No treatments).

(b) Individuals prevented cooperation norms from declining in Phases 2 to 6 when they could vote on sanctions (I-Voting-M, I-Voting-ST), unlike the I-No treatment. However, the individuals experienced significantly lower payoffs in Phases 2 to 6 than in Phase 1 due to punishment loss.

(c) The effects of voting were stronger among teams than individuals. Specifically, teams achieved strong cooperation norms when they could vote on sanctions (T-Voting-M and T-Voting-ST). They did not experience a drop in payoffs in Phases 2 to 6 relative to Phase 1.

[Between-groups treatment comparison:]

## B. Average contributions in phase 1 [randomization check]

| | Pooled data | | Each treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Indiv Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Indiv Voting | --- | 0.2007 | --- | --- | --- | --- | --- | --- |
| I-No | 0.1882 | 0.7051 | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.5179 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.1095 | 0.1999 | --- | --- | --- | --- |
| T-No | 0.0715* | 0.7051 | 0.9310 | 0.3558 | 0.0312** | --- | --- | --- |
| T-Voting-M | --- | --- | 0.7818 | 0.7674 | 0.1005 | 0.8777 | --- | --- |
| T-Voting-ST | --- | --- | 0.7119 | 0.9215 | 0.1227 | 0.6225 | 0.5327 | --- |

*Notes*: two-sided *p*-values. Mann-Whitney tests based on observations of group means. Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: No treatment differences were found for all comparisons except one comparison (T-No versus I-Voting-ST).

## C. Average contributions from phases 2 to 6

| | Pooled data | | Each treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Indiv Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Indiv Voting | --- | 0.0074*** | --- | --- | --- | --- | --- | --- |
| I-No | 0.2273 | 0.0016*** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.5180 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.1569 | 0.3754 | --- | --- | --- | --- |
| T-No | 0.3305 | 0.0000*** | 0.6861 | 0.7583 | 0.1757 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.0097*** | 0.0611* | 0.3085 | 0.0021*** | --- | --- |
| T-Voting-ST | --- | --- | 0.0051*** | 0.0052*** | 0.0706* | 0.0001*** | 0.2000 | --- |

*Notes*: two-sided *p*-values. Mann-Whitney tests based on observations of group means. Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: Voting on sanction schemes had no effects on improving cooperation when the decision-making units were individuals. By contrast, voting had strong positive effects when the decision-making units were teams.

## D. Average loss per period in phases 2 to 6 due to having the sanction schemes

| | | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | Avg. loss per period[#1] | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 5.202 | 0.3661 | --- | --- | --- | --- |
| I-Voting-M | 5.553 | --- | --- | --- | --- | --- |
| I-Voting-ST | 4.851 | --- | 0.6224 | --- | --- | --- |
| Team Voting | 4.477 | --- | --- | --- | --- | --- |
| T-Voting-M | 4.295 | --- | 0.2642 | 0.4905 | --- | --- |
| T-Voting-ST | 4.658 | --- | 0.5767 | 0.8696 | 0.6695 | --- |

*Notes*: The per unit losses include any reductions of payoffs under the IS scheme, and sanction payment, cost of imposing sanctions and administrative cost under the FS scheme. Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. [#1] Average loss is equal to all losses in a given treatment divided by (60 × the number of groups). Here, 60 = 3 decision-making units/group × 20 periods. [#2] Mann-Whitney tests based on observations of group means. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: Average per period loss a decision-making unit incurred due to sanction schemes did not differ by treatment.

## E. Average loss per decision-making unit and period under the IS scheme in phases 2 to 6

| | Avg. per unit loss in the IS scheme per period | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 5.579 | 0.5391 | --- | --- | --- | --- |
| I-Voting-M | 6.308 | --- | --- | --- | --- | --- |
| I-Voting-ST | 4.947 | --- | 0.4057 | --- | --- | --- |
| Team Voting | 4.234 | --- | --- | --- | --- | --- |
| T-Voting-M | 3.708 | --- | 0.2050 | 0.9719 | --- | --- |
| T-Voting-ST | 4.998 | --- | 0.6963 | 0.9611 | 0.6833 | --- |

*Notes*: The per unit losses include any reduction of payoffs (including the punisher and the punished) under the IS scheme. Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments [#1] Mann-Whitney tests based on observations of group means. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

## F. Realized group sanction rates under the FS scheme in phases 2 to 6

| | Average realized group sanction rates | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 0.24 | 0.0116** | --- | --- | --- | --- |
| I-Voting-M | 0.11 | --- | --- | --- | --- | --- |
| I-Voting-ST | 0.39 | --- | 0.7345 | --- | --- | --- |
| Team Voting | 0.79 | --- | --- | --- | --- | --- |
| T-Voting-M | 0.64 | --- | 0.0747* | 0.3984 | --- | --- |
| T-Voting-ST | 0.89 | --- | 0.0099*** | 0.0651* | 0.1106 | --- |

*Notes*: Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. [#1] Mann-Whitney tests based on observations of group means. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

| | % of cases in which group selected a sanction rate of 0.0 in phases 2 to 6 | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 64.35 % | 0.0080*** | --- | --- | --- | --- |
| I-Voting-M | 70.69 % | --- | --- | --- | --- | --- |
| I-Voting-ST | 57.00 % | --- | 0.8841 | --- | --- | --- |
| Team Voting | 19.20 % | --- | --- | --- | --- | --- |
| T-Voting-M | 29.09 % | --- | 0.0667* | 0.2632 | --- | --- |
| T-Voting-ST | 14.39 % | --- | 0.0083*** | 0.0670* | 0.3188 | --- |

*Notes*: Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. [#1] Group-level Mann-Whitney tests. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

| | % of cases in which group selected a sanction rate of 1.2 in phases 2 to 6 | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 12.96% | 0.0319** | --- | --- | --- | --- |
| I-Voting-M | 0.00% | --- | --- | --- | --- | --- |
| I-Voting-ST | 28.00% | --- | 0.0826* | --- | --- | --- |
| Team Voting | 49.55% | --- | --- | --- | --- | --- |
| T-Voting-M | 31.52% | --- | 0.0169** | 0.6277 | --- | --- |
| T-Voting-ST | 62.12% | --- | 0.0110** | 0.3534 | 0.1745 | --- |

*Notes*: Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. [#1] Group-level Mann-Whitney tests. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

## G. Average payoffs in phases 2 to 6

| | Pooled data | | Each treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Indiv Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Indiv Voting | --- | 0.0514* | --- | --- | --- | --- | --- | --- |
| I-No | 0.2343 | 0.2273 | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.0848* | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.7583 | 0.4118 | --- | --- | --- | --- |
| T-No | 0.3130 | 0.0661* | 0.6861 | 0.0848* | 1.0000 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.4790 | 0.0710* | 0.6224 | 0.3559 | --- | --- |
| T-Voting-ST | --- | --- | 0.1757 | 0.0138** | 0.4905 | 0.0267** | 0.3754 | --- |

*Notes*: two-sided *p*-values. Mann-Whitney tests based on observations of group means. Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: Voting had no effects on improving payoff when the decision-making units were individuals. By contrast, voting had positive effects when the decision-making units were teams and punishment is strong.

## H. Average contribution and payoff by Scheme in Phases 2 to 6

| | Average contribution under FS in phases 2-6 | | | | | Average payoff under FS in phases 2-6 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Individual Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST | Individual Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Team Voting | 0.0003*** | --- | --- | --- | --- | 0.0004*** | --- | --- | --- | --- |
| I-Voting-ST | --- | 0.7003 | --- | --- | --- | --- | 0.7728 | --- | --- | --- |
| T-Voting-M | --- | 0.0124** | 0.0305** | --- | --- | --- | 0.0209** | 0.0243** | --- | --- |
| T-Voting-ST | --- | 0.0100** | 0.0055*** | 0.5401 | --- | --- | 0.0129** | 0.0055*** | 0.6242 | --- |

| | Average contribution under IS in phases 2-6 | | | | | Average payoff under IS in phases 2-6 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Individual Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST | Individual Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Team Voting | 0.0554* | --- | --- | --- | --- | 0.3061 | --- | --- | --- | --- |
| I-Voting-ST | --- | 0.4963 | --- | --- | --- | --- | 0.3258 | --- | --- | --- |
| T-Voting-M | --- | 0.1389 | 0.6215 | --- | --- | --- | 0.1392 | 0.9719 | --- | --- |
| T-Voting-ST | --- | 0.0248** | 0.1568 | 0.5551 | --- | --- | 0.2831 | 0.7325 | 0.7857 | --- |

*Notes*: two-sided *p*-values. Mann-Whitney tests based on observations of group means. Individual Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively. Average contributions and payoffs under the FS or IS scheme can be found in Part A (columns (iii) and (iv)).

Result: (a) Average contributions and payoffs under the FS scheme were both significantly larger for teams than for individuals. (b) Average contributions and payoffs under the IS scheme were also larger for teams than for individuals. However, the differences were small. For example, the average payoffs were not significantly different between the teams and individuals, regardless of the size of punishment strength.

## I. Average contributions, phase by phase

Phase 2

| | Pooled data | | Each treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0011*** | --- | --- | --- | --- | --- | --- |
| I-No | 0.7051 | 0.0582* | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.1659 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.4598 | 0.0451** | --- | --- | --- | --- |
| T-No | 0.6138 | 0.0018*** | 0.8173 | 0.1027 | 0.4413 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.2058 | 0.0138** | 0.2240 | 0.0524* | --- | --- |
| T-Voting-ST | --- | --- | 0.0488** | 0.0009*** | 0.0323** | 0.0007*** | 0.2118 | --- |

Phase 3

| | Pooled data | | Each treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0273** | --- | --- | --- | --- | --- | --- |
| I-No | 0.7578 | 0.0296** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.9754 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.5742 | 0.6677 | --- | --- | --- | --- |
| T-No | 0.2481 | 0.0004*** | 0.4699 | 0.3559 | 0.2942 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.2635 | 0.2629 | 0.6165 | 0.0192** | --- | --- |
| T-Voting-ST | --- | --- | 0.0092*** | 0.0030*** | 0.1047 | 0.0002*** | 0.0614* | --- |

Phase 4

| | Pooled data | | Each treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0019*** | --- | --- | --- | --- | --- | --- |
| I-No | 0.6905 | 0.0015*** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.8533 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.6193 | 0.6677 | --- | --- | --- | --- |
| T-No | 0.8710 | 0.0000*** | 0.6032 | 0.8777 | 0.6658 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.0177** | 0.0311** | 0.1260 | 0.0016*** | --- | --- |
| T-Voting-ST | --- | --- | 0.0029*** | 0.0038*** | 0.0323** | 0.0001*** | 0.2094 | --- |

| | Phase 5 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0119** | --- | --- | --- | --- | --- | --- |
| I-No | 0.1477 | 0.0013*** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.5792 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.0543* | 0.1555 | --- | --- | --- | --- |
| T-No | 0.2952 | 0.0002*** | 0.5635 | 0.9020 | 0.0557* | --- | --- | --- |
| T-Voting-M | --- | --- | 0.0155** | 0.0375** | 0.7621 | 0.0066*** | --- | --- |
| T-Voting-ST | --- | --- | 0.0023*** | 0.0030*** | 0.0856* | 0.0003*** | 0.1405 | --- |

| | Phase 6 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0104** | --- | --- | --- | --- | --- | --- |
| I-No | 0.0097*** | 0.0001*** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.1315 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.0035*** | 0.0973* | --- | --- | --- | --- |
| T-No | 0.0198** | 0.0000*** | 0.4884 | 0.2300 | 0.0054*** | --- | --- | --- |
| T-Voting-M | --- | --- | 0.0006*** | 0.0053*** | 0.3344 | 0.0001*** | --- | --- |
| T-Voting-ST | --- | --- | 0.0009*** | 0.0068*** | 0.4963 | 0.0000*** | 0.7337 | --- |

*Notes*: The numbers are *p*-values (two-sided). Mann-Whitney tests based on observations of group means. Individual Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: (a) Voting on sanctions did not improve cooperation, even if subjects gained experience, when the decision-making units were individuals and the punishment strength was modest (see the comparisons between I-Voting-M versus I-No treatment). By contrast, subjects learned to cooperate with strong punishment in the I-Voting-ST treatment from phase to phase. The impact was significant at the 10% and 1% levels in phases 5 and 6, respectively, in the I-Voting-ST treatment.

(b) Regardless of whether the punishment strength was strong, voting on sanctions improved cooperation from the very first voting phase, and then the positive effects were sustained throughout the entire experiment when the decision-making units were teams.

## J. Average payoff, phase by phase

| | Phase 2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0367** | --- | --- | --- | --- | --- | --- |
| I-No | 0.0661* | 0.8008 | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.1238 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.1095 | 0.9738 | --- | --- | --- | --- |
| T-No | 0.0049*** | 0.2639 | 0.8173 | 0.0097*** | 0.0267** | --- | --- | --- |
| T-Voting-M | --- | --- | 0.5382 | 0.2786 | 0.3754 | 0.1756 | --- | --- |
| T-Voting-ST | --- | --- | 0.8535 | 0.0235** | 0.1077 | 0.5795 | 0.4502 | --- |

| | Phase 3 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0624* | --- | --- | --- | --- | --- | --- |
| I-No | 0.1534 | 0.4893 | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.0519* | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.6206 | 0.3747 | --- | --- | --- | --- |
| T-No | 0.2958 | 0.1482 | 0.4699 | 0.1239 | 0.8054 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.8766 | 0.1140 | 0.7164 | 0.4980 | --- | --- |
| T-Voting-ST | --- | --- | 0.3050 | 0.0250** | 0.3032 | 0.0737* | 0.3004 | --- |

| | Phase 4 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0122** | --- | --- | --- | --- | --- | --- |
| I-No | 0.2783 | 0.1095 | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.2673 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.4578 | 0.5983 | --- | --- | --- | --- |
| T-No | 0.3484 | 0.0214** | 0.6032 | 0.2954 | 0.5793 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.2653 | 0.0651* | 0.1842 | 0.1564 | --- | --- |
| T-Voting-ST | --- | --- | 0.1068 | 0.0208** | 0.1250 | 0.0124** | 0.5465 | --- |

| | Phase 5 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0523* | --- | --- | --- | --- | --- | --- |
| I-No | 0.6910 | 0.0454** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.2179 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.5778 | 0.1299 | --- | --- | --- | --- |
| T-No | 0.4707 | 0.0484** | 0.5635 | 0.1396 | 0.8053 | --- | --- | --- |
| T-Voting-M | --- | --- | 0.2656 | 0.1066 | 0.6907 | 0.3883 | --- | --- |
| T-Voting-ST | --- | --- | 0.0215** | 0.0226** | 0.2552 | 0.0123** | 0.3671 | --- |

| | Phase 6 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pooled data | | Each treatment | | | | | |
| | Individual Voting | Team Voting | I-No | I-Voting-M | I-Voting-ST | T-No | T-Voting-M | T-Voting-ST |
| Individual Voting | --- | 0.0825* | --- | --- | --- | --- | --- | --- |
| I-No | 0.3383 | 0.0010*** | --- | --- | --- | --- | --- | --- |
| I-Voting-M | --- | --- | 0.9754 | --- | --- | --- | --- | --- |
| I-Voting-ST | --- | --- | 0.1078 | 0.0598* | --- | --- | --- | --- |
| T-No | 0.3036 | 0.0002*** | 0.4884 | 0.8777 | 0.0558* | --- | --- | --- |
| T-Voting-M | --- | --- | 0.006*** | 0.0084*** | 0.9734 | 0.0011*** | --- | --- |
| T-Voting-ST | --- | --- | 0.0042*** | 0.0137** | 0.8162 | 0.0017*** | 0.6194 | --- |

*Notes*: The numbers are *p*-values (two-sided). Mann-Whitney tests based on observations of group means. Individual Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: (a) Parallel to the results in Part E, voting on sanctions did not improve payoff, even if subjects gained experience, when the decision-making units were individuals and the punishment strength was modest.

(b) When the punishment strength was strong, voting on sanctions improved payoffs of teams gradually from phase to phase, and the positive effects were significant for phases 4-6 (the T-Voting-ST treatment). When the punishment strength was weak, voting did not have a positive effect on improving payoffs before phase 6, but it had a positive effect in phase 6 (the T-Voting-M treatment).

## K. Scheme Choices in phases 2 to 6

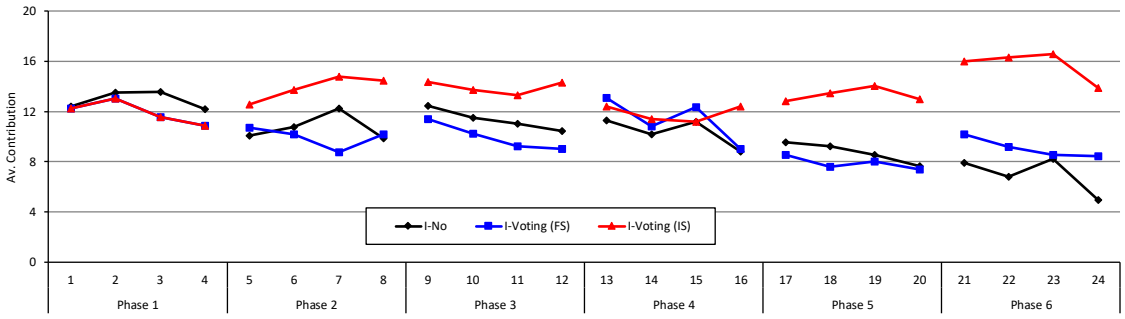
| | (a) Percentages of Times that Decision-Making Units Voted for the IS Scheme | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 50.30% | 0.7149 | --- | --- | --- | --- |
| I-Voting-M | 47.27% | --- | --- | --- | --- | --- |
| I-Voting-ST | 53.33% | --- | 0.5756 | --- | --- | --- |
| Team Voting | 54.55% | --- | --- | --- | --- | --- |
| T-Voting-M | 63.03% | --- | 0.3717 | 0.3540 | --- | --- |
| T-Voting-ST | 46.06% | --- | 0.7670 | 0.6204 | 0.1451 | --- |

| | (b) Percentages of Times that Groups Selected the IS Scheme | Two-sided *p*-values (treatment differences)[#1] | | | | |
|---|---|---|---|---|---|---|
| | | Team Voting | I-Voting-M | I-Voting-ST | T-Voting-M | T-Voting-ST |
| Indiv Voting | 50.91% | 0.7825 | --- | --- | --- | --- |
| I-Voting-M | 47.27% | --- | --- | --- | --- | --- |
| I-Voting-ST | 54.55% | --- | 0.6614 | --- | --- | --- |
| Team Voting | 49.09% | --- | --- | --- | --- | --- |
| T-Voting-M | 58.18% | --- | 0.5634 | 0.7096 | --- | --- |
| T-Voting-ST | 40.00% | --- | 0.4409 | 0.3477 | 0.1652 | --- |

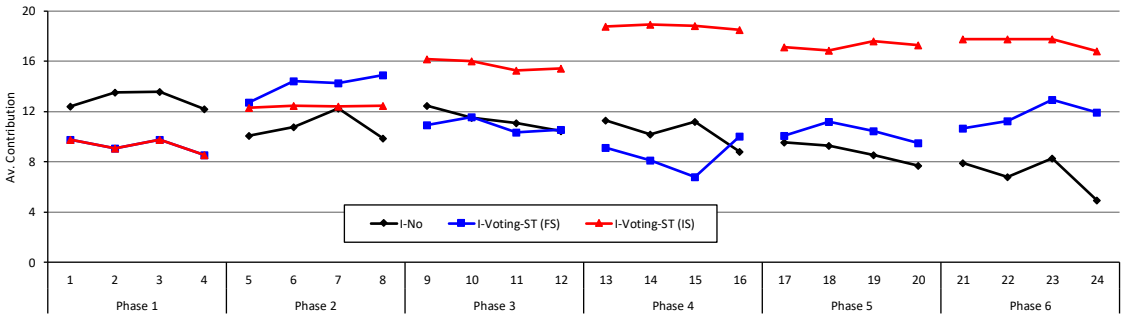
*Notes*: Indiv Voting includes the I-Voting-M and I-Voting-ST treatments. Team Voting includes the T-Voting-M and T-Voting-ST treatments. [#1] Group-level Mann-Whitney tests. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

## Appendix A.B: Additional Figures and Tables

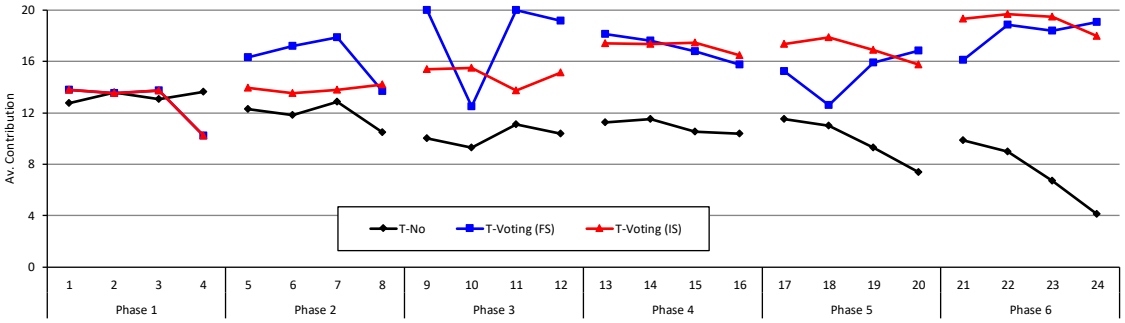
**Figure B.1:** *Average Contribution (Period by Period) by Scheme*



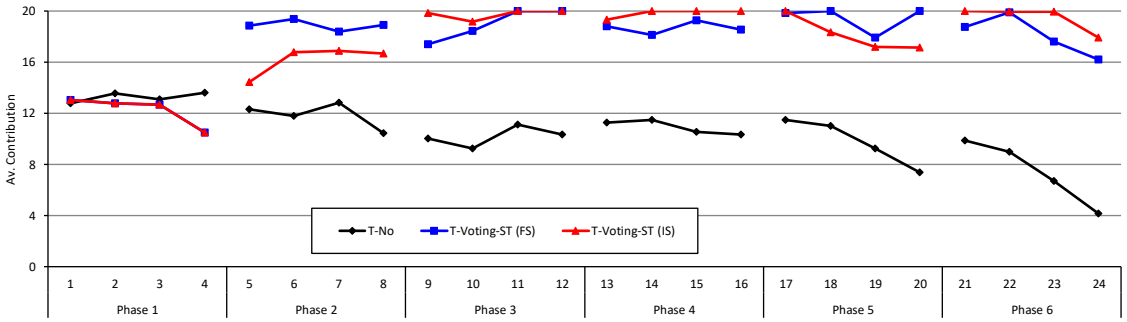

(a) I-Voting-M treatment




(b) I-Voting-ST treatment

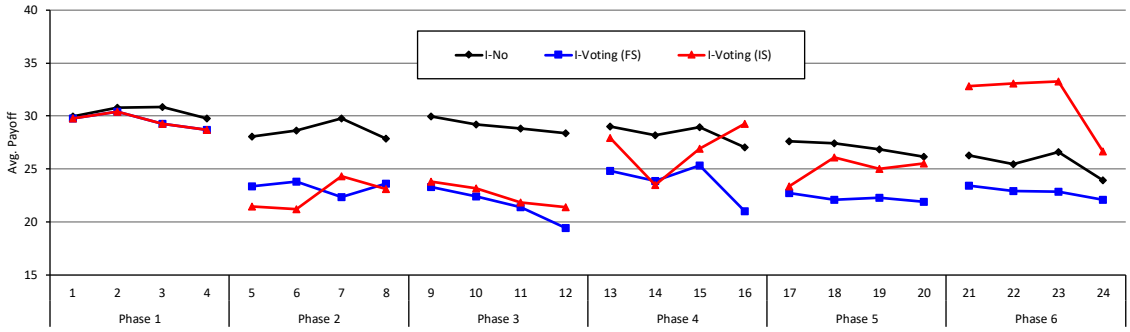



(c) T-Voting-M treatment

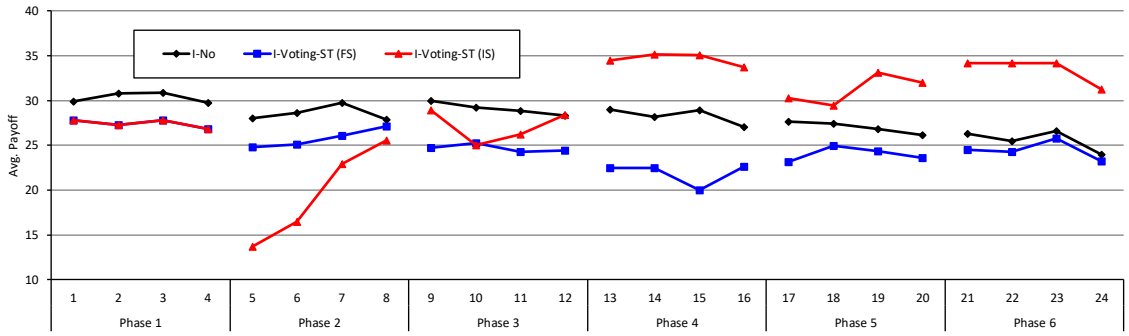



(d) T-Voting-ST treatment

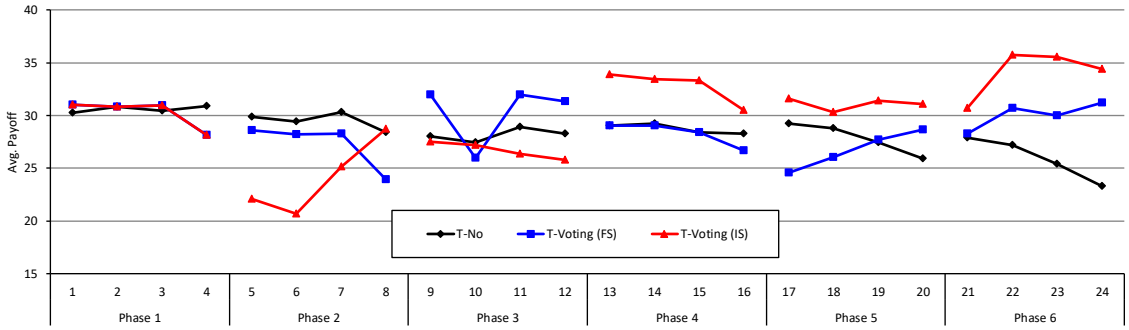
**Figure B.2:** *Average Payoff (Period by Period) by Scheme*



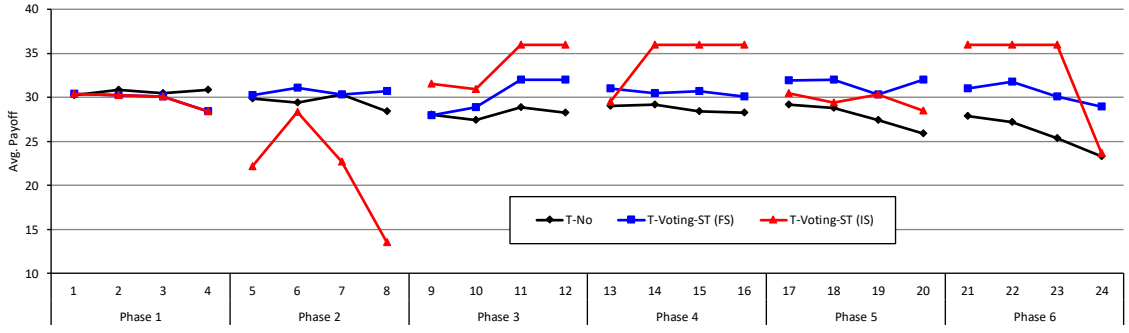

(a) I-Voting-M treatment




(b) I-Voting-ST treatment

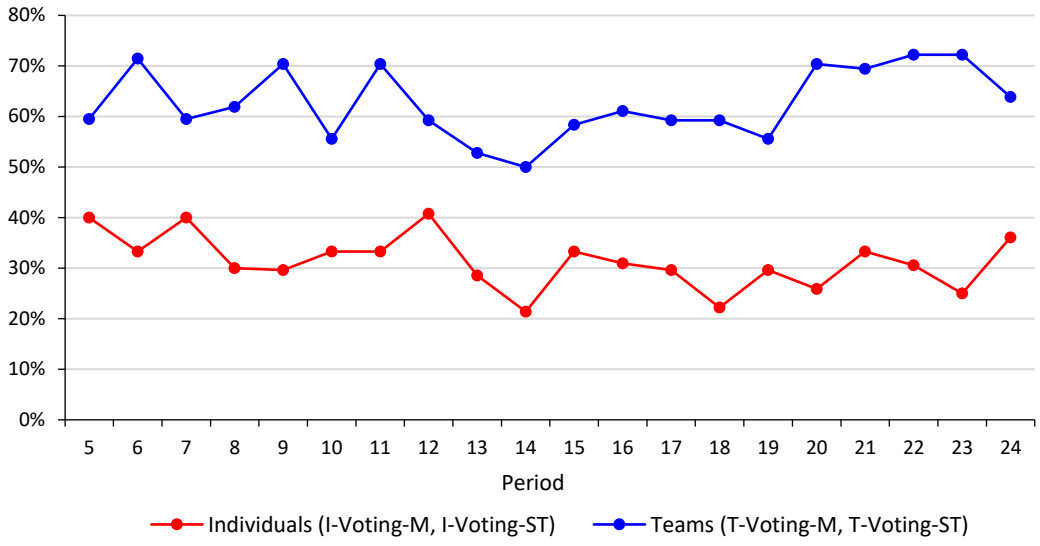



(c) T-Voting-M treatment

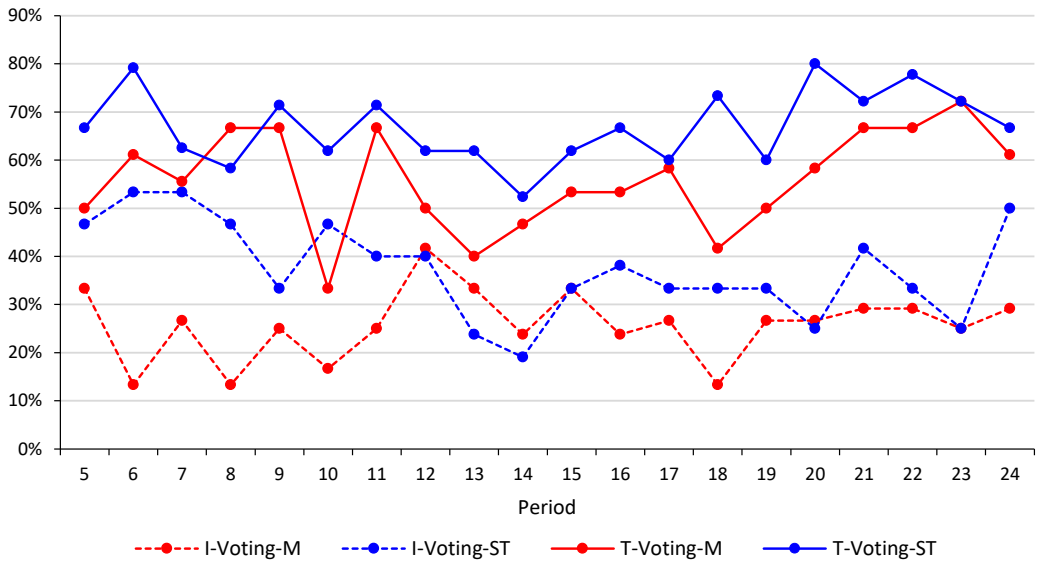



(d) T-Voting-ST treatment

**Figure B.3:** *Percentage of Decision-Making Units that Voted for a Deterrent Sanction Rate (greater than or equal to 0.4), Period by Period*



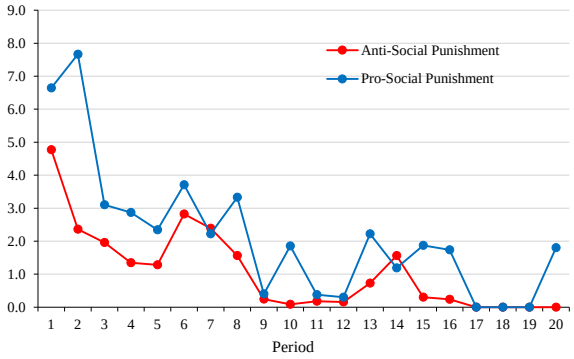

(a) Average Percentage using Pooled Data



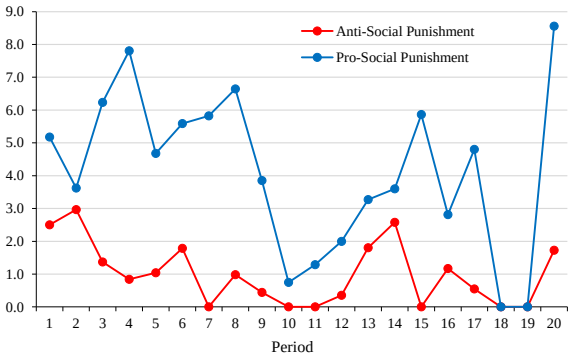

(b) Average Percentage by Treatment

**Figure B.4:** *Average Punishment Loss by the Type of Punishment, Period by Period*

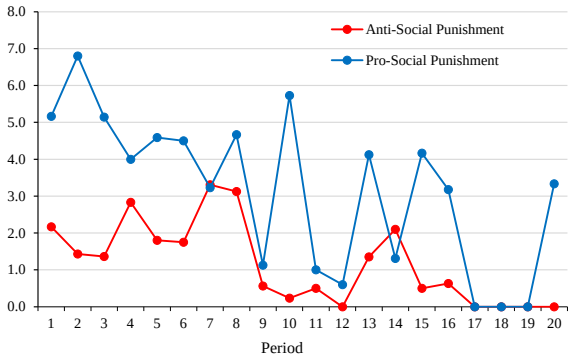
[A. Definition of Anti-social versus Pro-social punishment]



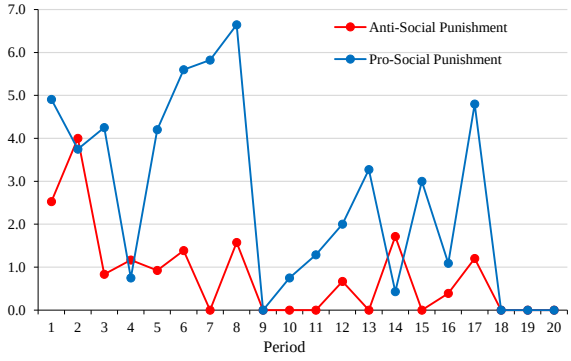

a. Individuals (I-Voting-M, I-Voting-ST)[#1]



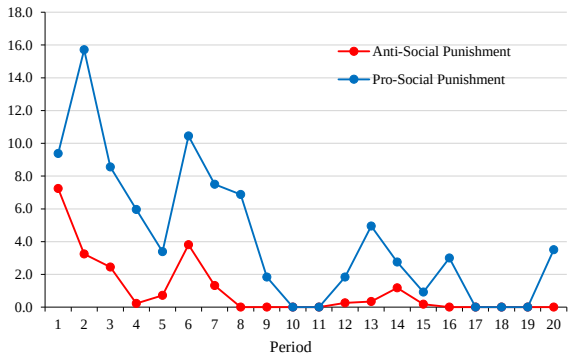

b. Teams (T-Voting-M, T-Voting-ST)[#1]



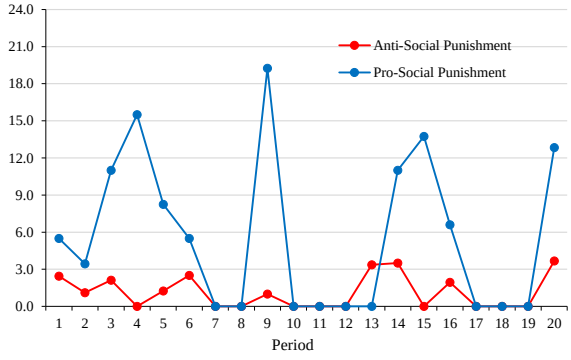

c. I-Voting-M




d. T-Voting-M



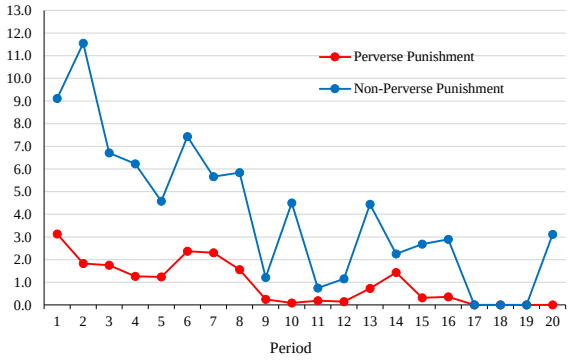

e. I-Voting-ST



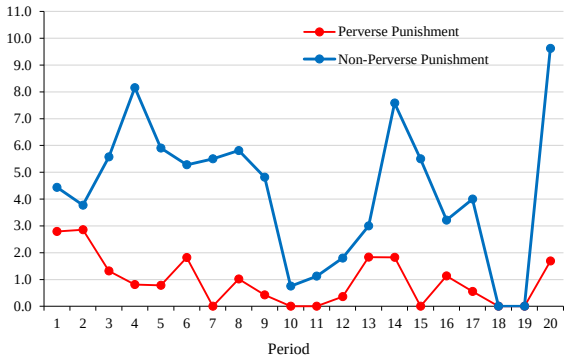

f. T-Voting-ST

*Notes*: Average anti-social (pro-social) punishment loss is calculated as the sum of anti-social (pro-social) punishment received by each unit playing the IS scheme in a given period, divided by the sum of opportunities to anti-socially (pro-socially) punish units in that period. [#1]Two treatments are pooled to calculate the average.

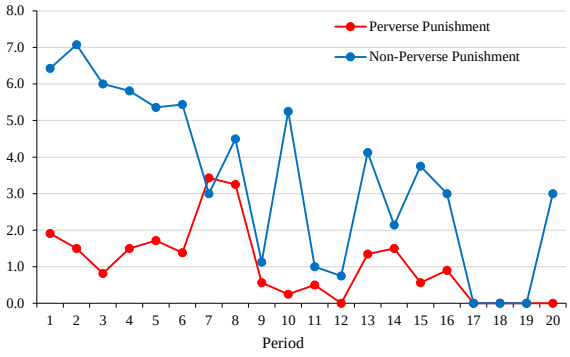
[B. Definition of Perverse versus Non-perverse punishment]



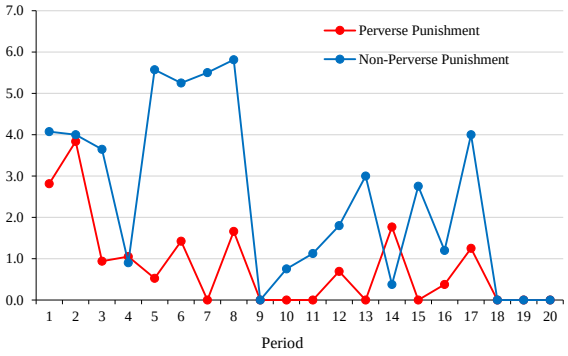

a. Individuals (I-Voting-M, I-Voting-ST)[#1]



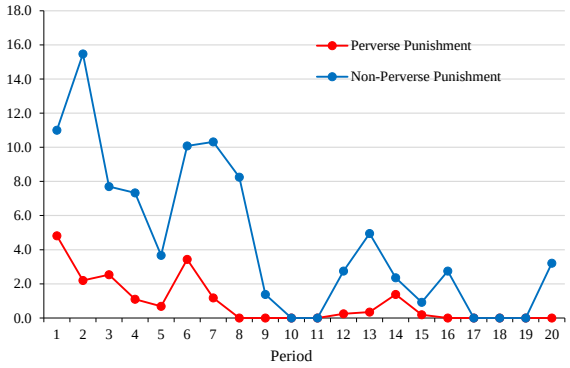

b. Teams (T-Voting-M, T-Voting-ST)[#1]



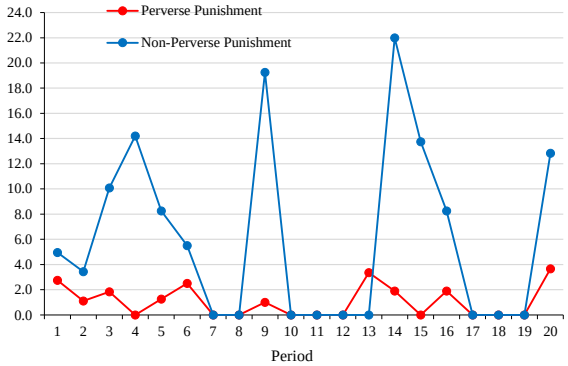

c. I-Voting-M




d. T-Voting-M




e. I-Voting-ST




f. T-Voting-ST

*Notes*: Average perverse (non-perverse) punishment loss is calculated as the sum of perverse (non-perverse) punishment received by each unit playing the IS scheme in a given period, divided by the sum of opportunities to perversely (non-perversely) punish units in that period. [#1] Two treatments are pooled to calculate the average.

**Table B.1:** *Cooperation Trends in Part 2 of the Experiment*

Dependent variable: Contribution amount of decision-making unit *i* in period *t*, where $5 \leq t \leq 24$.

| Independent Variable: | I-No | I-Voting-M | | | I-Voting-ST | | | T-No | T-Voting-M | | | T-Voting-ST | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All data (1) | All data (2a) | FS (2b) | IS (2c) | All data (3a) | FS (3b) | IS (3c) | All data (4) | All data (5a) | FS (5b) | IS (5c) | All data (6a) | FS (6b) | IS (6c) |
| Period within Phases {=1, 2, 3, 4} | -1.49*** (0.52) | -0.49* (0.27) | -1.13*** (0.40) | 0.16 (0.29) | 0.18 (0.29) | 0.25 (0.39) | -0.03 (0.30) | -1.12*** (0.28) | -0.45 (0.51) | 0.48 (1.09) | -0.86** (0.43) | -0.26 (1.02) | 1.49 (1.46) | -2.00* (1.19) |
| The phase number in Part 2[#1] {= 1, 2, 3, 4, 5} | -2.40*** (0.42) | -0.45** (0.21) | -1.48*** (0.32) | 0.65** (0.28) | 1.24*** (0.23) | 0.33 (0.39) | 1.21*** (0.27) | -1.37*** (0.22) | 2.26*** (0.42) | 2.11** (1.00) | 0.18 (0.44) | 1.88** (0.82) | 0.40 (1.12) | 2.80** (1.17) |
| Constant | 20.94*** (4.05) | 16.31*** (1.99) | 19.67*** (2.65) | 11.07*** (2.06) | 15.43*** (2.53) | 11.43*** (2.68) | 15.07*** (2.43) | 17.07*** (1.49) | 23.67*** (3.36) | 24.64*** (5.19) | 25.78*** (3.27) | 37.18*** (4.20) | 38.06*** (5.41) | 29.66*** (5.80) |
| # of observations | 720 | 660 | 348 | 312 | 660 | 300 | 360 | 720 | 660 | 276 | 384 | 660 | 396 | 264 |
| # of left(right)-censored observations | 237(235) | 75(205) | 57(76) | 18(129) | 46(311) | 38(95) | 8(216) | 133(119) | 27(441) | 15(195) | 12(246) | 16(568) | 11(341) | 5(227) |
| Log likelihood | -1284.60 | -1507.95 | -847.83 | -601.45 | -1202.59 | -651.16 | -503.92 | -1904.68 | -930.71 | -375.44 | -517.86 | -509.66 | -308.72 | -187.74 |
| Wald $\chi^2$ | 40.48 | 7.95 | 29.64 | 5.85 | 28.44 | 1.12 | 19.79 | 53.20 | 30.02 | 4.55 | 4.19 | 5.33 | 1.21 | 7.42 |
| Prob > Wald $\chi^2$ | 0.0000*** | 0.0188** | 0.0000*** | 0.0536* | 0.0000*** | 0.5702 | 0.0001*** | 0.0000*** | 0.0000*** | 0.1030 | 0.1233 | 0.0696* | 0.5472 | 0.0245** |

*Notes:* Tobit regressions. The reason for using a tobit model is that a large fraction of contribution decisions were censored. Decision-making unit random effects were included to control for the panel structure. The numbers in parentheses are standard errors. Observations in periods 5 to 24 are used in the regressions. [#1] This variable equals the phase number minus 1. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: *(a) When the sanctioning schemes were unavailable, decision-making units (whether individuals or teams) decreased contribution amounts from phase to phase. They also decreased contribution amounts over the periods within phases. (b) Individuals in the I-Voting-M treatment followed a similar declining trend as those in the I-No treatment, driven by free riding dynamics in the FS scheme. However, (c) individuals in the I-Voting-ST treatment increased contribution amounts from phase to phase because they learned to cooperate gradually under the IS scheme. (d) In both the T-Voting-M and T-Voting-ST treatments, teams' contribution amounts rose from phase to phase.*

Remark: As a robustness check, a random effects tobit regression was further conducted using group averages as observations, considering that decision-making units' decisions to contribute may have been correlated within group, confirming that the dynamics included in above Results (a) to (d) are all significant. The results are omitted to conserve space.

**Table B.2:** *Payoff Trends in Part 2 of the Experiment*

Dependent variable: Payoff of decision-making unit *i* in period *t*, where $5 \leq t \leq 24$.

| Independent Variable: | I-No | I-Voting-M | | | I-Voting-ST | | | T-No | T-Voting-M | | | T-Voting-ST | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All data (1) | All data (2a) | FS (2b) | IS (2c) | All data (3a) | FS (3b) | IS (3c) | All data (4) | All data (5a) | FS (5b) | IS (5c) | All data (6a) | FS (6b) | IS (6c) |
| Period within Phases {=1, 2, 3, 4} | -0.41** (0.17) | -0.31 (0.31) | -0.58** (0.23) | -0.02 (0.67) | 0.45 (0.32) | 0.04 (0.25) | 0.80 (0.53) | -0.62*** (0.13) | 0.09 (0.35) | -0.04 (0.40) | 0.18 (0.48) | -0.21 (0.20) | 0.12 (0.25) | -0.72 (0.53) |
| The phase number in Part 2[#1] {= 1, 2, 3, 4, 5} | -0.81** (0.30) | 0.61 (0.57) | 0.60 (0.36) | 0.88 (0.80) | 1.81*** (0.47) | 0.14 (0.31) | 1.89* (0.89) | -0.74*** (0.19) | 1.42** (0.50) | 1.04 (0.75) | 0.56* (0.26) | 0.59 (0.41) | 0.10 (0.29) | 1.13** (0.33) |
| Constant | 31.17*** (1.19) | 22.73*** (1.94) | 26.26*** (1.34) | 22.44*** (3.66) | 20.19*** (2.09) | 23.47*** (1.32) | 21.21*** (3.97) | 31.81*** (0.64) | 24.59*** (1.54) | 25.28*** (1.45) | 27.47*** (1.38) | 29.49*** (1.23) | 30.05*** (1.08) | 28.98*** (1.80) |
| # of observations | 720 | 660 | 348 | 312 | 660 | 300 | 360 | 720 | 660 | 276 | 384 | 660 | 396 | 264 |
| F | 4.31 | 1.21 | 4.30 | 1.56 | 10.48 | 0.11 | 2.56 | 20.49 | 4.55 | 5.56 | 2.51 | 1.41 | 0.16 | 7.42 |
| Prob > F | 0.0415** | 0.3375 | 0.0606* | 0.2623 | 0.0035*** | 0.8967 | 0.1320 | 0.0002*** | 0.0394** | 0.0307** | 0.1308 | 0.2886 | 0.8571 | 0.0239** |
| R-squared (Overall) | 0.0271 | 0.0097 | 0.0040 | 0.0433 | 0.0716 | 0.0061 | 0.1659 | 0.0396 | 0.0558 | 0.0114 | 0.1020 | 0.0138 | 0.0025 | 0.0371 |

*Notes:* Linear regressions. A tobit model was not used unlike in Table B.1 since the payoff data are not censored. Decision-making unit fixed effects were included to control for the panel structure. The numbers in parentheses are robust standard errors clustered by group ID. Observations in periods 5 to 24 are used in the regressions. [#1] This variable equals the phase number minus 1.

*, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Result: *(a) In the I-No and T-No treatments (where the sanctioning schemes were unavailable), decision-making units' payoffs decreased significantly from phase to phase. The payoffs also decreased over the periods within phases. (b) By sharp contrast, teams enjoyed high payoffs when the sanctioning schemes were available. First, the average payoff of a team increased from phase to phase in the T-Voting-M treatment. Second, the payoff of a team was high from the beginning of Part 2 and the high level stayed stable throughout Part 2 in the T-Voting-ST treatment. (c) While the payoff of an individual stayed low in the I-Voting-M treatment, it increased from phase to phase in the I-Voting-ST treatment.*

**Table B.3:** *Determinants of Votes between Formal and Informal Schemes in Phase 6*

Dependent variable: A dummy that equals 1(0) if a subject *i* votes for the FS (IS) regime.

| Data: | I-Voting-M (1) | I-Voting-ST (2) | T-Voting-M (3) | T-Voting-ST (4) | Two treatments with mild punishment (5) | Two treatments with strong punishment (6) | All four voting treatments (7) |
|---|---|---|---|---|---|---|---|
| Relative payoff ratio[#1] | 0.03<br>(0.02) | 0.09*<br>(0.04) | -0.02<br>(0.03) | 0.01*<br>(0.01) | 0.02<br>(0.02) | 0.02**<br>(0.01) | 0.01**<br>(0.01) |
| Gave anti-social pun dummy[#2] | 0.20<br>(0.15) | -0.12<br>(0.25) | 0.13<br>(0.19) | 0.39<br>(0.22) | 0.27**<br>(0.12) | 0.16<br>(0.17) | 0.26**<br>(0.09) |
| Received anti-social pun dummy[#3] | 0.02<br>(0.02) | 0.43*<br>(0.20) | 0.26<br>(0.26) | 0.50*<br>(0.22) | 0.28<br>(0.16) | 0.42**<br>(0.15) | 0.37***<br>(0.11) |
| Constant | 0.72<br>(0.19)*** | 0.06<br>(0.12) | 0.37<br>(0.25) | -0.21<br>(0.20) | 0.37*<br>(0.20) | 0.00<br>(0.11) | 0.17<br>(0.11) |
| # of observations | 21 | 24 | 27 | 18 | 48 | 42 | 90 |
| # of groups | 7 | 8 | 9 | 6 | 16 | 14 | 30 |
| F | 1.05 | 15.27 | 1.46 | 20.22 | 2.26 | 15.84 | 12.27 |
| Prob > F | 0.4351 | 0.0019*** | 0.2957 | 0.0032*** | 0.1237 | 0.0001*** | 0.0000*** |

*Notes:* Linear probability model with robust standard errors clustered by group ID. The numbers in parentheses are standard errors. Only groups who played under both the FS and IS regimes were used as data. [#1] Relative payoff ratio = (Avg. payoff under formal scheme before Phase 6)/(avg. payoff under informal scheme before Phase 6). [#2] Gave anti-social pun dummy equals 1 if a subject *i* has ever punished a group member who contributed more than *i* in the group when the IS was in place before Phase 6; 0 otherwise. [#3] Received anti-social pun dummy equals 1 if a subject *i* has ever been anti-socially punished when the IS was in place before Phase 6; 0 otherwise. Results are qualitatively similar even if the "gave perverse pun" dummy (which equals 1 if a subject *i* has ever punished a group member who contributed more than the group average contribution; 0 otherwise) or "received perverse pun" dummy, instead of variables #2 and #3, are used. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table B.4:** *Relationship between Realized Sanction Rates and Contributions under the FS Scheme*

Dependent variable: Contribution amount of decision-making unit *i* in period *t* in the FS scheme

| | I-Voting-M | | I-Voting-ST | | T-Voting-M | | T-Voting-ST | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Sanction rate {= 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2} | 16.24*** (3.22) | --- | 15.17*** (2.10) | --- | 30.97*** (4.06) | --- | 32.64*** (5.36) | --- |
| Deterrent dummy {= 1(0) if sanction rate was[#1] deterrent (non-deterrent)} | --- | 8.47*** (1.70) | --- | 8.28*** (1.84) | --- | 22.34*** (3.99) | --- | 24.66*** (5.05) |
| Period within Phases {=1, 2, 3, 4} | -1.05*** (0.39) | -1.06*** (0.39) | -0.12 (0.38) | 0.14 (0.38) | 0.40 (0.86) | -0.48 (1.00) | 2.31* (1.28) | 1.95 (1.38) |
| The phase number in Part 2[#2] {= 1, 2, 3, 4, 5} | -1.12*** (0.32) | -1.16*** (0.32) | 0.09 (0.37) | 0.21 (0.38) | -0.15 (0.86) | 0.82 (0.93) | -1.07 (1.04) | 0.41 (1.10) |
| Constant | 15.96*** (2.40) | 16.12*** (2.31) | 7.55*** (1.86) | 9.16*** (2.23) | 15.21*** (4.62) | 18.92*** (4.85) | 12.37** (4.82) | 17.43*** (5.28) |
| # of observations | 348 | 348 | 300 | 300 | 276 | 276 | 396 | 396 |
| # of left-censored observations | 57 | 57 | 38 | 38 | 15 | 15 | 11 | 11 |
| # of right-censored observations | 76 | 76 | 95 | 95 | 195 | 195 | 341 | 341 |
| Log likelihood | -861.83 | -862.28 | -626.90 | -640.66 | -329.51 | -354.78 | -259.40 | -283.57 |
| Wald $\chi^2$ | 54.26 | 53.51 | 52.95 | 21.28 | 60.44 | 34.55 | 37.13 | 24.13 |
| Prob > Wald $\chi^2$ | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. The numbers in parentheses are standard errors. Observations in periods 5 to 24 are used in the regressions. [#1] A sanction rate is deterrent (non-deterrent) if SR ≥ (<) 0.4. [#2] This variable equals the phase number minus 1. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

## Appendix A.C: Coding Procedure for the Communication Contents

### *C.1. Procedure*

Two coders were hired to judge each group's communication content, period by period, and then to assign relevant codes (summarized in Section C.2) to it in each period. Each coder read the content and decide on the code(s) starting from the early periods and working in ascending order. They assigned as many codes as they deemed appropriate, based on communication in a given period, or a stage within a period where there were multiple incidences of communication.

A total of 34 Excel files, simply numbered 1 to 34, were provided to each coder; these were termed "coding sheets." Each excel file had designated sections for each code type in every communication stage (e.g., for the T-No treatment, codes starting with A were in the first block, followed by codes starting with G). Along with each coding sheet, the coder was given the relevant communication file and a further excel file containing contextual information. The contextual information was intended to

provide a limited but necessary view of the group's behaviour to help make sense of their communication. These documents consisted of data from 12 groups, 11 groups, and 11 groups in the T-No treatment, T-Voting, and T-Voting-ST treatments, respectively. Coding took place in the order of files provided and working from top to bottom in each excel file.

Coding was conducted treatment by treatment, broken into four blocks as below, across approximately three months. Each coder worked on coding *without* knowing the other coder's identity for the entire process (hence, they were *not* able to communicate with each other).

First block (first 3 weeks):

The list of codes and the experiment instructions for the T-No treatment were provided on the first day. On that day, a meeting was scheduled by the researchers with each of the coders to explain the coding process and the first treatment. The coders were not made aware of the purpose of the research, other treatments, or any of the data analysis/results throughout the coding process.

To reduce the likelihood of problems and to give the coders feedback, the data from one group in the T-No treatment were provided as a sample to code before moving onto the remaining sessions. After coding the first session, a researcher met with each of the coders independently to discuss any problems or difficulties they had. This initial practice process and feedback took seven days. After that, the coders had a total of 14 days to code the remaining 11 T-No treatment groups. To monitor the progress of coding, group files were provided in two subblocks (five group data, and six group data). The first five files had to be completed before the researcher sent the remaining 6 files for coding. Once the coder had returned the group files to the researcher, they could not change the coding (unless there was some misunderstanding or confusion in the coding practice). Feedback was not given to the coders for their coding practice for the 11 files.

Second block (next 3 weeks):

At the onset of the second block, the coders were given the list of codes for the voting treatments and the experiment instructions for the T-Voting/T-Voting-ST treatment, and had a meeting with the researchers regarding the coding and the treatment (any questions could be clarified). In the second coding block, one coder worked on coding the T-Voting treatment, while the other coder worked on coding the T-Voting-ST treatment, to control for possible order effects of coding. Immediately after that, coders were given a sample from one of the sessions. The rest of the procedure is the same for the coding practice of the T-No treatment. Each coder was given one week to try coding the sample session, after which the coder had a meeting with one of the researchers. Once all questions were cleared in the

meeting, the first block of five group files was sent to be completed and returned before the remaining files were sent. Feedback was not given to the coders for their coding practice.

Third block (Third 3 weeks):

The coders were given the experiment instructions for the remaining treatments (e.g., those for the T-Voting-ST treatment if a coder coded sessions for the T-Voting treatment in the second block period) and the data. A quick meeting was scheduled to explain the difference between the two treatments. Each coder was then also given a chance to have a meeting with the researcher if they had any issues or questions. As before, the coders had to complete and return the first block of five group files before the remaining six files were received.

There was no sample coding in the third block as the list of codes is the same as in the second block. Again, feedback was not given to the coders for their coding practice.

Fourth block (final 3 weeks):

The coding results were compared for discrepancies by the researchers. The discrepancies were then highlighted in Excel spreadsheets and a copy given to each coder. The coders were asked to re-evaluate these discrepancies, with the additional knowledge of the other coder's codes, and either confirm or alter their initial findings. Each coder was also informed that his codes would be sent to the other coder and the other coder simultaneously re-evaluates these discrepancies. The coders neither communicated nor become aware of each other's identity, at any stage. This re-consideration process was used in van Elten and Penczynski (2020)[#1], confirming its effectiveness.

*Note*: [#1] van Elten, J., & Penczynski, S. (2020) Coordination games with asymmetric payoffs: An experimental study with intra-group communication. *J. Econ. Behav. Organ.*, *169*, 158-188.

*C.2. Full List of Codes*

(a) T-No treatment

| **Stage & codes** | **Description** | **Interpretation (this column was <u>not</u> shown to coders)** |
|---|---|---|
| **Note: A coder can assign whatever codes fit with teams' communication in a given stage (multiple choices); A coder may choose not to assign a code if communication does not fit any available code.** | | |
| **Part 1 & Part 2** | | |
| **Allocation:** | | |
| A1 | Recognizes that everyone receives a high payoff if they all contribute large amounts | Knowledge of the game |
| A2 | Suggests a high contribution or increasing contribution regardless of others behavior/mentions no strategy or reason | Unconditional cooperation |
| A3 | Suggests a high contribution or increasing contribution as a strategy to encourage other teams to cooperate or to avoid discouraging already cooperative teams | Rational cooperation; conditional cooperation (trust) |

| | | |
|---|---|---|
| A4 | Suggests a low contribution or reducing contribution regardless of others behavior/mentions no strategy or reason | Strict Nash; Unconditional non-cooperation |
| A5 | Suggests a low contribution or reducing contribution as a strategy out of distrust for the other teams/safety | Conditional cooperation (trust) |
| A6 | Suggests a low contribution or reducing contribution because of uncertainty over other teams' behaviors | Risk Aversion |
| A7 | Suggests a low contribution or reducing contribution regardless of others behavior using the logic of payoff maximisation | Strict Nash |
| A8 | Suggests contributing more/less based on contributions of others in previous rounds | Conditional cooperation (trust) |
| A9 | Discusses strategy from point of view of other teams | Conditional cooperation (trust); Rational cooperation |
| A10 | Discusses predictions for future rounds e.g. expectations of other teams' future contributions | Conditional cooperation (trust); Rational cooperation |
| A11 | Discusses when they change to defection (until when they attempt to foster cooperation norms) | Rational cooperation |
| A12 | Confusion, errors, mistakes (e.g., failing to understand anonymity condition, a matching protocol, or dominant strategy) | Confusion |
| A13 | Suggests a specific amount (e.g., around half) without writing any reasons | |
| A14 | Suggests around half to see how others respond before committing to high/low strategy | Learning |
| A15 | Suggests strategies that are successful in their daily social interactions | Learning |
| A16 | No communication or communication of unrelated things (boredom or established order) | N/A |
| A17 | Suggests just the same strategy they adopted in the last period (e.g., strategy in period $t$ now is the same as strategy in period $t-1$). This code is only applicable for a period greater than 1. | |
| | | |
| **General:** | | |
| G1 | Team members agree on a socially desirable option through discussion, despite two members initially having different opinions (i.e., higher contribution, higher sanction rates, punishment of free riders). | Group polarization (SCT) |
| G2 | Either a) a team agrees on an extreme option presented by one team member after initially preferring another option, or b) a team reaches an extreme decision through discussion despite all three team members preferring other options initially. | Group polarization (PAT) |
| G3 | Team makes a compromise choice between the two members. | Group polarization (mean reverting (e.g. Sunstein, 2007)) |
| G4 | Asks for team members to make a suggestion/offer their opinion | Team behavior |
| G5 | Collaborates decisions with team mates (e.g., checking whether calculations are correct, whether their memory of previous period information is correct) | Team behavior |
| G6 | Discuss changing behavior (contribution/scheme, etc) out of boredom | Team behavior |
| G7 | One or more team member disagrees with the others until the end of a given chat stage | Team behavior |
| G8 | Uses loaded words that were not used in the instructions (e.g., 'contribute', 'punish', 'retaliate', 'donate', 'tax' etc.) | Ideological reasoning |
| G9 | Shows knowledge of game e.g. discusses game theory, public goods, rationality, etc. | Knowledge of the game |
| G10 | Suggests that other teams do not understand game/mechanism | Knowledge of the game |
| G11 | Suggests that other teams are 'good' or 'bad'/'trustworthy' or 'untrustworthy' | Rational cooperation, Conditional cooperation (trust) |
| G12 | Expresses that the group situation is positive/ordered or chaotic/negative | Emotions |
| G13 | Expresses a positive emotion e.g. they are enjoying the experiment or are happy | Emotions |
| G14 | Expresses a negative emotion e.g. they are angry or annoyed | Emotions |
| G15 | Relates game to political ideology e.g. capitalism, communism, etc | Ideological reasoning |

| | | |
|---|---|---|
| G16 | Expresses belief that other teams/players are not real or are simulations | Confusion |
| G17 | Discusses or refers to the length of the chat session being sufficient or long | |

[The frequency that given codes were marked by coders:]

| Codes | % of teams that were assigned codes at least once[#1] | Avg. % of periods per team the code was marked[#2] |
|---|---|---|
| A1 | 77.78% | 0.08 |
| A2 | 91.67% | 0.14 |
| A3 | 97.22% | 0.17 |
| A4 | 88.89% | 0.14 |
| A5 | 75.00% | 0.11 |
| A6 | 55.56% | 0.04 |
| A7 | 77.78% | 0.10 |
| A8 | 94.44% | 0.21 |
| A9 | 100.00% | 0.69 |
| A10 | 100.00% | 0.33 |
| A11 | 75.00% | 0.09 |
| A12 | 58.33% | 0.07 |
| A13 | 80.56% | 0.08 |
| A14 | 61.11% | 0.04 |
| A15 | 0.00% | 0.00 |
| A16 | 38.89% | 0.05 |
| A17 | 100.00% | 0.29 |

| Codes | % of teams that were assigned codes at least once | Avg. % of periods per team the code was marked |
|---|---|---|
| G1 | 80.56% | 0.10 |
| G2 | 94.44% | 0.11 |
| G3 | 80.56% | 0.10 |
| G4 | 100.00% | 0.88 |
| G5 | 97.22% | 0.27 |
| G6 | 8.33% | 0.00 |
| G7 | 47.22% | 0.04 |
| G8 | 16.67% | 0.01 |
| G9 | 13.89% | 0.01 |
| G10 | 30.56% | 0.02 |
| G11 | 88.89% | 0.21 |
| G12 | 80.56% | 0.08 |
| G13 | 63.89% | 0.08 |
| G14 | 77.78% | 0.12 |
| G15 | 11.11% | 0.00 |
| G16 | 13.89% | 0.02 |
| G17 | 5.56% | 0.00 |

*Note*: [#1] Calculated as the number of teams that were assigned a given code at least once divided by the number of teams with the opportunity to be assigned a given code. The denominator here means that if they did not play the FS scheme, then they would not be included in the number of teams with opportunity (for example, only 27 of 33 teams played FS in the T-Voting treatment, so the % for C1 was calculated as 8/27). [#2] Calculated as the total amount of time a code was marked divided by {the number of periods in which it could be marked* times the number of teams with the opportunity to be assigned a given code in a given period}. Here, the term * means that, for example, A codes could only be assigned in periods 1-4 in the voting treatments (four periods), or B codes in scheme choice rounds (5 periods). So, A1 in the voting treatments are calculated as (total marked A1 code/4)/33.

## (b) T-Voting and T-Voting-ST treatments

| Stage & codes | Description | |
|---|---|---|
| **Note: A coder can assign whatever codes fit with teams' communication in a given stage (multiple choices); A coder may choose not to assign a code if communication does not fit any available code.** | | |
| | | |
| **Part 1** | | |
| **Allocation – Part 1 only** | | |
| A1 | Recognizes that everyone receives a high payoff if they all contribute large amounts | Knowledge of the game |
| A2 | Suggests high contribution or increasing contribution regardless of others behavior/mentions no strategy or reason | Unconditional cooperation |
| A3 | Suggests high contribution or increasing contribution as a strategy to encourage other teams to cooperate or to avoid discouraging already cooperative teams | Rational cooperation; conditional cooperation (trust) |
| A4 | Suggests a low contribution or reducing contribution regardless of others behavior/mentions no strategy or reason | Strict Nash; Unconditional non-cooperation |
| A5 | Suggests a low contribution or reducing contribution as a strategy out of distrust for the other teams/safety | Conditional cooperation (trust) |
| A6 | Suggests a low contribution or reducing contribution because of uncertainty over | Risk Aversion |

| | | |
|---|---|---|
| | other teams' behaviors | |
| A7 | Suggests a low contribution or reducing contribution regardless of others behavior using the logic of payoff maximisation | Strict Nash |
| A8 | Suggests contributing more/less based on contributions of others in previous rounds | Conditional cooperation (trust) |
| A9 | Discusses strategy from point of view of other teams | Conditional cooperation (trust); Rational cooperation |
| A10 | Discusses predictions for future rounds e.g. expectations of other teams' future contributions | Conditional cooperation (trust); Rational cooperation |
| A11 | Discusses when they change to defection (until when they attempt to foster cooperation norms) | Rational cooperation |
| A12 | Confusion, errors, mistakes (e.g., failing to understand anonymity condition, a matching protocol, or dominant strategy) | Confusion |
| A13 | Suggests a specific amount (e.g., around half) without writing any reasons | |
| A14 | Suggests around half to see how others respond before committing to high/low strategy | Learning |
| A15 | Suggests strategies that are successful in their daily social interactions | Learning |
| A16 | No communication or communication of unrelated things (boredom or established order) | N/A |
| A17 | Suggests just the same strategy they adopted in the last period (e.g., strategy in period $t$ now is the same as strategy in period $t - 1$). This code is only applicable for a period greater than 1. | |
| | | |
| **Part 2** | | |
| **Scheme Choice decision (first period of each phase only)** | | |
| B1 | Preference for FS based on the assumption of self-interest (i.e., full cooperation with a deterrent sanction rate versus low cooperation with the IS scheme as no punishment is inflicted in the latter scheme) | Strict Nash |
| B2 | Preference for FS to protect from extreme/variable/unpredictable punishment under IS | Risk aversion |
| B3 | Preference for FS to construct the NS by selecting a sanction rate of 0.0 (same as Phase 1) | |
| B4 | Preference for IS to avoid repeated administrative cost | |
| B5 | Preference for IS to avoid being fined (e.g. where planning to contribute low) | |
| B6 | Preference for IS for more control over punishment | Risk aversion |
| B7 | Preference for IS to avoid risk of high shared cost of fines towards other teams (e.g. if others contribute low) | Risk aversion |
| B8 | Preference based on ideology e.g. liberal/anti-tax/anti-punishment | Ideological reasoning |
| B9 | Selects one of the schemes randomly due to spiteful motives (e.g., to confuse other teams, gaming) | Spitefulness |
| B10 | Selects one of the schemes randomly from motives other than B9 (e.g., to see how it goes) | Learning |
| B11 | Preference for either scheme based on experience/contributions/behaviors in previous phases | Learning |
| B12 | Preference based on perceived simplicity of scheme e.g. easier to understand | Cognitive Load |
| B13 | Confusion, errors, mistakes (e.g., failing to understand the conditions of each scheme, such as the presence of the administrative cost and punishment technology in the IS scheme) | Learning |
| B14 | Discusses voting based on how other teams vote | Other (strategic voting) |
| B15 | No communication or communication of unrelated things (boredom or established order) | N/A |
| B16 | Suggests just the same strategy they adopted in the last phase (e.g., strategy in phase $t$ now is the same as strategy in phase $t - 1$). This code is only applicable only when second to sixth voting phases. | |
| | | |
| **Sanction Rate decision (Formal Scheme [Group-determined fines] only)** | | |
| 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 | Ideological reasoning |

| | | |
|---|---|---|
| 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) | Confusion |
| C3 | Discusses rate based on own contribution plans e.g. low sanction rate for low contributors | |
| C4 | Discusses rate based on other teams' contributions e.g. high sanction rate for low contributors/to encourage cooperation | Conditional cooperation (Trust for other units) |

| | | |
|---|---|---|
| C5[63] | 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 | Strict Nash |
| C6 | Discusses effects of a strong sanction rate, other than deterrence (e.g., why 1.2 is preferred to 0.8) | |
| C7 | Discusses increasing/decreasing rate as a reaction to experiences with previous rates | Learning |
| C8 | Selects one of the sanction rates randomly to see how it goes | Learning |
| C9 | Discusses voting based on how other teams vote | Other (strategic voting) |
| C10 | No communication or communication of unrelated things (boredom or established order) | N/A |
| C11 | Suggests just the same strategy they adopted in the last period (e.g., strategy in period *t* now is the same as strategy in period $t - 1$). This code is only applicable only when the period number within phase is 2 to 4. | |
| | | |
| **Allocation decision (Formal Scheme only)** | | |
| D1 | Suggests high contributions regardless of sanction rate or other cooperation behavior, without any reasoning | Unconditional cooperation |
| D2 | Suggests low contributions regardless of sanction rate or other cooperation behavior, without any reasoning | Unconditional non-cooperation |
| D3 | Suggests high contribution as a strategy to encourage other teams to cooperate or to avoid discouraging already cooperative teams | Rational cooperation; conditional cooperation (trust); Strict Nash, dependent on a realized sanction rate |
| D4 | Suggests a low contribution as a strategy out of distrust for the other teams/safety [this coding option is available only when sanction rate is 0.0, or 0.2.] | Conditional cooperation (trust) |
| D5 | Suggests contributing more/less based on contributions of others in previous rounds [this coding option is available only when sanction rate is 0.0, or 0.2.][64] | Conditional cooperation (trust) |
| D6 | Suggests contributing more/less based on contributions of others in previous rounds, despite the current period sanction rate being more than or equal to 0.4 (deterrent), due to confusion. | Confusion |
| D7 | Discusses strategy from point of view of other teams | Conditional cooperation (trust); Rational cooperation |
| D8 | Suggests low contributions to increase others' per capita share of imposing fine (similar to anti-social punishment or revenge) | Spitefulness |
| D9 | Discusses contribution to avoid fines e.g. suggests high contribution to avoid fines | Ideological reasoning; Cognitive load |
| D10 | Discusses contribution based on material motives (i.e., contribute large amounts if SR is deterrent; contribute little if SR is non- deterrent) | Strict Nash |
| D11 | Confusion, errors, mistakes (e.g., failing to understand which allocation should be subject to penalty, how per capita share of imposing fine is calculated) | Confusion |
| D12 | No communication or communication of unrelated things (boredom or established order) | N/A |
| D13 | Suggests just the same strategy they adopted in the last period (e.g., strategy in period *t* now is the same as strategy in period $t - 1$). This code is only applicable only when the period number within phase is 2 to 4. | |
| | | |
| **Allocation decision (Informal Scheme [Team Reduction Decisions] only)** | | |
| E1 | Suggests high contribution regardless of punishment received or any other behavior/mentions no strategy or reason | Unconditional cooperation |

[63] It should be noted that the deterrent level in C5 was incorrectly given on the code sheet in that 0.4 is a deterrent sanction rate. To confirm that this is unlikely to have affected coders' coding, all communication files were reviewed for use of the C5 code. It was found that coders used the code more generally to refer to discussion of deterring others by selecting a higher sanction rate (at any level) rather than trying to work out the mathematical threshold of deterrent rates. In particular, no teams were found to discuss whether the specific rate of 0.4 was a deterrent to others or not. Furthermore, as the code required coding of any "discussion" of the deterrent level, the code would have been used regardless as to what the correct deterrent level was. As such, a visual inspection of the data and the nature of the code highly suggest that this error had little to no effect on coding.

[64] The instruction to the two coders was incorrectly given regarding codes D4 and D5. They were instructed to consider whether these two codes were relevant when the sanction rate was 0.4, in addition to 0.0 and 0.2. As this is just an error of instruction, the authors were able to correct for potential over-use of these codes by simply removing assignments when the sanction rate was 0.4.

| | | |
|---|---|---|
| E2 | Suggests a low contribution regardless of punishment received or any other behavior/mentions no strategy or reason | Strict Nash; Unconditional non-cooperation |
| E3 | Suggests high contribution as a strategy to encourage other teams to cooperate or to avoid discouraging already cooperative teams | Rational cooperation; conditional cooperation (trust) |
| E4 | Suggests a low contribution as a strategy out of distrust for the other teams (e.g., punishment in previous rounds did not affect others' contributions) | Conditional cooperation (trust) |
| E5 | Discusses contribution relative to potential punishment beliefs e.g. suggests high contribution to avoid punishment or low contribution if they don't think others will punish | Conditional cooperation (trust) |
| E6 | Suggests maintaining contribution at same level when not punished in previous rounds | Conditional cooperation (trust) |
| E7 | Suggests decreasing contribution when not punished in previous rounds | Strict Nash |
| E8 | Suggests high contribution when pro-socially punished in previous period | Learning |
| E9 | Suggests low contribution when pro-socially punished in previous period | Spitefulness, or Confusion |
| E10 | Suggests low contribution when anti-socially punished in previous period | Learning |
| E11 | Discusses defection in the end period in a given phase | |
| E12 | Suggests around half or some amount intuitively to see what allocation strategies others use | Learning |
| E13 | Suggests around half or some amount intuitively to see what punishment strategies others use | Learning |
| E14 | Confusion, errors, mistakes (e.g., failing to understand the consequence of punishment) | Confusion |
| E15 | No communication or communication of unrelated things (boredom or established order) | N/A |
| E16 | Suggests just the same strategy they adopted in the last period (e.g., strategy in period $t$ now is the same as strategy in period $t-1$). This code is only applicable only when the period number within phase is 2 to 4. | |
| | | |
| **Punishment decision (Informal Scheme only)** | | |
| F1 | Suggests punishment for a contribution higher than their own (anti-social) | Spitefulness |
| F2 | Suggests no punishment for a contribution higher than their own (pro-social) | Rational Cooperation; Conditional cooperation (trust) |
| F3 | Suggests punishment for a contribution lower than their own (pro-social) | Rational Cooperation; Conditional cooperation (trust) |
| F4 | Suggests no punishment for a contribution lower than their own | Strict Nash |
| F5 | Suggests punishment for a contribution higher than the group average (perverse) | Spitefulness |
| F6 | Suggests no punishment for a contribution higher than the group average (non-perverse) | Strict Nash; Conditional cooperation (trust) |
| F7 | Suggests punishment for a contribution lower than the group average (non-perverse) | Rational Cooperation; Conditional cooperation (trust) |
| F8 | Suggests no punishment for a contribution lower than the group average | Strict Nash |
| F9 | Suggests punishment based on absolute contribution e.g. below or above a specific number | |
| F10 | Expresses desire to avoid punishment regardless of contribution due to ideological reason | Ideological Reasoning |
| F11 | Expresses desire to avoid punishment regardless of contribution due to the cost in imposing punishment | Strict Nash |
| F12 | Suggests punishment to those who may punish themselves (retaliation) | |
| F13 | Expresses desire to avoid punishment to prevent retaliation | |
| F14 | Expresses that punishment seems to be random/irrational/perverse (anti-social/perverse) | |
| F15 | Suggests punishment as a revenge for previous-round punishment received | Betrayal Aversion; Other (blind revenge [e.g., Ostrom *et al* 1992]) |
| F16 | Suggests punishment as an emotional response | Emotions |
| F17 | Suggests that punishment strength is too strong (received punishment or punishing others) | |
| F18 | Suggests punishment to a team as a response to that team's previous contribution decision | Other (delayed punishment) |
| F19 | Confusion, errors, mistakes (e.g., failing to understand the punishment cost) | Confusion |

| | | |
|---|---|---|
| F20 | No communication or communication of unrelated things (boredom or established order) | N/A |
| F21 | Discusses increasing (decreasing) punishment when group average contribution decreased (increased) in the current periodcompared with in the last period. Note. This code is only applicable only when the period number within a phase is 2 to 4. | Other (Conformity) |
| F22 | Suggests just the same strategy they adopted in the last period (e.g., strategy in period *t* now is the same as the strategy in period *t* – 1). This code is only applicable only when the period number within a phase is 2 to 4. | |
| | | |
| **General (both for Part 1 and Part 2):** | | |
| G1 | Team members agree on a socially desirable option through discussion, despite two members initially having different opinions (i.e., higher contribution, higher sanction rates, punishment of free riders). | Group polarization (SCT) |
| G2 | Either a) a team agrees on an extreme option presented by one team member after initially preferring another option, or b) a team reaches an extreme decision through discussion despite all three team members preferring other options initially. | Group polarization (PAT) |
| G3 | Team makes a compromise choice between the two members. | Group polarization (mean reverting (e.g. Sunstein, 2007)) |
| G4 | Asks for team members to make a suggestion/offer their opinion | Team behavior |
| G5 | Collaborates decisions with team mates (e.g., checking whether calculations are correct, whether their memory of previous period information is correct) | Team behavior |
| G6 | Discuss changing behavior (contribution/scheme, etc) out of boredom | Team behavior |
| G7 | One or more team member disagrees with the others until the end of a given chat stage | Team behavior |
| G8 | Uses loaded words that were not used in the instructions (e.g., 'contribute', 'punish', 'retaliate', 'donate', 'tax' etc.) | Ideological reasoning |
| G9 | Shows knowledge of game e.g. discusses game theory, public goods, rationality, etc. | Knowledge of the game |
| G10 | Suggests that other teams do not understand game/mechanism | Knowledge of the game |
| G11 | Suggests that other teams are 'good' or 'bad'/'trustworthy' or 'untrustworthy' | Rational cooperation, Conditional cooperation (trust) |
| G12 | Expresses that the group situation is positive/ordered or chaotic/negative | Emotions |
| G13 | Expresses a positive emotion e.g. they are enjoying the experiment or are happy | Emotions |
| G14 | Expresses a negative emotion e.g. they are angry or annoyed | Emotions |
| G15 | Relates game to political ideology e.g. capitalism, communism, etc | Ideological reasoning |
| G16 | Expresses belief that other teams/players are not real or are simulations | Confusion |
| G17 | Discusses or refers to the length of the chat session being sufficient or long | |

[The frequency that given codes were marked by coders:]

| | **T-Voting treatment** | | | **T-Voting-ST treatment** | |
|---|---|---|---|---|---|
| **Codes** | **% of teams that were assigned codes at least once** | **Avg. % of periods per team the code was marked** | | **% of teams that were assigned codes at least once** | **Avg. % of periods per team the code was marked** |
| A1 | 54.55% | 0.22 | | 54.55% | 0.22 |
| A2 | 36.36% | 0.11 | | 54.55% | 0.18 |
| A3 | 66.67% | 0.29 | | 51.52% | 0.20 |
| A4 | 33.33% | 0.12 | | 36.36% | 0.11 |
| A5 | 36.36% | 0.13 | | 36.36% | 0.11 |
| A6 | 9.09% | 0.03 | | 12.12% | 0.04 |
| A7 | 42.42% | 0.11 | | 36.36% | 0.19 |
| A8 | 45.45% | 0.16 | | 66.67% | 0.30 |
| A9 | 84.85% | 0.54 | | 84.85% | 0.65 |
| A10 | 78.79% | 0.37 | | 69.70% | 0.36 |
| A11 | 21.21% | 0.07 | | 18.18% | 0.07 |
| A12 | 12.12% | 0.04 | | 18.18% | 0.07 |
| A13 | 15.15% | 0.05 | | 21.21% | 0.06 |
| A14 | 18.18% | 0.05 | | 27.27% | 0.10 |
| A15 | 0.00% | 0.00 | | 0.00% | 0.00 |

| | | | | | |
|---|---|---|---|---|---|
| A16 | 6.06% | 0.02 | | 6.06% | 0.02 |
| A17 | 78.79% | 0.37 | | 69.70% | 0.33 |
| | | | | | |
| B1 | 9.09% | 0.02 | | 24.24% | 0.05 |
| B2 | 12.12% | 0.03 | | 42.42% | 0.13 |
| B3 | 18.18% | 0.04 | | 15.15% | 0.03 |
| B4 | 51.52% | 0.15 | | 42.42% | 0.10 |
| B5 | 3.03% | 0.01 | | 6.06% | 0.02 |
| B6 | 21.21% | 0.04 | | 24.24% | 0.05 |
| B7 | 0.00% | 0.00 | | 0.00% | 0.00 |
| B8 | 0.00% | 0.00 | | 3.03% | 0.01 |
| B9 | 0.00% | 0.00 | | 0.00% | 0.00 |
| B10 | 54.55% | 0.14 | | 57.58% | 0.21 |
| B11 | 69.70% | 0.22 | | 69.70% | 0.21 |
| B12 | 24.24% | 0.05 | | 18.18% | 0.04 |
| B13 | 6.06% | 0.02 | | 24.24% | 0.05 |
| B14 | 0.00% | 0.00 | | 15.15% | 0.04 |
| B15 | 27.27% | 0.10 | | 42.42% | 0.16 |
| B16 | 87.88% | 0.38 | | 93.94% | 0.35 |
| | | | | | |
| C1 | 29.63% | 0.04 | | 26.67% | 0.04 |
| C2 | 51.85% | 0.04 | | 36.67% | 0.04 |
| C3 | 29.63% | 0.03 | | 53.33% | 0.07 |
| C4 | 81.48% | 0.11 | | 83.33% | 0.09 |
| C5 | 51.85% | 0.05 | | 73.33% | 0.07 |
| C6 | 14.81% | 0.01 | | 36.67% | 0.03 |
| C7 | 59.26% | 0.06 | | 66.67% | 0.08 |
| C8 | 66.67% | 0.07 | | 56.67% | 0.06 |
| C9 | 14.81% | 0.01 | | 40.00% | 0.03 |
| C10 | 40.74% | 0.08 | | 53.33% | 0.20 |
| C11 | 81.48% | 0.17 | | 86.67% | 0.19 |
| | | | | | |
| D1 | 88.89% | 0.15 | | 80.00% | 0.10 |
| D2 | 37.04% | 0.04 | | 46.67% | 0.03 |
| D3 | 33.33% | 0.04 | | 46.67% | 0.03 |
| D4 | 29.63% | 0.02 | | 33.33% | 0.02 |
| D5 | 11.11% | 0.01 | | 20.00% | 0.01 |
| D6 | 7.41% | 0.01 | | 23.33% | 0.02 |
| D7 | 62.96% | 0.11 | | 76.67% | 0.11 |
| D8 | 0.00% | 0.00 | | 3.33% | 0.00 |
| D9 | 55.56% | 0.06 | | 80.00% | 0.10 |
| D10 | 81.48% | 0.10 | | 83.33% | 0.12 |
| D11 | 22.22% | 0.01 | | 40.00% | 0.04 |
| D12 | 33.33% | 0.06 | | 60.00% | 0.22 |
| D13 | 70.37% | 0.10 | | 86.67% | 0.15 |
| | | | | | |
| E1 | 66.67% | 0.07 | | 61.90% | 0.06 |
| E2 | 48.48% | 0.06 | | 23.81% | 0.02 |
| E3 | 48.48% | 0.04 | | 61.90% | 0.05 |
| E4 | 24.24% | 0.02 | | 23.81% | 0.01 |
| E5 | 21.21% | 0.03 | | 47.62% | 0.03 |
| E6 | 60.61% | 0.09 | | 33.33% | 0.02 |
| E7 | 36.36% | 0.04 | | 33.33% | 0.02 |
| E8 | 36.36% | 0.03 | | 38.10% | 0.03 |
| E9 | 18.18% | 0.02 | | 14.29% | 0.01 |
| E10 | 33.33% | 0.02 | | 28.57% | 0.01 |
| E11 | 30.30% | 0.03 | | 23.81% | 0.04 |
| E12 | 21.21% | 0.01 | | 19.05% | 0.01 |
| E13 | 18.18% | 0.01 | | 23.81% | 0.01 |
| E14 | 15.15% | 0.01 | | 28.57% | 0.02 |

| | | | | | |
|---|---|---|---|---|---|
| E15 | 39.39% | 0.17 | | 66.67% | 0.25 |
| E16 | 87.88% | 0.13 | | 71.43% | 0.17 |
| | | | | | |
| F1 | 42.42% | 0.04 | | 19.05% | 0.02 |
| F2 | 51.52% | 0.11 | | 52.38% | 0.05 |
| F3 | 63.64% | 0.08 | | 66.67% | 0.06 |
| F4 | 48.48% | 0.10 | | 28.57% | 0.03 |
| F5 | 39.39% | 0.04 | | 23.81% | 0.02 |
| F6 | 66.67% | 0.15 | | 47.62% | 0.05 |
| F7 | 63.64% | 0.08 | | 71.43% | 0.06 |
| F8 | 57.58% | 0.15 | | 33.33% | 0.03 |
| F9 | 51.52% | 0.04 | | 9.52% | 0.01 |
| F10 | 9.09% | 0.01 | | 23.81% | 0.01 |
| F11 | 39.39% | 0.03 | | 57.14% | 0.05 |
| F12 | 24.24% | 0.02 | | 28.57% | 0.02 |
| F13 | 54.55% | 0.06 | | 38.10% | 0.04 |
| F14 | 9.09% | 0.01 | | 9.52% | 0.00 |
| F15 | 21.21% | 0.01 | | 28.57% | 0.02 |
| F16 | 9.09% | 0.00 | | 19.05% | 0.01 |
| F17 | 12.12% | 0.01 | | 42.86% | 0.02 |
| F18 | 15.15% | 0.01 | | 14.29% | 0.01 |
| F19 | 12.12% | 0.01 | | 9.52% | 0.00 |
| F20 | 54.55% | 0.19 | | 61.90% | 0.24 |
| F21 | 3.03% | 0.00 | | 23.81% | 0.02 |
| F22 | 48.48% | 0.06 | | 57.14% | 0.14 |
| | | | | | |
| G1 | 96.97% | 0.13 | | 84.85% | 0.09 |
| G2 | 78.79% | 0.10 | | 63.64% | 0.06 |
| G3 | 36.36% | 0.04 | | 54.55% | 0.04 |
| G4 | 100.00% | 1.20 | | 100.00% | 0.93 |
| G5 | 100.00% | 1.25 | | 100.00% | 0.99 |
| G6 | 57.58% | 0.09 | | 45.45% | 0.04 |
| G7 | 33.33% | 0.03 | | 33.33% | 0.02 |
| G8 | 39.39% | 0.09 | | 12.12% | 0.01 |
| G9 | 18.18% | 0.01 | | 18.18% | 0.02 |
| G10 | 36.36% | 0.05 | | 54.55% | 0.05 |
| G11 | 84.85% | 0.20 | | 81.82% | 0.15 |
| G12 | 69.70% | 0.10 | | 69.70% | 0.12 |
| G13 | 72.73% | 0.11 | | 75.76% | 0.11 |
| G14 | 69.70% | 0.09 | | 66.67% | 0.08 |
| G15 | 12.12% | 0.01 | | 18.18% | 0.02 |
| G16 | 6.06% | 0.01 | | 9.09% | 0.01 |
| G17 | 57.58% | 0.06 | | 57.58% | 0.07 |

*C.3. Agreement Rate and Cohen's Kappa*

The average Cohen's Kappas for the initial coding were 0.28, 0.29 and 0.38 in the T-No, T-Voting and T-Voting-ST treatments, respectively. The reconsideration step improved the Kappas. After the independent reconsideration process, the Kappas became 0.88, 0.90 and 0.87 in the T-No, T-Voting and T-Voting-ST treatments, respectively.

Remark: The overall agreement rates of coding between the two coders after (before) the reconsideration process were 86.4% (97.0%), 90.3% (98.0%), and 91.6% (97.6%) in the T-No, T-Voting and T-Voting-ST treatments, respectively.

The following summarizes the agreement rates and the Kappas before and after the reconsideration step for each code:

## (a) T-No Treatment

[Agreement Rate:]

| Agreement rate | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 94.8% | 85.5% | 84.1% | 84.4% | 90.4% | 94.6% | 89.4% | 77.3% | 47.5% | 69.7% | 91.6% | 90.2% | 91.6% | 94.8% | 100.0% | 95.1% | 81.7% |
| After reconsideration | 98.3% | 94.7% | 93.4% | 93.3% | 97.3% | 98.8% | 97.0% | 94.9% | 97.1% | 92.7% | 98.1% | 96.9% | 95.8% | 97.3% | 100.0% | 99.3% | 91.0% |

| Agreement rate | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 86.8% | 81.3% | 87.7% | 14.7% | 69.4% | 99.4% | 94.7% | 98.8% | 99.2% | 96.4% | 74.9% | 89.4% | 92.7% | 89.8% | 99.9% | 99.3% | 99.2% |
| After reconsideration | 97.1% | 92.7% | 96.2% | 99.8% | 95.0% | 99.9% | 97.8% | 99.7% | 99.5% | 98.3% | 93.2% | 97.5% | 97.1% | 98.6% | 100.0% | 99.9% | 99.4% |

[Cohen's Kappa:]

| Kappa | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.61 | 0.34 | 0.42 | 0.32 | 0.42 | 0.02 | 0.25 | 0.12 | 0.14 | 0.23 | 0.33 | 0.13 | 0.44 | 0.37 | n.a. | 0.33 | 0.57 |
| After reconsideration | 0.89 | 0.80 | 0.80 | 0.77 | 0.87 | 0.88 | 0.85 | 0.86 | 0.93 | 0.84 | 0.90 | 0.80 | 0.78 | 0.75 | n.a. | 0.94 | 0.80 |

| Kappa | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.12 | 0.10 | 0.23 | 0.01 | 0.02 | 0.00 | 0.21 | -0.01 | 0.58 | 0.20 | 0.15 | 0.00 | 0.48 | 0.37 | 0.86 | 0.75 | 0.00 |
| After reconsideration | 0.86 | 0.71 | 0.82 | 0.99 | 0.88 | 0.89 | 0.77 | 0.82 | 0.80 | 0.72 | 0.81 | 0.85 | 0.84 | 0.94 | 1.00 | 0.96 | 0.44 |

## (b) T-Voting Treatment

[Agreement Rate:]

| Agreement rate | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 90.2% | 85.6% | 72.0% | 86.4% | 87.1% | 95.5% | 90.2% | 85.6% | 70.5% | 69.7% | 95.5% | 94.7% | 92.4% | 95.5% | 100.0% | 98.5% | 89.4% |
| After reconsideration | 97.7% | 92.4% | 95.5% | 96.2% | 98.5% | 97.7% | 94.7% | 97.0% | 96.2% | 97.7% | 100.0% | 98.5% | 94.7% | 98.5% | 100.0% | 100.0% | 93.2% |

| Agreement rate | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B12 | B13 | B14 | B15 | B16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 97.6% | 97.6% | 96.4% | 94.5% | 98.2% | 95.8% | 100.0% | 100.0% | 100.0% | 84.8% | 77.6% | 97.6% | 98.8% | 99.4% | 94.5% | 88.5% |
| After reconsideration | 99.4% | 98.8% | 98.2% | 98.8% | 98.8% | 98.2% | 100.0% | 100.0% | 99.4% | 87.9% | 87.3% | 99.4% | 99.4% | 99.4% | 96.4% | 91.5% |

| Agreement rate | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 89.9% | 92.0% | 90.2% | 83.3% | 89.1% | 93.5% | 86.2% | 87.0% | 97.8% | 88.8% | 75.0% |
| After reconsideration | 96.4% | 96.4% | 95.3% | 98.2% | 96.7% | 94.9% | 94.6% | 93.8% | 99.3% | 95.7% | 79.3% |

| Agreement rate | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 | D13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 78.6% | 91.3% | 90.2% | 96.4% | 97.5% | 97.8% | 85.5% | 99.6% | 90.9% | 82.2% | 96.7% | 91.3% | 76.8% |
| After reconsideration | 95.3% | 96.7% | 94.6% | 98.9% | 98.2% | 98.6% | 95.7% | 99.6% | 97.1% | 98.2% | 98.6% | 94.6% | 76.8% |

| Agreement rate | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 89.8% | 88.8% | 94.5% | 95.8% | 95.6% | 86.2% | 95.8% | 96.9% | 97.4% | 97.9% | 94.5% | 98.7% | 97.9% | 96.4% | 88.5% | 81.3% |
| After reconsideration | 97.9% | 97.4% | 97.7% | 99.0% | 98.2% | 99.2% | 98.4% | 99.2% | 98.7% | 99.5% | 99.0% | 99.2% | 99.5% | 97.7% | 90.9% | 83.1% |

| Agreement rate | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 | F17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 96.6% | 83.3% | 96.9% | 87.2% | 94.0% | 74.0% | 87.5% | 74.7% | 92.7% | 97.4% | 94.0% | 97.4% | 89.6% | 98.4% | 98.2% | 98.7% | 98.7% |
| After reconsideration | 99.0% | 97.9% | 99.0% | 99.5% | 100.0% | 99.7% | 100.0% | 100.0% | 99.5% | 98.4% | 97.4% | 99.2% | 98.2% | 99.2% | 99.2% | 99.5% | 100.0% |

| Agreement rate | F18 | F19 | F20 | F21 | F22 |
|---|---|---|---|---|---|
| Before reconsideration | 97.4% | 97.1% | 67.4% | 75.8% | 70.8% |
| After reconsideration | 99.0% | 99.0% | 94.5% | 99.0% | 72.9% |

| Agreement rate | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 93.5% | 94.6% | 97.3% | 42.2% | 38.1% | 95.6% | 98.6% | 95.7% | 99.6% | 98.5% | 89.7% | 94.3% | 94.1% | 96.4% | 99.8% | 99.9% | 96.3% |
| After reconsideration | 98.8% | 98.8% | 98.8% | 99.2% | 97.0% | 99.8% | 99.6% | 99.8% | 100.0% | 99.8% | 98.9% | 99.3% | 98.2% | 99.1% | 99.9% | 100.0% | 99.0% |

[Cohen's Kappa:]

| Kappa | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.68 | 0.34 | 0.19 | 0.18 | 0.15 | 0.23 | 0.53 | 0.31 | 0.42 | 0.22 | 0.48 | -0.02 | 0.41 | 0.38 | n.a. | 0.00 | 0.78 |
| After reconsideration | 0.94 | 0.71 | 0.89 | 0.84 | 0.94 | 0.72 | 0.78 | 0.89 | 0.92 | 0.95 | 1.00 | 0.83 | 0.64 | 0.85 | n.a. | 1.00 | 0.86 |

| Kappa | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B12 | B13 | B14 | B15 | B16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.00 | 0.59 | 0.48 | 0.76 | 0.39 | 0.44 | n.a. | n.a. | n.a. | 0.50 | 0.38 | 0.70 | 0.66 | 0.00 | 0.73 | 0.76 |
| After reconsideration | 0.85 | 0.83 | 0.81 | 0.95 | 0.66 | 0.81 | n.a. | n.a. | 0.00 | 0.63 | 0.69 | 0.94 | 0.85 | 0.00 | 0.82 | 0.83 |

| Kappa | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.17 | 0.41 | 0.18 | 0.35 | 0.27 | 0.22 | 0.31 | 0.45 | -0.01 | 0.56 | 0.50 |
| After reconsideration | 0.80 | 0.79 | 0.71 | 0.95 | 0.84 | 0.48 | 0.79 | 0.78 | 0.80 | 0.85 | 0.59 |

| Kappa | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 | D13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.40 | 0.39 | 0.29 | 0.36 | 0.35 | 0.39 | 0.50 | 0.00 | 0.50 | 0.26 | 0.38 | 0.64 | 0.50 |
| After reconsideration | 0.89 | 0.82 | 0.69 | 0.85 | 0.61 | 0.66 | 0.88 | 0.00 | 0.87 | 0.95 | 0.79 | 0.79 | 0.50 |

| Kappa | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.43 | 0.17 | 0.54 | 0.25 | 0.52 | 0.16 | 0.67 | 0.68 | 0.43 | 0.59 | 0.06 | 0.70 | -0.01 | 0.20 | 0.74 | 0.56 |
| After reconsideration | 0.91 | 0.87 | 0.84 | 0.88 | 0.84 | 0.97 | 0.89 | 0.93 | 0.79 | 0.92 | 0.89 | 0.84 | 0.85 | 0.60 | 0.79 | 0.61 |

| Kappa | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 | F17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.68 | 0.29 | 0.86 | 0.37 | 0.20 | 0.00 | 0.10 | -0.01 | 0.05 | -0.01 | 0.41 | 0.43 | 0.19 | -0.01 | 0.45 | -0.01 | 0.00 |
| After reconsideration | 0.92 | 0.94 | 0.96 | 0.98 | 1.00 | 0.99 | 1.00 | 1.00 | 0.96 | 0.56 | 0.81 | 0.88 | 0.91 | 0.72 | 0.82 | 0.75 | 1.00 |

| Kappa | F18 | F19 | F20 | F21 | F22 |
|---|---|---|---|---|---|
| Before reconsideration | -0.01 | -0.01 | 0.17 | -0.01 | 0.27 |
| After reconsideration | 0.71 | 0.77 | 0.88 | 0.33 | 0.31 |

| Kappa | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.19 | 0.15 | 0.28 | 0.01 | 0.02 | 0.00 | 0.31 | 0.09 | 0.50 | 0.56 | 0.12 | 0.00 | 0.31 | 0.45 | 0.33 | 0.83 | 0.00 |
| After reconsideration | 0.91 | 0.88 | 0.77 | 0.98 | 0.94 | 0.97 | 0.87 | 0.98 | 1.00 | 0.95 | 0.94 | 0.93 | 0.85 | 0.90 | 0.89 | 1.00 | 0.84 |

## (c) T-Voting-ST Treatment

[Agreement Rate:]

| Agreement rate | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 93.2% | 81.1% | 82.6% | 87.9% | 87.1% | 94.7% | 85.6% | 73.5% | 76.5% | 67.4% | 94.7% | 93.2% | 93.9% | 90.9% | 100.0% | 98.5% | 85.6% |
| After reconsideration | 98.5% | 94.7% | 93.2% | 97.0% | 94.7% | 98.5% | 94.7% | 93.9% | 99.2% | 93.9% | 98.5% | 98.5% | 97.0% | 97.0% | 100.0% | 100.0% | 92.4% |

| Agreement rate | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B12 | B13 | B14 | B15 | B16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 93.9% | 89.1% | 96.4% | 95.2% | 97.0% | 97.0% | 100.0% | 97.6% | 99.4% | 86.1% | 76.4% | 93.3% | 93.9% | 96.4% | 93.9% | 84.8% |
| After reconsideration | 97.6% | 97.6% | 98.8% | 98.2% | 98.8% | 99.4% | 100.0% | 98.2% | 99.4% | 91.5% | 91.5% | 95.8% | 97.0% | 99.4% | 95.2% | 86.7% |

| Agreement rate | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 93.2% | 93.4% | 89.4% | 87.9% | 91.7% | 95.2% | 88.9% | 90.9% | 95.7% | 91.9% | 82.8% |
| After reconsideration | 97.0% | 98.2% | 97.5% | 98.2% | 97.2% | 99.7% | 94.4% | 94.9% | 99.7% | 93.7% | 86.9% |

| Agreement rate | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 | D13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 84.3% | 95.5% | 96.7% | 96.5% | 97.0% | 97.5% | 84.1% | 99.0% | 89.1% | 82.6% | 94.2% | 90.7% | 80.6% |
| After reconsideration | 97.0% | 100.0% | 99.2% | 99.5% | 98.5% | 99.7% | 97.7% | 99.2% | 96.7% | 91.7% | 98.5% | 92.9% | 81.6% |

| Agreement rate | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 92.4% | 96.6% | 93.2% | 97.3% | 93.2% | 96.2% | 97.3% | 97.3% | 98.1% | 99.2% | 93.2% | 98.5% | 97.3% | 97.0% | 91.3% | 84.5% |
| After reconsideration | 98.9% | 99.2% | 97.7% | 99.2% | 97.0% | 99.6% | 98.9% | 98.9% | 100.0% | 100.0% | 98.5% | 99.6% | 98.5% | 98.5% | 93.2% | 86.0% |

| Agreement rate | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 | F17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 97.7% | 95.1% | 98.9% | 96.2% | 96.2% | 91.7% | 90.9% | 95.5% | 98.1% | 92.0% | 93.6% | 98.1% | 93.2% | 99.2% | 97.0% | 97.3% | 96.6% |
| After reconsideration | 99.2% | 98.5% | 99.2% | 98.9% | 100.0% | 100.0% | 100.0% | 100.0% | 99.2% | 94.3% | 98.1% | 99.6% | 98.9% | 100.0% | 99.2% | 98.9% | 100.0% |

| Agreement rate | F18 | F19 | F20 | F21 | F22 |
|---|---|---|---|---|---|
| Before reconsideration | 96.2% | 98.9% | 91.7% | 96.2% | 86.4% |
| After reconsideration | 97.7% | 98.9% | 92.8% | 98.9% | 87.1% |

| Agreement rate | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 92.9% | 96.0% | 97.4% | 48.8% | 46.6% | 97.4% | 98.6% | 99.1% | 99.1% | 97.5% | 89.7% | 92.5% | 94.6% | 96.3% | 99.3% | 99.6% | 94.9% |
| After reconsideration | 97.2% | 98.7% | 98.9% | 93.5% | 92.8% | 99.1% | 99.4% | 99.4% | 99.4% | 99.2% | 96.8% | 98.2% | 97.9% | 99.1% | 99.9% | 99.9% | 98.5% |

[Cohen's Kappa:]

| Kappa | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.79 | 0.28 | 0.40 | 0.15 | 0.30 | -0.03 | 0.49 | 0.27 | 0.54 | 0.19 | 0.51 | 0.27 | 0.47 | 0.41 | n.a. | 0.49 | 0.68 |
| After reconsideration | 0.96 | 0.84 | 0.81 | 0.86 | 0.78 | 0.83 | 0.84 | 0.86 | 0.98 | 0.87 | 0.89 | 0.89 | 0.78 | 0.85 | n.a. | 1.00 | 0.84 |

| Kappa | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | B12 | B13 | B14 | B15 | B16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.25 | 0.38 | 0.24 | 0.71 | -0.01 | 0.60 | n.a. | n.a. | n.a. | 0.60 | 0.18 | 0.23 | 0.34 | 0.00 | 0.79 | 0.69 |
| After reconsideration | 0.79 | 0.90 | 0.83 | 0.90 | 0.74 | 0.94 | n.a. | 0.39 | 0.00 | 0.77 | 0.77 | 0.61 | 0.75 | 0.92 | 0.84 | 0.73 |

| Kappa | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.31 | 0.29 | 0.24 | 0.31 | 0.53 | 0.08 | 0.49 | 0.46 | 0.00 | 0.81 | 0.62 |
| After reconsideration | 0.77 | 0.87 | 0.88 | 0.93 | 0.87 | 0.97 | 0.79 | 0.75 | 0.97 | 0.86 | 0.72 |

| Kappa | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 | D13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.22 | 0.09 | 0.54 | 0.21 | -0.01 | 0.00 | 0.23 | 0.33 | 0.50 | 0.42 | 0.35 | 0.80 | 0.56 |
| After reconsideration | 0.89 | 1.00 | 0.92 | 0.93 | 0.66 | 0.95 | 0.92 | 0.57 | 0.88 | 0.77 | 0.88 | 0.85 | 0.59 |

| Kappa | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.44 | 0.17 | 0.40 | -0.01 | 0.27 | -0.02 | 0.62 | 0.68 | 0.00 | 0.80 | 0.27 | 0.33 | 0.35 | 0.32 | 0.82 | 0.65 |
| After reconsideration | 0.94 | 0.89 | 0.85 | 0.83 | 0.76 | 0.95 | 0.86 | 0.88 | 1.00 | 1.00 | 0.89 | 0.89 | 0.71 | 0.77 | 0.86 | 0.69 |

| Kappa | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 | F17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.56 | 0.60 | 0.93 | 0.56 | 0.00 | 0.00 | 0.13 | 0.00 | -0.01 | 0.00 | 0.48 | 0.61 | -0.02 | 0.00 | 0.19 | -0.01 | 0.18 |
| After reconsideration | 0.89 | 0.90 | 0.96 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 0.75 | 0.42 | 0.88 | 0.94 | 0.90 | 1.00 | 0.87 | 0.72 | 1.00 |

| Kappa | F18 | F19 | F20 | F21 | F22 |
|---|---|---|---|---|---|
| Before reconsideration | -0.02 | 0.57 | 0.83 | -0.01 | 0.66 |
| After reconsideration | 0.56 | 0.57 | 0.85 | 0.82 | 0.68 |

| Kappa | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.08 | 0.00 | 0.15 | 0.01 | 0.02 | 0.08 | 0.21 | 0.00 | 0.57 | 0.32 | 0.05 | 0.04 | 0.41 | 0.40 | 0.56 | 0.00 | 0.00 |
| After reconsideration | 0.75 | 0.80 | 0.77 | 0.87 | 0.86 | 0.81 | 0.75 | 0.57 | 0.78 | 0.85 | 0.81 | 0.86 | 0.83 | 0.89 | 0.94 | 0.80 | 0.82 |

*C.4. Regression Results*

**(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 |
| c3 dummy | 0.140 | 0.166 | -0.144 | 0.223 | 0.404 | 0.255 |
| c4 dummy | 0.027 | 0.148 | 0.266 | 0.179 | -0.221 | 0.251 |
| 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 |
| c7 dummy | 0.132 | 0.152 | 0.015 | 0.191 | -0.047 | 0.246 |
| c8 dummy | -0.243 | 0.188 | -0.155 | 0.221 | 0.118 | 0.328 |
| c9 dummy | 0.461 | 0.294 | 0.710 | 0.490 | 0.640 | 0.400 |
| c10 dummy | 0.044 | 0.215 | -0.542** | 0.256 | 1.012*** | 0.369 |
| c11 dummy | 0.102 | 0.168 | -0.276 | 0.197 | 0.837*** | 0.293 |
| g1 dummy | 0.476** | 0.242 | 0.450* | 0.271 | 0.812 | 0.494 |
| g2 dummy | -0.198 | 0.266 | 0.512 | 0.320 | -0.282 | 0.461 |
| g3 dummy | -0.376 | 0.306 | -0.126 | 0.326 | -1.366** | 0.583 |
| g4 dummy | 0.073 | 0.141 | -0.136 | 0.170 | 0.285 | 0.224 |
| g5 dummy | -0.054 | 0.142 | -0.083 | 0.173 | -0.013 | 0.225 |
| g6 dummy | -0.794*** | 0.228 | -0.470* | 0.245 | -1.248*** | 0.437 |
| g7 dummy | 1.746* | 0.929 | 1.265 | 0.908 | 10.827 | 486.277 |
| g8 dummy | -0.167 | 0.618 | 0.079 | 0.603 | -0.243 | 1.161 |
| g9 dummy | -0.825 | 1.008 | (omitted)#1 | --- | -1.025 | 1.151 |
| g10 dummy | 0.078 | 0.268 | 0.247 | 0.304 | 0.049 | 0.462 |
| g11 dummy | -0.031 | 0.171 | -0.321 | 0.199 | 0.637* | 0.334 |
| g12 dummy | -0.214 | 0.194 | -0.184 | 0.242 | -0.354 | 0.311 |
| g13 dummy | 0.151 | 0.165 | 0.105 | 0.183 | 0.383 | 0.323 |
| g14 dummy | -0.095 | 0.242 | -0.049 | 0.275 | 0.001 | 0.440 |
| g15 dummy | 0.878 | 0.745 | (omitted) #1 | --- | 0.510 | 0.864 |
| g16 dummy | 4.196 | 130.348 | 3.303 | 66.432 | (omitted) #1 | --- |
| g17 dummy | 0.089 | 0.281 | 0.391 | 0.336 | -0.255 | 0.427 |
| phase2 dummy[#2] | -0.711*** | 0.165 | -0.830*** | 0.199 | -0.275 | 0.267 |
| phase3 dummy | -0.483*** | 0.156 | -0.688*** | 0.216 | -0.106 | 0.234 |
| phase4 dummy | -0.533*** | 0.143 | -0.765*** | 0.168 | -0.135 | 0.234 |
| phase5 dummy | -0.148 | 0.154 | -0.286 | 0.179 | -0.101 | 0.256 |
| Period within phases {= 1,2,3,4} | -0.021 | 0.042 | 0.009 | 0.053 | 0.005 | 0.066 |
| Constant | 1.235*** | 0.288 | 1.446*** | 0.349 | 0.277 | 0.477 |
| # 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 $\chi^2$ | 136.33 | --- | 75.53 | --- | 78.49 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] omitted due to collinearity. [#2] The reference group is observations in phase 6. *, **, 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* {= 0, 1, 2, …, or 20}

[When using codes with the definition of anti-social/pro-social punishment:]

| | (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 |
| f10 dummy | -4.014 | 4.139 | -32.299 | 709.925 | -3.816** | 1.645 |
| f11 dummy | -3.259** | 1.443 | -0.019 | 1.990 | -8.875*** | 1.507 |
| f12 dummy | 2.566 | 1.572 | -4.570* | 2.694 | 8.080*** | 1.620 |
| f13 dummy | -3.736** | 1.592 | -5.046* | 2.741 | 0.775 | 1.076 |
| f14 dummy | -0.972 | 3.247 | -1.107 | 4.573 | 4.838 | 143.091 |
| f15 dummy | 2.722* | 1.616 | 8.339*** | 2.891 | -1.728 | 1.496 |
| f16 dummy | 6.103*** | 2.100 | -5.132 | 3.805 | 9.854*** | 1.519 |
| f17 dummy | 0.815 | 1.646 | 2.939 | 2.498 | 1.131 | 1.024 |
| f18 dummy | 1.723 | 1.770 | 1.245 | 3.125 | 8.806*** | 1.965 |
| f19 dummy | 7.964*** | 1.741 | 6.153*** | 2.065 | 12.468*** | 2.736 |
| f20 dummy | -3.250** | 1.557 | -3.302 | 2.045 | -7.215*** | 2.134 |
| f21 dummy | ---[#2] | --- | ---[#2] | --- | -5.949** | 2.587 |
| f22 dummy | ---[#2] | --- | ---[#2] | --- | -10.364*** | 2.435 |
| g1 dummy | -0.415 | 1.187 | -1.629 | 1.853 | 2.032** | 0.934 |
| g2 dummy | -0.103 | 1.608 | 0.139 | 1.837 | 9.513*** | 1.811 |
| g3 dummy | 0.265 | 1.691 | -2.292 | 2.377 | 1.270 | 1.597 |
| g4 dummy | 1.838* | 0.993 | 1.408 | 1.266 | -0.363 | 0.641 |
| g5 dummy | -0.524 | 1.077 | -1.343 | 1.383 | -6.542*** | 1.084 |
| g6 dummy | 0.888 | 1.958 | -2.008 | 2.516 | -4.544 | 143.112 |
| g7 dummy | -0.010 | 1.750 | 3.420 | 2.185 | -8.293*** | 1.078 |
| g8 dummy | 2.413 | 1.852 | 1.593 | 2.233 | 6.168** | 2.465 |
| g9 dummy | -9.287 | 1276.133 | -12.646 | 6696.507 | -12.334 | . |
| g10 dummy | 0.784 | 2.260 | -4.494 | 4.166 | 9.718*** | 1.903 |
| g11 dummy | -0.052 | 1.307 | -1.710 | 1.982 | -2.583** | 1.034 |
| g12 dummy | -2.037 | 1.538 | 3.533 | 2.394 | -5.033*** | 1.124 |
| g13 dummy | 0.983 | 1.813 | -0.166 | 2.838 | 2.542* | 1.310 |
| g14 dummy | 1.352 | 1.627 | 6.676*** | 2.413 | -0.026 | 1.164 |
| g15 dummy | -16.834 | 1037.905 | 11.036 | 5118.976 | -15.709 | . |
| g16 dummy | -12.539 | 2030.692 | (omitted)[#1] | --- | -6.936 | . |
| g17 dummy | -0.002 | 2.642 | -0.851 | 3.068 | -38.307 | . |
| phase2 dummy[#3] | 1.381 | 1.387 | 3.649 | 2.375 | -0.433 | 1.216 |
| phase3 dummy | 0.932 | 1.347 | 3.808 | 2.326 | 4.755*** | 1.299 |
| phase4 dummy | 0.181 | 1.543 | 2.546 | 2.530 | -5.211*** | 1.262 |
| phase5 dummy | 1.573 | 1.421 | 3.687 | 2.431 | 3.977** | 1.971 |
| Period within phases {= 1,2,3,4} | -0.489 | 0.303 | 0.174 | 0.452 | -0.116 | 0.210 |
| Constant | -6.031*** | 2.005 | -9.701*** | 2.818 | -0.448 | 2.667 |
| # 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 $\chi^2$ | 172.55 | --- | 150.91 | --- | n.a. | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | n.a. | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] omitted due to collinearity. [#2] omitted because Kappas were less than 0.4 in the T-Voting treatment. [#3] The reference group is observations in phase 6. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

[When using codes with the definition of perverse/non-perverse punishment:]

| | (1) Pooled data | | (2) T-Voting | | (3) T-Voting-ST | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| f5 dummy | 5.566*** | 1.194 | 5.879*** | 1.431 | 5.297*** | 0.544 |
| f6 dummy | -0.628 | 1.045 | 0.247 | 1.295 | -1.781*** | 0.239 |
| f7 dummy | 8.067*** | 1.042 | 7.779*** | 1.166 | 10.494*** | 0.683 |
| f8 dummy | -3.396*** | 1.211 | -5.081*** | 1.348 | 6.742*** | 0.685 |
| f9 dummy | 3.585*** | 1.119 | 4.049*** | 1.380 | 11.588*** | 0.824 |
| f10 dummy | -5.079 | 4.121 | -36.714 | 1394.598 | -2.833*** | 0.899 |
| f11 dummy | -3.200** | 1.461 | -1.897 | 1.899 | -7.053*** | 0.805 |
| f12 dummy | 2.288 | 1.565 | -3.763 | 2.599 | 2.362*** | 0.429 |
| f13 dummy | -4.254*** | 1.573 | -6.942** | 2.743 | 0.289 | 0.213 |
| f14 dummy | 1.029 | 2.997 | 0.493 | 4.271 | 5.164 | 79.104 |
| f15 dummy | 4.045** | 1.571 | 6.880** | 2.736 | 4.085*** | 0.364 |
| f16 dummy | 4.908** | 2.042 | -4.467 | 3.757 | 9.434*** | 0.388 |
| f17 dummy | 0.014 | 1.662 | 0.322 | 2.528 | -0.518** | 0.214 |
| f18 dummy | -0.737 | 1.711 | 1.031 | 3.003 | -2.459*** | 0.318 |
| f19 dummy | 8.275*** | 1.757 | 6.693*** | 2.101 | 8.891*** | 0.984 |
| f20 dummy | -2.740* | 1.519 | -3.434* | 2.063 | -2.417*** | 0.923 |
| f21 dummy | ---[#2] | --- | ---#2 | --- | -0.770* | 0.404 |
| f22 dummy | ---[#2] | --- | ---#2 | --- | -3.374*** | 0.529 |
| g1 dummy | 0.251 | 1.145 | 0.000 | 1.776 | 1.349*** | 0.284 |
| g2 dummy | 1.937 | 1.566 | 2.281 | 1.799 | 2.753*** | 0.500 |
| g3 dummy | 1.439 | 1.667 | 0.840 | 2.397 | 0.020 | 0.302 |
| g4 dummy | 1.689* | 0.983 | 1.367 | 1.244 | -0.763*** | 0.195 |
| g5 dummy | 0.183 | 1.036 | 0.796 | 1.289 | -4.642*** | 0.316 |
| g6 dummy | 1.390 | 1.942 | 0.346 | 2.285 | 4.188 | 79.108 |
| g7 dummy | -0.308 | 1.795 | 2.490 | 2.324 | -14.875*** | 0.388 |
| g8 dummy | -0.009 | 1.678 | -1.430 | 1.956 | -5.168*** | 0.873 |
| g9 dummy | -8.823 | 1242.918 | -12.782 | 5496.452 | -0.741 | . |
| g10 dummy | 2.456 | 2.112 | -0.484 | 3.787 | 6.521*** | 0.877 |
| g11 dummy | 0.178 | 1.254 | -0.950 | 1.871 | 3.170*** | 0.198 |
| g12 dummy | -1.105 | 1.509 | 2.003 | 2.302 | -4.409*** | 0.328 |
| g13 dummy | 1.387 | 1.715 | 1.276 | 2.589 | -0.685** | 0.326 |
| g14 dummy | 1.012 | 1.599 | 5.811** | 2.361 | -6.661*** | 0.234 |
| g15 dummy | -14.488 | 1149.064 | 12.931 | 24570.070 | 0.438 | . |
| g16 dummy | -11.939 | 2158.098 | (omitted) #1 | --- | 2.190 | . |
| g17 dummy | -1.925 | 2.945 | -3.338 | 3.600 | -6.662 | . |
| phase2 dummy[#3] | 1.861 | 1.376 | 3.415 | 2.300 | 3.146*** | 0.297 |
| phase3 dummy | 0.939 | 1.335 | 2.877 | 2.250 | -0.001 | 0.509 |
| phase4 dummy | 0.544 | 1.525 | 2.367 | 2.425 | 0.748 | 0.461 |
| phase5 dummy | 2.372* | 1.419 | 3.792 | 2.351 | 1.662*** | 0.326 |
| Period within phases {= 1,2,3,4} | -0.350 | 0.301 | -0.062 | 0.432 | 0.483*** | 0.062 |
| Constant | -7.124*** | 2.018 | -9.398*** | 2.712 | -4.370*** | 0.939 |
| # of observations | 648 | --- | 384 | --- | 264 | --- |
| # of left-censored observations (0) | 535 | --- | 315 | --- | 220 | --- |
| # of right-censored observations (20) | 5 | --- | 3 | --- | 2 | --- |
| Log likelihood | -362.239 | --- | -212.749 | --- | -58.9492 | --- |
| Wald $\chi^2$ | 170.52 | --- | 154.79 | --- | n.a. | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | n.a. | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] Omitted due to collinearity. [#2] Omitted because Kappas were less than 0.4 in the T-Voting treatment. [#3] The reference group is observations in phase 6. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

## (c) Team contribution decisions

### c.1. Contribution decisions in the FS scheme

Dependent variable: contribution amount of team *i* in period *t* in the FS scheme {= 0, 1, …, or 20}

| | (1) Pooled data | | (2) T-Voting | | (3) T-Voting-ST | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| d1 dummy | 8.916*** | 2.049 | 9.381*** | 2.427 | 6.428** | 3.226 |
| d2 dummy | -6.524*** | 2.107 | -5.688** | 2.380 | -6.073* | 3.589 |
| d3 dummy | 2.380 | 2.594 | 6.734* | 3.752 | -5.595 | 3.744 |
| d4 dummy | -12.415*** | 2.700 | -14.238*** | 4.049 | -12.204*** | 3.778 |
| d5 dummy | 0.197 | 3.174 | 3.935 | 3.414 | -3.426 | 4.691 |
| d6 dummy | -3.670 | 3.636 | 1.539 | 8.667 | -5.170 | 5.312 |
| d7 dummy | 1.511 | 1.757 | 3.146 | 2.276 | -1.052 | 2.572 |
| d8 dummy | ---[#1] | | ---[#1] | | -22.120* | 12.208 |
| d9 dummy | 9.203*** | 2.388 | 11.086*** | 2.961 | 5.333 | 3.381 |
| d10 dummy | -2.624 | 2.046 | -7.232*** | 2.635 | 0.367 | 3.257 |
| d11 dummy | -4.381* | 2.392 | -8.087** | 3.979 | -3.171 | 2.952 |
| d12 dummy | 19.345*** | 3.441 | 11.902*** | 4.208 | 17.299*** | 5.021 |
| d13 dummy | 17.095*** | 2.546 | 12.184*** | 3.056 | 13.907*** | 3.788 |
| g1 dummy | -0.352 | 2.397 | -1.234 | 2.970 | 2.743 | 3.750 |
| g2 dummy | -10.301*** | 2.354 | -5.209** | 2.364 | -14.793*** | 4.927 |
| g3 dummy | -8.155** | 3.187 | -6.883** | 3.043 | -1.971 | 6.197 |
| g4 dummy | 0.239 | 1.680 | 0.853 | 1.764 | -4.764* | 2.883 |
| g5 dummy | -0.376 | 1.765 | -3.169 | 2.198 | 5.303* | 2.954 |
| g6 dummy | -1.524 | 3.683 | -2.571 | 4.734 | 7.206 | 5.735 |
| g7 dummy | -16.782*** | 5.939 | -19.264* | 11.075 | -4.026 | 9.304 |
| g8 dummy | -1.643 | 7.339 | -13.266 | 10.500 | 50.032 | 5649.600 |
| g9 dummy | -3.461 | 5.377 | -9.425 | 11.787 | 0.858 | 6.800 |
| g10 dummy | 0.450 | 5.973 | 53.275 | 1702.725 | -3.305 | 6.651 |
| g11 dummy | 0.236 | 2.602 | -0.348 | 3.309 | 1.837 | 3.912 |
| g12 dummy | -0.939 | 2.626 | 9.298 | 5.613 | -6.561* | 3.551 |
| g13 dummy | 1.792 | 3.016 | -3.345* | 3.169 | 16.118** | 7.025 |
| g14 dummy | -0.532 | 3.846 | 6.781 | 6.429 | 2.178 | 5.525 |
| g15 dummy | 1.041 | 6.307 | -2.677 | 7.288 | 41.294 | 1874.158 |
| g16 dummy | -3.724 | 9.433 | 24.708 | 4006.715 | -14.400 | 14.338 |
| g17 dummy | -3.741 | 3.394 | -1.572 | 3.901 | 1.455 | 6.821 |
| phase2 dummy[#2] | 2.760 | 2.025 | 1.733 | 2.514 | 8.467*** | 3.165 |
| phase3 dummy | -0.213 | 2.283 | -4.379 | 2.936 | 5.655 | 3.646 |
| phase4 dummy | 2.075 | 1.896 | 0.130 | 1.898 | 7.011** | 3.321 |
| phase5 dummy | -0.152 | 1.991 | -2.187 | 2.011 | 6.086 | 3.914 |
| Period within phases {= 1,2,3,4} | 0.195 | 0.565 | 0.508 | 0.635 | 0.577 | 0.939 |
| Constant | 18.079*** | 3.215 | 19.117*** | 4.100 | 15.140*** | 4.775 |
| # of observations | 672 | --- | 276 | --- | 396 | --- |
| # of left-censored observations (0) | 26 | --- | 15 | --- | 11 | --- |
| # of right-censored observations (20) | 536 | --- | 195 | --- | 341 | --- |
| Log likelihood | -507.868 | --- | -267.685 | --- | - 201.987 | --- |
| Wald $\chi^2$ | 212.50 | --- | 160.79 | --- | 87.47 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] Omitted because Kappas were less than 0.4 in the T-Voting treatment. [#2] The reference group is observations in phase 6. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

## c.2. Contribution decisions in the IS scheme

Dependent variable: contribution amount of team *i* in period *t* in the IS scheme {= 0, 1, …, or 20}

| | (1) Pooled data | | (2) T-Voting | | (3) T-Voting-ST | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| e1 dummy | 3.015** | 1.201 | 4.284*** | 1.259 | 4.842 | 3.384 |
| e2 dummy | -7.058*** | 1.143 | -6.276*** | 1.142 | -7.466** | 3.285 |
| e3 dummy | 5.558*** | 1.570 | 5.700*** | 1.709 | 14.813*** | 5.349 |
| e4 dummy | -7.788*** | 1.608 | -5.453*** | 1.687 | -14.852*** | 5.547 |
| e5 dummy | 0.886 | 1.303 | 2.042 | 1.453 | -6.524** | 3.291 |
| e6 dummy | 0.393 | 1.125 | 1.463 | 1.089 | 23.590 | 903.617 |
| e7 dummy | -1.075 | 1.250 | -0.582 | 1.263 | -8.604** | 4.292 |
| e8 dummy | 2.412* | 1.323 | 3.106** | 1.390 | 13.494** | 5.809 |
| e9 dummy | -1.466 | 1.586 | 0.383 | 1.628 | -10.694** | 4.646 |
| e10 dummy | 1.955 | 1.702 | 0.135 | 1.770 | 8.358* | 4.291 |
| e11 dummy | 0.500 | 1.396 | 1.253 | 1.576 | -4.601 | 4.032 |
| e12 dummy | 3.567* | 2.143 | 3.048 | 2.246 | 2.181 | 4.312 |
| e13 dummy | -6.793*** | 2.161 | -3.878 | 2.591 | -0.341 | 4.336 |
| e14 dummy | 0.205 | 1.578 | 0.038 | 1.923 | -0.721 | 3.865 |
| e15 dummy | 7.602*** | 1.941 | 6.068* | 2.367 | 5.023 | 4.904 |
| e16 dummy | 3.004*** | 1.074 | 1.841 | 1.134 | 18.976*** | 4.595 |
| g1 dummy | 1.664 | 1.385 | 1.582 | 1.416 | -2.057 | 6.148 |
| g2 dummy | -4.334*** | 1.269 | -2.488* | 1.319 | -8.709* | 5.183 |
| g3 dummy | 0.104 | 1.600 | -0.411 | 1.729 | 10.553** | 5.139 |
| g4 dummy | -1.633 | 1.050 | -1.314 | 1.080 | -7.651 | 5.199 |
| g5 dummy | -3.273*** | 1.014 | -2.466** | 1.023 | -6.119* | 3.557 |
| g6 dummy | -0.892 | 1.706 | -2.641 | 1.701 | 7.006 | 7.276 |
| g7 dummy | 0.988 | 2.667 | -0.891 | 4.474 | 12.631 | 8.005 |
| g8 dummy | 0.040 | 1.629 | -0.578 | 1.496 | 18.070 | 7938.304 |
| g9 dummy | 1.751 | 5.114 | 4.008 | 5.918 | 24.931 | 17439.750 |
| g10 dummy | -2.816 | 2.279 | -0.189 | 2.791 | 4.762 | 5.671 |
| g11 dummy | 3.156*** | 1.161 | 2.252* | 1.242 | 10.179** | 4.036 |
| g12 dummy | -3.597*** | 1.343 | -4.223*** | 1.488 | -2.399 | 4.184 |
| g13 dummy | -0.569 | 1.410 | 0.182 | 1.627 | -7.741* | 4.501 |
| g14 dummy | 0.563 | 1.258 | 1.648 | 1.392 | -4.650 | 2.942 |
| g15 dummy | 36.392 | 1919.199 | 30.194 | 963.699 | 16.201 | 9100.331 |
| g16 dummy | 12.004 | 3323.158 | (omitted)[#1] | | -21.635 | 33428.210 |
| g17 dummy | 2.141 | 2.788 | 2.963 | 3.815 | -4.780 | 5.790 |
| phase2 dummy[#2] | -0.122 | 1.389 | 1.548 | 1.494 | -16.173*** | 5.033 |
| phase3 dummy | 1.801 | 1.351 | 1.953 | 1.450 | -10.693* | 5.631 |
| phase4 dummy | -1.063 | 1.353 | -0.292 | 1.419 | -0.882 | 8.634 |
| phase5 dummy | 0.560 | 1.352 | 1.495 | 1.475 | -9.319* | 4.637 |
| Period within phases {= 1,2,3,4} | -0.726** | 0.311 | -0.272 | 0.327 | -2.730*** | 0.783 |
| Constant | 24.582*** | 2.343 | 20.872*** | 2.744 | 44.670*** | 8.953 |
| # of observations | 648 | --- | 384 | --- | 264 | --- |
| # of left-censored observations (0) | 17 | --- | 12 | --- | 5 | --- |
| # of right-censored observations (20) | 473 | --- | 246 | --- | 227 | --- |
| Log likelihood | -588.738 | --- | -433.879 | --- | -118.003 | --- |
| Wald $\chi^2$ | 254.42 | --- | 203.11 | --- | 108.58 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] Omitted due to collinearity. [#2] The reference group is observations in phase 6. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

### c.3. Contribution decisions without sanctioning scheme

Dependent variable: contribution amount of team *i* in period *t* without schemes {= 0, 1, …, or 20} (Since the institutional environment is identical for all three team treatments when there is no scheme, data [T-No treatment and Phase 1 of the T-Voting and T-Voting-ST treatments] is pooled in the regression).

| | Coefficient | S.E. |
|---|---|---|
| a1 dummy | 1.137* | 0.685 |
| a2 dummy | 4.954*** | 0.617 |
| a3 dummy | 5.428*** | 0.612 |
| a4 dummy | -4.098*** | 0.625 |
| a5 dummy | -4.429*** | 0.714 |
| a6 dummy | -3.534*** | 0.980 |
| a7 dummy | -3.771*** | 0.708 |
| a8 dummy | 0.023 | 0.506 |
| a9 dummy | -0.256 | 0.543 |
| a10 dummy | -0.242 | 0.485 |
| a11 dummy | 2.282*** | 0.730 |
| a12 dummy | -0.650 | 0.771 |
| a13 dummy | 0.816 | 0.719 |
| a14 dummy | -0.518 | 0.867 |
| a16 dummy | 0.855 | 1.182 |
| a17 dummy | 2.239*** | 0.494 |
| g1 dummy | 3.651*** | 0.673 |
| g2 dummy | -3.370*** | 0.651 |
| g3 dummy | 0.376 | 0.634 |
| g4 dummy | 0.138 | 0.705 |
| g5 dummy | -0.156 | 0.482 |
| g6 dummy | -6.615** | 3.095 |
| g7 dummy | -1.235 | 0.968 |
| g8 dummy | -4.117* | 2.183 |
| g9 dummy | -1.818 | 1.752 |
| g10 dummy | -0.193 | 1.224 |
| g11 dummy | -0.325 | 0.580 |
| g12 dummy | -0.434 | 0.829 |
| g13 dummy | 0.105 | 0.827 |
| g14 dummy | 0.946 | 0.754 |
| g15 dummy | -2.180 | 3.430 |
| g16 dummy | -1.627 | 1.880 |
| g17 dummy | -0.532 | 3.394 |
| phase1 dummy | 6.508*** | 0.766 |
| phase2 dummy | 4.741*** | 0.776 |
| phase3 dummy | 3.723*** | 0.772 |
| phase4 dummy | 4.739*** | 0.763 |
| phase5 dummy[#1] | 3.012*** | 0.751 |
| Period within phases {= 1,2,3,4} | -0.536*** | 0.177 |
| Constant | 7.704*** | 1.295 |
| # of observations | 1128 | --- |
| # of left-censored observations (0) | 170 | --- |
| # of right-censored observations (20) | 253 | --- |
| Log likelihood | -2636.596 | --- |
| Wald $\chi^2$ | 749.45 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] The reference group is observations in phase 6. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

**(d) Team scheme choice**

Dependent variable: scheme vote of team *i* in phase *k* {= 1(0) for voting in favor of FS(IS)}

| | (1) Pooled data | | (2) T-Voting | | (3) T-Voting-ST | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| b1 dummy | 1.329 | 0.920 | (omitted)[#1] | | 2.581 | 1.997 |
| b2 dummy | 2.233*** | 0.793 | (omitted)[#1] | | 3.322* | 1.843 |
| b3 dummy | 2.832*** | 1.049 | 5.600** | 2.326 | (omitted)[#1] | |
| b4 dummy | -1.230** | 0.592 | -2.048 | 1.255 | -1.637 | 1.652 |
| b5 dummy | (omitted)[#1] | | (omitted)[#1] | | (omitted)[#1] | |
| b6 dummy | -0.971 | 1.076 | 2.336 | 2.148 | -3.422 | 2.396 |
| b10 dummy | 0.377 | 0.562 | 1.531 | 1.213 | 0.283 | 1.259 |
| b11 dummy | 1.206*** | 0.442 | 2.245* | 1.167 | 0.289 | 0.722 |
| b12 dummy | -0.501 | 0.625 | 0.192 | 1.222 | 0.971 | 1.501 |
| b13 dummy | 0.581 | 0.779 | 0.881 | 1.939 | 0.943 | 1.677 |
| b14 dummy | ---[#2] | | ---[#2] | | 0.396 | 1.643 |
| b15 dummy | 0.381 | 0.665 | 0.294 | 1.451 | 0.532 | 1.500 |
| b16 dummy | 0.297 | 0.525 | 0.100 | 1.164 | 0.495 | 1.229 |
| g1 dummy | -3.318*** | 1.268 | (omitted)[#1] | | -8.130** | 3.696 |
| g2 dummy | 1.061 | 1.289 | 1.739 | 1.775 | 2.355 | 7.052 |
| g3 dummy | (omitted)[#1] | | 0.000 | (omitted) | (omitted)[#1] | |
| g4 dummy | 0.239 | 0.342 | 0.469 | 0.671 | 0.372 | 0.697 |
| g5 dummy | 0.096 | 0.382 | 0.150 | 0.866 | -0.089 | 0.692 |
| g6 dummy | -1.025 | 0.932 | (omitted)[#1] | | -1.455 | 1.517 |
| g7 dummy | (omitted)[#1] | | (omitted)[#1] | | (omitted)[#1] | |
| g8 dummy | -2.250** | 1.072 | (omitted)[#1] | | -4.251 | 5.348 |
| g9 dummy | -1.079 | 1.615 | (omitted)[#1] | | -2.159 | 2.184 |
| g10 dummy | -0.451 | 1.012 | -0.174 | 1.604 | 1.030 | 5.137 |
| g11 dummy | -0.100 | 0.560 | 0.500 | 1.031 | -2.031 | 1.651 |
| g12 dummy | -0.009 | 0.754 | -0.239 | 1.348 | 1.581 | 2.182 |
| g13 dummy | 0.112 | 0.949 | 35.186 | 29409.480 | -0.743 | 2.227 |
| g14 dummy | -0.053 | 0.902 | -1.599 | 1.585 | 1.817 | 2.105 |
| g15 dummy | (omitted)[#1] | | (omitted)[#1] | | (omitted)[#1] | |
| g16 dummy | (omitted)[#1] | | (omitted)[#1] | | (omitted)[#1] | |
| g17 dummy | 1.172 | 0.765 | -23.351 | 29376.130 | 1.785 | 1.307 |
| phase2 dummy | 0.297 | 0.592 | -1.899 | 1.307 | 1.504 | 1.135 |
| phase3 dummy | -1.426*** | 0.461 | -2.894*** | 1.018 | -0.686 | 0.911 |
| phase4 dummy | -0.494 | 0.374 | -1.167 | 0.830 | 0.540 | 0.708 |
| phase5 dummy | -0.627* | 0.379 | -0.807 | 0.792 | -0.337 | 0.617 |
| Constant | -0.525 | 0.657 | -1.274 | 1.355 | -0.231 | 1.498 |
| # of observations | 317 | --- | 137 | --- | 153 | --- |
| Log likelihood | -134.894 | --- | -47.914 | --- | -63.803 | --- |
| Wald $\chi^2$ | 41.51 | --- | 16.25 | --- | 12.55 | --- |
| Prob > Wald $\chi^2$ | 0.048 | --- | 0.803 | --- | 0.995 | --- |

*Notes*: Probit regressions. Decision-making unit random effects were included to control for the panel structure. [#1] Omitted due to collinearity. [#2] Omitted because Kappas were less than 0.4. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

## Appendix A.D: Implementation and Sample Instructions Used in the Experiment

*D.1. Implementation*

The experiment was conducted at the EXEC laboratory in the University of York from July 2019 through January 2020.[65] Observations of 11 or 12 groups were collected for each treatment condition by conducting six or seven (two) sessions in each Team (Individual) 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 2.1 of the paper. 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.[66] 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.

**References:**

Bock, O., Ingmar B., & Andreas, N. (2014) hroot: Hamburg Registration and Organization Online Tool. *Eur. Econ. Rev., 71*, 117-120.

Fischbacher, U. (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ. 10*(2), 171-178.

[65] All the sessions were completed before the Covid-19 pandemic (which began in March 2020 in the United Kingdom).

[66] 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-M and I-Voting-ST treatments (the T-No, T-Voting-M 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).

*D.2. Sample Instructions*

## [Individual Treatments:]

Part 1 is the same for the three individual treatments (the I-No, I-Voting-M and I-Voting-ST treatments). At the beginning of the experiment, the following instructions were read aloud to subjects. The subjects were also given printed copies of the instructions:

## Instructions

You are now taking part in a decision-making experiment. Depending on your decisions and the decisions of other participants, you will be able to earn money in addition to the £3 guaranteed for your participation. Please read the following instructions carefully.

During the experiment you are not allowed to communicate with other participants. If you have questions, raise your hand. One of us will come to answer your question.

During the experiment your earnings will be calculated in points. At the end of the experiment your points will be converted to U.K. pounds at the following rate:

**62.5 points = £1**

(or each point will be exchanged for 1.6 pence of real money). At the end of the experiment your total earnings (including the **£3** participation fee) will be paid out to you in cash. Your payment will be rounded to the nearest 10 pence (e.g., £15.30 if it is £15.33; and £15.40 if it is £15.37).

At the beginning of the experiment, you are randomly assigned to a group of three and interact with each other. **You will be part of the same group throughout the entire experiment**. This experiment consists of two parts. Part 1 consists of 4 periods, while Part 2 has five phases each consisting of 4 periods (in total, 20 periods for Part 2). Thus there are six phases (a total of 24 periods) in the experiment. In each period you will be required to make at least one decision. The time allocated for each decision is displayed in the top right corner of the screen in seconds. Please make your decision and click the submit/continue button before the timer reaches 0.

We will first explain the detail of Part 1. We will distribute the instructions for Part 2 once Part 1 is over.

## PART 1

### *Your decision in each period:*

In each period, you and your two group members are each given **an endowment of 20 points** and simultaneously make allocation decisions. There are two accounts to allocate points to, the **private account and the group account**. Specifically, you are asked how many points you want to allocate to the group account. The remaining points (that is, 20 minus your allocation to the group account) will be automatically allocated to your private account. Your earnings in a given period depend on **(a) the number of points in your private account and (b) the total amount allocated to the group account**.

### *How to calculate your earnings:*

Your earnings in a given period are calculated as in the following formula:

**(sum of points in your private account) + 0.6 × (sum of points allocated by you and your group members to the group account)**

In other words, your earnings from your private account **are equal to the number of points you allocated to the private account** (20 minus your allocation to your group account). The points you allocate to your private account do not affect the earnings of your group members.

By contrast, your earnings from the group account equal the **sum** of points allocated to the group account by you and your two group members multiplied **by 0.6**. In other words, if you allocate 1 point to the group account, your earnings from your allocation is $0.6 \times 1 = 0.6$ points, which is less than 1 point. However, by allocating 1 point to the group account, the earnings of each of your group members also increase by 0.6 points. Therefore, the total earnings in this case are 1.8 points, which is greater than 1 point. Note that you also obtain earnings of 0.6 points for each point your other group members allocate to your group account.

Once all group members make allocation decisions, you will be informed of the interaction outcomes (your earnings, along with each of the two group members' allocation decisions anonymously and in a random order).

If you have any questions, please raise your hand. When all questions are answered, we will move on to comprehension questions.

*Comprehension questions*

Please answer the following questions. Raise your hand if you need help.

**1.** Suppose that all three members in your group allocate 0 points to the group account. How much does each member earn? _____

**2.** Suppose that all three members in your group allocate 20 points to the group account. How much does each member earn? _____

**3.** Suppose that the other two members in your group in total allocate 15 points to the group. Answer the following:

a) How much do you earn if you allocate 0 points to the group account? _____

b) How much do you earn if you allocate 10 points to the group account? _____

c) How much do you earn if you allocate 20 points to the group account? _____

Any questions? When all questions are answered, we will move on to Part 1.

[Once everyone finished answering the comprehension questions and the experimenter explained the answers, the experiment began.]

As soon as Part 1 was over, subjects moved on to Part 2. Part 2 differ by treatment.

**I-No treatment:** At the onset of Part 2, the following instructions were distributed and were read aloud (subjects also had printed copies of the instructions):

## Instructions for Part 2

As explained, you have **five** phases each consisting of 4 periods (in total, 20 periods) in Part 2.

The five phases are each separated by a break of 40 seconds. The structure of each phase is identical to that of Part 1 (1 point = 1.6 pence).

The group composition in Part 2 is the same as Part 1.

If you have any questions, please raise your hand. When all questions are answered, we will start Part 2.

**I-Voting-M treatment:** At the onset of Part 2, the following instructions were distributed and were read aloud (subjects also had printed copies of the instructions):

## Instructions for Part 2

As explained, you have **five** phases each consisting of 4 periods (in total, 20 periods) in Part 2.

You will continue to interact with the same two individuals. In each period you will make a decision about allocating 20 points to either a private account or a group account, with the same immediate payment consequence (see the instructions for Part 1). The conversion rate is the same: 1 point = 1.6 pence.

However, there is a significant difference in that each period consists of two stages. In the first stage, you make your allocation decision and learn the decisions of the other group members along with your earnings. In the second stage, your earnings from the allocation stage can be reduced. There are two possible schemes governing the second stage of each period. **At the beginning of each phase, your group will determine by majority vote which of the two schemes will be used during the four periods of that phase.** You can select different schemes in different phases.

Of the two possible schemes, one is a scheme in which the group votes on the rules of a fine (which we call "**Group-determined fines**"); the other is a scheme in which individuals can reduce others' earnings after learning of their allocations (which we call "**Individual reduction decisions**").

### Scheme 1: Group-determined fines

In this scheme, earnings from the allocation stage can be reduced by a fine rule. When a rule is in place, allocations to the **private** account are subject to a fine.

At the beginning of each period, your group chooses a fine rate (the amount of the fine per point allocated to the private account) by voting. Possible fine rates are 0, 0.2, 0.4, 0.6, 0.8, 1.0 and 1.2 points per point allocated to the private account.

For each point that is lost by a member who is fined, the group also incurs a cost of 0.6 points to impose that fine. This cost is interpreted as an administrative cost in imposing a fine. For example, if an individual is fined a total of 5 points, this costs the group 3 (=5×0.6) points, with each group member (including the fine recipient) being equally charged 1 (=3/3) point as his or her share of that cost. More generally, for each 1 point of fines imposed on any group member, each group member pays 0.6*(1/3) = 0.2 points as his or her per capita cost of imposing the fine. In the example of an

individual fined 5 points, that individual thus loses both the 5 points and his or her per capita share of the cost, 1 point, for a total loss of 6 (= 5 + 1) points. Notice that since the person fined loses a total of 6 points while the other group members pay 2 points in the aggregate (i.e., 2×1) in imposing the fine, **the ultimate cost ratio is 1:3 (= 2:6)**.

*Fixed charge:* In addition to the fines and costs based on the fine rule your group chooses, at the end of a period, a fixed cost of 4 points is also deducted from the earnings of each group member. This can be thought of as the fixed administrative cost of having a fine scheme in operation, a cost that doesn't depend on how frequently or infrequently fines are in fact imposed.

Fines in the present phase cannot bring an individual's earnings for a period to less than zero. However, you always incur the per capita share of the cost of imposing fines and the fixed charge, even if it brings one's earnings for the period to less than zero.

This means that your earnings for a period can be calculated as follows:

Part A: Earnings from the allocation stage minus your fine, or 0 if it is negative

**minus**

Part B: Your part of the cost of administering the fine scheme
= your per capita share of imposing fine {= 0.2×total fines imposed} + 4

As mentioned, you incur the cost of Part B even if it causes your net earnings for the period to be negative.

The fine rate in a given period will be determined based on the median of three votes casted by group members. For example, if three members enter choices of 0.6, 0, and 0.2 as their preferred rate, then the fine rate will be 0.2 in your group. For another example, if three members enter choices of 0.4, 0, and 0.4 as their preferred rate, then the fine rate will be 0.4.

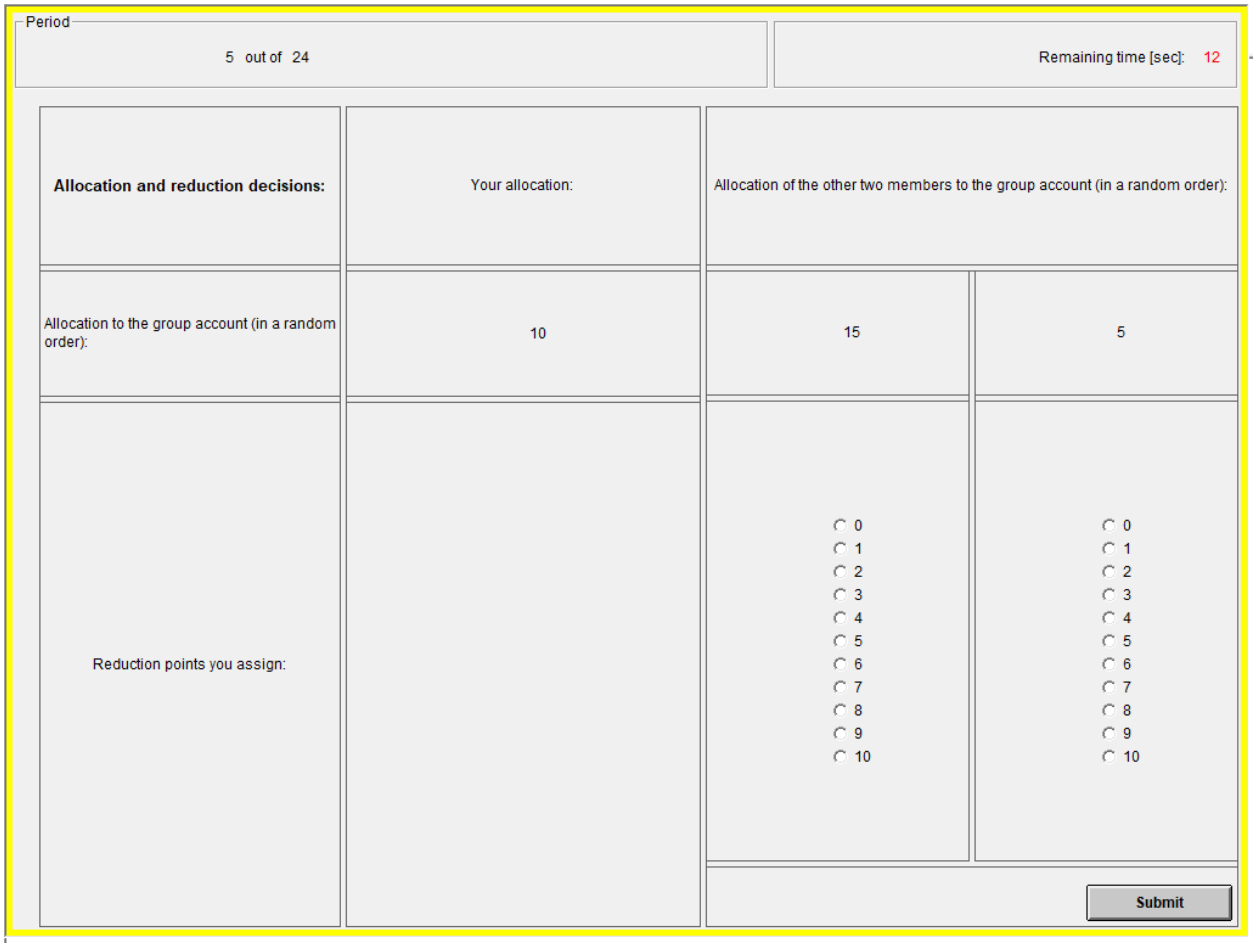
Note that there is effectively no fine if your group chooses a fine rate of 0. Also, if the fine rate is positive, earnings at the end of a period may be unchanged from those at the end of the allocation stage if no member allocates points to the private account.

Summary: In each period of this phase, your group will first vote on the fine rate. You will be informed of the vote outcome, and will then decide how to allocate between your private and group account.

## Scheme 2: Individual reduction decisions

In this scheme, you have an opportunity in stage 2 of each period to reduce the earnings of others in your group at a cost to your own earnings. You can assign reduction points to each of your group members.

Each reduction point you allocate to reducing another's earnings reduces your own earnings by 1 point and reduces that individual's earnings by 3 points. Thus, the **cost ratio is 1:3** as in the Group-determined fines explained above. Your own earnings can be reduced in the same way by the decisions of others in your group. You are free to leave any or all others' earnings unchanged by entering 0's in the relevant boxes.




*Note*: Numbers are for illustration only

Earnings reductions directed at you cannot bring your earnings for the period to less than zero. However, you always incur the cost of giving reductions to others even if it makes your period earnings negative. (If you lose points in a period, they are deducted from those you accumulate in other periods.) Thus, earnings in each period of this phase can be calculated as follows:

Part A: Earnings from the allocation stage minus reductions by others in your group, or 0 if it is negative

**minus**

Part B: Points you use to reduce others' earnings

Note that you incur the cost in Part B even if it causes your net earnings for the period to be negative.

Example: Suppose that you use 0 points to reduce the earnings of the first group member whose allocation appears on the screen, and you use 1 point to reduce the earnings of the second. Suppose further that these two individuals use 1, and 3 points to reduce your earnings. Then the second individual's earnings for the period will be reduced by 3 points in addition to any reductions due to the decisions of the third individual. Your own earnings for the period will be reduced by 1 point [i.e., your cost to impose reductions on others], plus (1×3)+(3×3)=12 points [i.e., the reductions imposed on your earnings by others]. At the end of the reduction stage, you will learn that others decided to reduce your earnings by a total of 12 points, but you will not be told which individuals

reduced your earnings or by how much any given individual reduced your earnings. Others will also not know who in particular reduced their earnings, or by how much.

In addition to the fact that earnings from the allocation stage and reductions received cannot go below zero, the earnings reduction process is subject to two requirements. First, your reduction points must be an integer. Second, you cannot assign more than 10 reduction points to any one individual in your group.

Remember that if no reductions are imposed (the reduction boxes are filled in with 0's), earnings after the reduction stage are the same as those before it.

## Summary of Part 2 (phases 2 to 6):

At the beginning of the first period in every 4 period phase, you will vote on two schemes: "Group-determined fines" versus "Individual reduction decisions."

Whichever scheme gets the most votes (≥ 2 votes) will be in effect for four periods.

(i) When Group-determined fines is chosen:

In each period, you will vote on the fine rate. The median vote is used for the fine rate in your group. Under the chosen fine rate, you and your group members simultaneously decide allocations between your private and group accounts.

(ii) When Individual reduction decision is chosen:

In each period, you will make your decision on allocating points to your private or group account. After that, you will make a decision about whether to reduce the earnings of others or not, and by what amount you reduce them if so.

You will vote **5 times** in total on the scheme to be used by your group—once for each of phases 2 – 6.

*Comprehension questions:*

Please answer the following questions. Raise your hand if you need help.

1. About voting between the two schemes:

a) How many periods do you have in Part 2 of the experiment? _____

b) How many times do you have the opportunity to vote on which scheme is used?

_____

__

c) If your group selects the scheme of group-determined fines in Phase 3 (periods 9 – 12), can it select a different scheme in Phase 4 (periods 13 – 16)? _____

2. Suppose that the scheme of group-determined fines is in place in a given phase.

a) What is the fixed charge each period for operating the fine scheme? _____

b) Suppose that the three votes on fine rate in your group are: 0.2, 0.6, 0. What is the fine per point in your group? _____

c) Suppose that your group selected a fine rate of 0.4, and suppose that you allocate 15 points to the group account. How many points will you lose in the form of a fine*? _____ points
* Note: do not include your share of the cost of imposing this fine in your answer.
What will be your share of the cost of imposing that fine? _____ points

3. Suppose that the scheme of individual reduction decisions is in place.

How much does it cost you to reduce the earnings of another group member by 6 points?
_____ points

Note: The instructions for the I-Voting-ST treatment are omitted to conserve space, because the instructions are identical to those for the I-Voting-M treatment (except for the differences in the punishment strength and numerical examples).

## [Team Treatments:]

Teams, as a decision-making unit, make decisions through communication. While some studies set the duration of each communication stage at much more than 60 seconds, prior papers such as Kagel (2018) and Kamei (2019b) set the duration to 60 seconds or less.

Part 1 is the same for the three team treatments (T-No, T-Voting-M and T-Voting-ST treatments). At the beginning of the experiment, the following instructions were read aloud to subjects. The subjects were also given printed copies of the instructions:

## Instructions

You are now taking part in a decision-making experiment. Depending on your decisions and the decisions of other participants, you will be able to earn money in addition to the £3 guaranteed for your participation. Please read the following instructions carefully.

During the experiment you are not allowed to communicate with other participants. If you have questions, raise your hand. One of us will come to answer your question.

During the experiment your earnings will be calculated in points. At the end of the experiment your points will be converted to U.K. pounds at the following rate:

**62.5 points = £1**

(or each point will be exchanged for 1.6 pence of real money). At the end of the experiment your total earnings (including the **£3** participation fee) will be paid out to you in cash. Your payment will be rounded to the nearest 10 pence (e.g., £15.30 if it is £15.33; and £15.40 if it is £15.37).

At the beginning of the experiment, you are randomly assigned to a team with two other participants. The team is the decision-making unit in the experiment. **The team composition stays the same throughout the entire experiment**. Your team is then randomly assigned to a group with two other teams, and interact with each other. This means that you are in a group with 8 other participants (two in the same team, and six in the other teams). **You will be part of the same group throughout the entire experiment**. No one knows which other teams are in their group, and no one will be informed which other teams were in which group after the experiment.

This experiment consists of two parts. Part 1 consists of 4 periods, while Part 2 has five phases each consisting of 4 periods (in total, 20 periods for Part 2). Thus there are six phases (a total of 24 periods) in the experiment. In each period your team will be required to make at least one joint decision. The time allocated for each decision is displayed in the top right corner of the screen in seconds. Please make your decision and click the submit/continue button before the timer reaches 0.

We will first explain the detail of Part 1. We will distribute the instructions for Part 2 once Part 1 is over.

## PART 1

*Your team's decision in each period:*

In each period, each team will be given an endowment of 20 points and will make an allocation decision based on the endowment. The other two teams in your group are also each given **an endowment of 20 points** and simultaneously make allocation decisions.

There are two possibilities:

1. You, as a team, can allocate points to a **group account**.
2. You, as a team, can allocate points to a **private account**.

Specifically, each team will be asked how many points they want to allocate to the group account. The remaining points (that is, 20 minus the allocation to the group account) will be automatically allocated to the team's private account. Your earnings in a given period depend on **(a) the number of points in your team's private account and (b) the total amount allocated to the group account in your group**.

*How to calculate your earnings:*

Your team's earnings in a given period are calculated as in the following formula:

**(sum of points in your team's private account) + 0.6 × (sum of points allocated by your team and the other two teams to the group account)**

In other words, your team's earnings from the private account **are equal to the number of points your team allocated to the private account** (20 minus your team's allocation to the group account). The points your team allocates to the private account do not affect the earnings of the other two teams in your group.

By contrast, your team's earnings from the group account equal the **sum** of points allocated to the group account by your team and the other two teams in your group members multiplied **by 0.6**. In other words, if your team allocates 1 point to the group account, your team's earnings from the allocation is $0.6 \times 1 = 0.6$ points, which is less than 1 point. However, by allocating 1 point to the group account, the earnings of each of the other two teams also increase by 0.6 points. Therefore, the total earnings in this case are 1.8 points, which is greater than 1 point. Note that your team also obtains earnings of 0.6 points for each point the other teams allocate to the group account.

Once three teams in your group make allocation decisions, you will be informed of the interaction outcomes (your team's earnings, along with each of the other two teams' allocation decisions anonymously and in a random order).

**You and two members in your team <u>each obtain the same earnings that your team obtained in each period (e.g. your earnings will be 25 if your team's earnings are 25)</u>.**

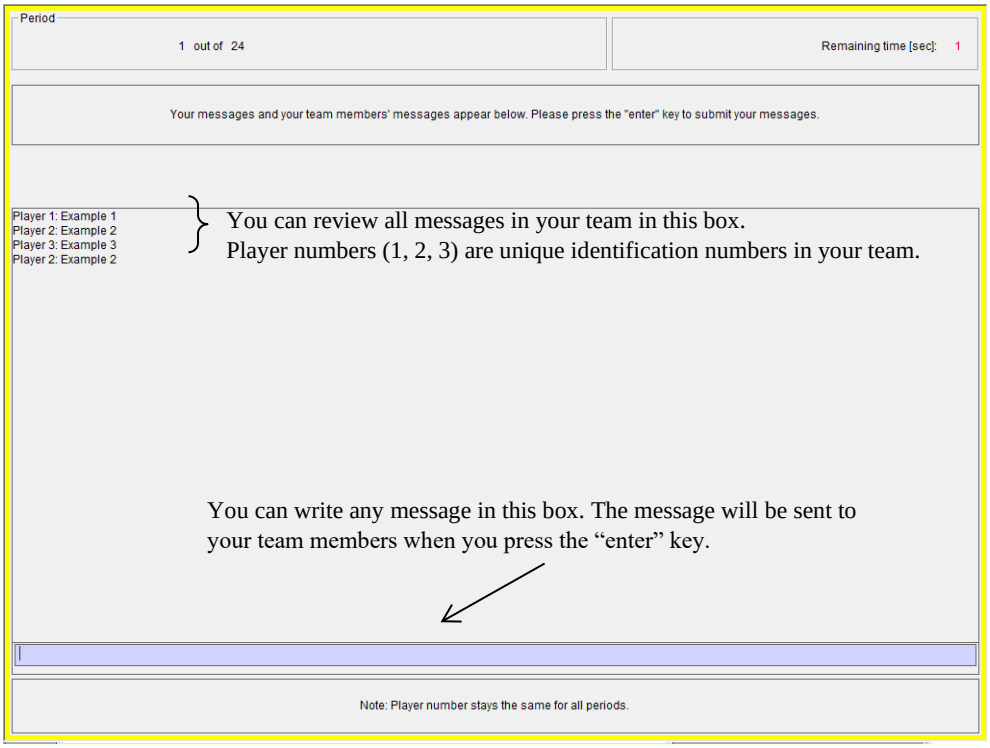
If you have any questions, please raise your hand. When all questions are answered, we will move on to comprehension questions.

*How to decide allocation amounts in your team:*

At the beginning of each period, you and your two team members have <u>1 minute to communicate</u> using the computer to jointly decide the allocation amount for the period. Specifically, you can send any messages via a chat window as illustrated below. In this stage as well, you are not allowed to

verbally communicate with anyone during the entire experiment except via the computer screen with the two members.

An example of the computer screen:




In the communication stage, any kind of offensive language is prohibited. Also, you are not allowed to convey any personal information nor information that can identify you including which seat you are sitting. With a clear violation of this rule you will be deducted 10 pounds from your today's payment.

Once the communication stage is over, you and the other two members in your team each submit your agreed joint allocation decision on your computer screen. In case that you do not agree what you allocate as a team, you can submit whatever amount you prefer to allocate as a team to the group account.

If all three members in your team submit the same (agreed) amount, then the amount becomes your team's joint allocation decision in this period. Otherwise, the median of the three submitted amounts will be used as your team's joint allocation decision. Once three team members press the "Submit" button to submit your team's allocation decision, you will be informed of what allocation amount the other two members in your team submitted before you are informed of the outcome of the allocation stage in the period.

If you have any questions, please raise your hand. When all questions are answered, we will move on to comprehension questions.

*Comprehension questions*

Please answer the following questions. Raise your hand if you need help.

**1.** Suppose that all three teams in your group allocate 0 points to the group account.

(a) How much does your team earn? _____

(b) How much do you earn? _____

(c) How much does another team in your group earn? _____

(d) How much does each member in the other team earn? _____

**2.** Suppose that all three teams in your group allocate 20 points to the group account.

(a) How much does your team earn? _____

(b) How much do you earn? _____

(c) How much does another team in your group earn? _____

(d) How much does each member in the other team earn? _____

**3.** Suppose that the other two teams in your group in total allocate 15 points to the group. Answer the following:

a) How much does your team earn if your team allocates 0 points to the group account? _____ In this case, how much do you earn? _____

b) How much does your team earn if your team allocates 10 points to the group account? _____ In this case, how much do you earn? _____

c) How much does your team earn if your team allocates 20 points to the group account? _____ In this case, how much do you earn? _____

Any questions? When all questions are answered, we will move on to Part 1.

[Once everyone finished answering the comprehension questions and the experimenter explained the answers, the experiment began.]

As soon as Part 1 was over, subjects moved on to Part 2. Part 2 differ by treatment.

**T-No treatment:** At the onset of Part 2, the following instructions were distributed and were read aloud (subjects had also printed copies of the instructions):

## Instructions for Part 2

As explained, you have **five** phases each consisting of 4 periods (in total, 20 periods) in Part 2.

The five phases are each separated by a break of 40 seconds. The structure of each phase is identical to that of Part 1 (1 point = 1.6 pence).

The team composition in Part 2 is the same as Part 1. The group composition in Part 2 is the same as Part 1.

If you have any questions, please raise your hand. When all questions are answered, we will start Part 2.

**T-Voting-M treatment:** At the onset of Part 2, the following instructions were distributed and were read aloud (subjects had also printed copies of the instructions):

## Instructions for Part 2

As explained, you have **five** phases each consisting of 4 periods (in total, 20 periods) in Part 2.

You will continue to be a part of the same team assigned in Part 1, and will interact with the same two teams in the group.

In this part as well, **teams are the decision-making unit** (i.e., all decisions are made jointly with two other members in a team). In making any team decision in Part 2, you will be given an opportunity to communicate with your two team members for 1 minute (using the electronic chat window). After that, you and the other two members in your team each submit your agreed decision on the computer screen. In case that you do not agree what to do as a team, you can submit whatever decision you prefer to make as a team. If all three team members submit the same (agreed) decision, then it becomes your team's decision. Otherwise, the median of the three submissions will be used as your team's joint decision.

In each period, each team will jointly make a decision about allocating 20 points to either a private account or a group account, with the same immediate payment consequence (see the instructions for Part 1). The conversion rate is the same: 1 point = 1.6 pence.

However, there is a significant difference from Part 1 in that each period consists of two stages. The first stage is the allocation stage we just explained. However, each team's earnings from the allocation stage can be reduced in the second stage. There are two possible schemes governing the second stage. **At the beginning of each phase, your group will determine by majority vote which of the two schemes will be used during the four periods of that phase.** You can select different schemes in different phases.

Of the two possible schemes, one is a scheme in which the group votes on the rules of a fine (which we call "**Group-determined fines**"); the other is a scheme in which teams can reduce the other teams' earnings after learning of their allocations (which we call "**Team reduction decisions**").

### Scheme 1: Group-determined fines

In this scheme, earnings from the allocation stage can be reduced by a fine rule. When a rule is in place, allocations to the **private** account are subject to a fine.

At the beginning of each period, your group chooses a fine rate (the amount of the fine per point allocated to the private account) by voting. Possible fine rates are 0, 0.2, 0.4, 0.6, 0.8, 1.0 and 1.2 points per point allocated to the private account.

For each point that is lost by a team who is fined, the group also incurs a cost of 0.6 points to impose that fine. This cost is interpreted as an administrative cost in imposing a fine. For example, if a team is fined a total of 5 points, this costs the group 3 (=5×0.6) points, with each team (including the team that receives the fine) being equally charged 1 (=3/3) point as a share of that cost. More generally, for each 1 point of fines imposed on any team, each team in the group pays 0.6*(1/3) = 0.2 points as their per capita cost of imposing the fine. In the example of a team fined 5 points, that team thus loses both the 5 points and their per capita share of the cost, 1 point, for a total loss of 6 (= 5 + 1) points. Notice that since the team fined loses a total of 6 points while the other teams pay 2 points in the aggregate (i.e., 2×1) in imposing the fine, **the ultimate cost ratio is 1:3 (= 2:6)**.

*Fixed charge:* In addition to the fines and costs based on the fine rule your group chooses, at the end of a period, a fixed cost of 4 points is also deducted from the earnings of each team in your group. This can be thought of as the fixed administrative cost of having a fine scheme in operation, a cost that doesn't depend on how frequently or infrequently fines are in fact imposed.

Fines in the present phase cannot bring a team's earnings for a period to less than zero. However, you always incur the per capita share of the cost of imposing fines and the fixed charge, even if it brings your team's earnings for the period to less than zero.

This means that your team's earnings for a period can be calculated as follows:

Part A: Earnings from the allocation stage minus your team's fine, or 0 if it is negative

**minus**

Part B: Your part of the cost of administering the fine scheme
= the per capita share of imposing fine {= 0.2×total fines imposed} + 4

As mentioned, a team incurs the cost of Part B even if it causes their net earnings for the period to be negative.

*Voting on a fine rate*: Under the group-determined fines, each period starts with teams' voting on a fine rate. The fine rate in a given period will be determined based on the median of the three votes submitted in your group. For example, if three teams enter choices of 0.6, 0, and 0.2 as their preferred rate, then the fine rate will be 0.2 in your group. For another example, if three teams enter choices of 0.4, 0, and 0.4 as their preferred rate, then the fine rate will be 0.4. Note that there is effectively no fine if your group chooses a fine rate of 0. Also, if the fine rate is positive, earnings at the end of a period may be unchanged from those at the end of the allocation stage if no member allocates points to the private account.

Summary: Each period under the group-determined fines consists of the following steps:

Step 1. You will communicate with your two team members to jointly decide which fine rate to vote.
Step 2. You and the two members simultaneously submit your team's decision (fine rate). After that, each of you will be informed of (a) your team's decision and (b) what the other two members submitted as a fine rate.
Step 3. Your team and the other two teams will be informed of the vote outcome in your group. The **median** of three votes cast in your group is implemented in the period.
Step 4. You will communicate with the two members in your team regarding your team's joint allocation decision under the determined fine rate.
Step 5. You and the two members simultaneously submit your team's decision (allocation decision). After that, each of you will be informed of (a) your team's joint allocation decision and (b) what the other two members submitted.
Step 6. You will be informed of the outcome of the allocation decision (e.g., earnings).

**You and two members in your team each obtain the same earnings that your team obtained in each period.**

## Scheme 2: Team reduction decisions

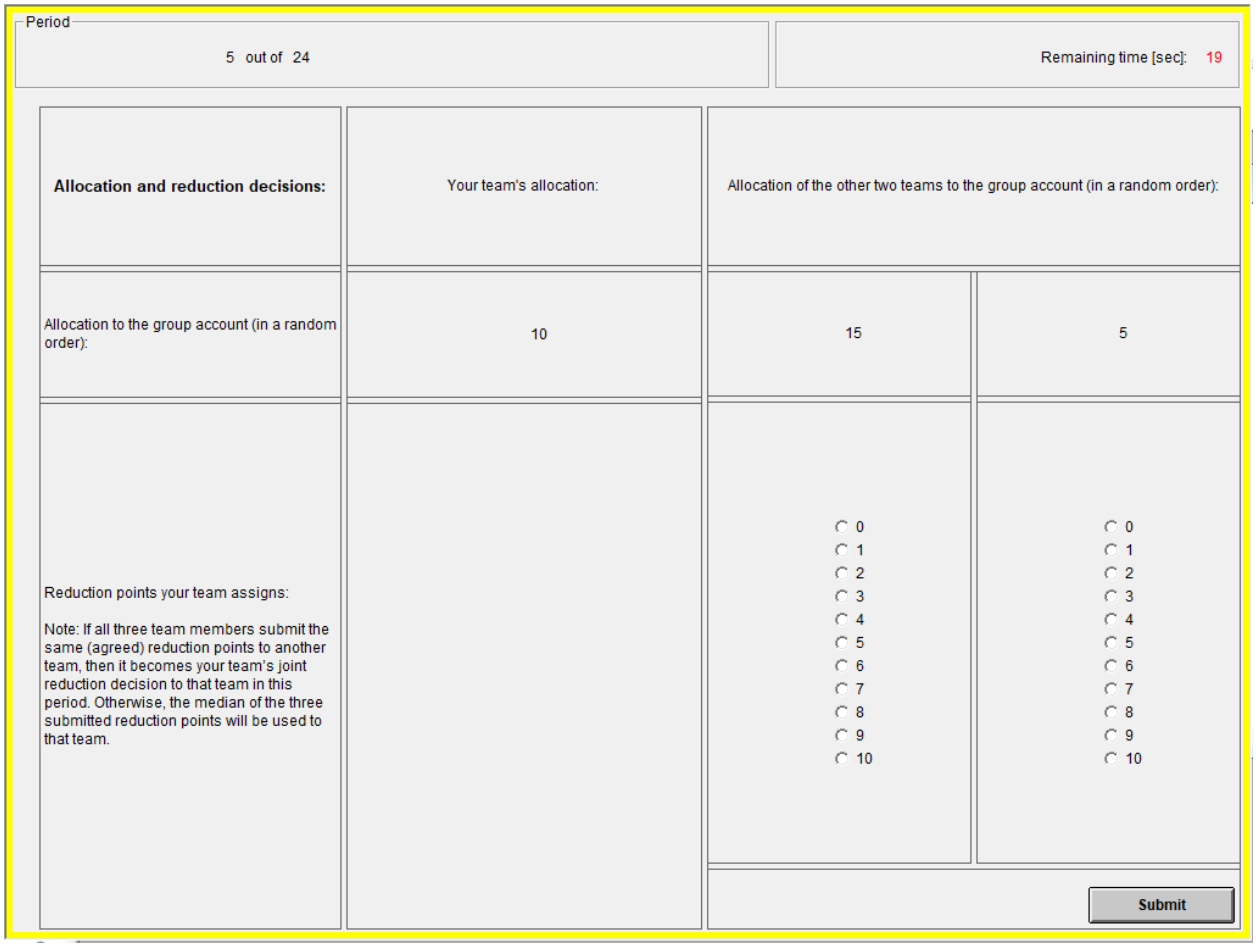
In this scheme, each period starts with your team's joint allocation decision as in Part 1.

However, unlike Part 1, there is a post-allocation stage in which you as a team have an opportunity to reduce the earnings of the other teams in your group at a cost to your earnings.

Specifically, your team can assign reduction points to each of the other two teams. Each reduction point you allocate to reducing another team's earnings reduces your earnings by 1 point and reduces the other team's earnings by 3 points. Thus, the **cost ratio is 1:3 (your team: the other team)** as in the group-determined fines explained above. The earnings of your team can also be reduced in the same way by the decisions of the other teams in your group. You are free to leave any or all other teams' earnings unchanged by assigning 0 reduction points.

The way to jointly decide reduction points to another team is the same as other team decision-making. You will be first given an opportunity to communicate with your two team members for 1 minute using an electronic chat window. Before the communication stage, you will be informed of the allocation decisions made by the two other teams in your group.

After the communication, you and the two members simultaneously submit your team's agreed reduction decisions. After that, each of you will be informed of (a) your team's joint reduction decision and (b) what the other two members submitted.




*Note*: Numbers are for illustration only

Reductions directed at your team cannot bring your earnings for the period to less than zero. However, you always incur the cost of giving reductions to other teams even if it makes your team's period earnings negative. (If your team loses points in a period, they are deducted from those they accumulate in other periods.) Thus, earnings in each period of this phase can be calculated as follows:

Part A: Earnings from the allocation stage minus reductions by other teams in your group, or 0 if it is negative

**minus**

Part B: Points your team's use to reduce other teams' earnings

Note that you incur the cost in Part B even if it causes your team's net earnings for the period to be negative.

Example: Suppose that your team uses 0 points to reduce the earnings of the first team whose allocation appears on the screen, and uses 1 point to reduce the earnings of the second. Suppose further that these two teams use 1 point and 3 points to reduce your team's earnings. Then the second team's earnings for the period will be reduced by 3 (=1×3) points in addition to any reductions due to the decisions of the other team. Your team's own earnings for the period will be reduced by 1 point [i.e., your cost to impose reductions on others], plus (1×3)+(3×3)=12 points [i.e., the reductions imposed on your team by others]. At the end of the reduction stage, your team will learn that other

teams decided to reduce your team's earnings by a total of 12 points, but your team will not be told which teams reduced your earnings or by how much any given team reduced your earnings. Others will also not know which team in particular reduced their earnings, or by how much.

In addition to the fact that earnings from the allocation stage and reductions received cannot go below zero, the earnings reduction process is subject to two requirements. First, the reduction points must be an integer. Second, you cannot assign more than 10 reduction points to any team in your group.

Remember that if no reductions are imposed, earnings after the reduction stage are the same as those before it.

**As in the other scheme, you and two members in your team each obtain the same earnings that your team obtained in each period.**

## Summary of Part 2 (phases 2 to 6):

At the beginning of the first period in every 4 period phase, each team will jointly vote on two schemes: "Group-determined fines" versus "Team reduction decisions."

Whichever scheme gets the most votes (≥ 2 votes) will be in effect for four periods.

(i) When Group-determined fines is chosen:

In each period, your team will jointly make a voting decision on the fine rate after one minute of communication.

The median of the three teams' votes in your group is used for the fine rate. Under the chosen fine rate, your team and the other two teams in your group simultaneously decide joint allocations amounts to the group account.

(ii) When Team reduction decisions is chosen:

Each period consists of an allocation stage and a reduction stage. You as a team will first jointly decide how to allocate points between your private and group account. After that, your team will make a decision about whether to reduce the earnings of other teams or not, and by what amount your team reduces them if so.

You will have **5 voting decisions** in total regarding the scheme to be used by your group—once for each of phases 2 – 6.

*Comprehension questions:*

Please answer the following questions. Raise your hand if you need help.

1. About voting between the two schemes:

a) How many periods do you have in Part 2 of the experiment? _____

b) How many times do you have the opportunity to vote on which scheme is used?

_____
__

c) If your group selects the scheme of group-determined fines in Phase 3 (periods 9 – 12), can it select a different scheme in Phase 4 (periods 13 – 16)? _____

2. Suppose that the scheme of group-determined fines is in place in a given phase.

a) What is the fixed charge each period for operating the fine scheme? _____

b) Suppose that votes on fine rate cast by the three teams in your group are: 0.2, 0.6, 0. What is the fine per point in your group? _____

c) Suppose that your group selected a fine rate of 0.4, and suppose that your team allocates 15 points to the group account. How many points will your team lose in the form of a fine*? _____ points
* Note: do not include your team's share of the cost of imposing this fine in your answer.
What will be each team's share of the cost of imposing that fine? _____ points

3. Suppose that the scheme of team reduction decisions is in place.

How much does it cost your team to reduce the earnings of another team by 6 points?
_____ points

Note: The instructions for the T-Voting-ST treatment are omitted to conserve space, because the instructions are identical to those for the T-Voting-M treatment (except for the differences in the punishment strength and numerical examples).

# Appendix B: Appendix for Chapter 3

## Appendix B.A: Experiment Procedure and Instructions Used in the Experiment

Sixty-two sessions (thirty-one sessions per treatment) were conducted online using the oTree software (Chen *et al.*, 2016) and Zoom from May 2021 through January 2022, following the same procedure as a standard laboratory experiment. All subjects' cameras were on during the session to make sure that they were alone, were paying attention to the experiment, and did not cheat when making decisions. While subjects were visible by the experimenter, they were unable to see the researcher on Zoom during the experiment. They also remained anonymous during the entire session without seeing other participants' faces, names, etc. Each session consisted of nine subjects. This means that each session consists of one group. However, full anonymity was retained in the experiment since, as already mentioned, subjects did not see the other students' names or faces, and they were recruited from a very large student population in the university. All experiment sessions were conducted using the subject pool and the experiment system in the EXEC (Centre for Experimental Economics) at the University of York. As all standard experiment protocols (such as the no deception rule) have been rigorously adopted for any experiment in this laboratory for more than 30 years, it can be assumed that subjects believed in the explanation provided in the experiment, although subjects do not see the presence of any subject physically.

All subjects were recruited using *hroot* (Bock *et al.*, 2014). Subjects did not participate in more than one session. A total of 552 students in the University of York participated in the experiment.[67] The instructions shown below were neutrally framed. Terms with positive or negative connotations, such as shirk, free ride or cooperate, were not used. Each session took between 90-120 minutes. This part of the Appendix includes the instructions for both the ENDO and EXO treatments as follows:

A.1: Instructions for phase 1 (identical for the ENDO and EXO treatments)

A.2: Instructions for phase 2 (the EXO treatment)

A.3: Instructions for phase 2 (the ENDO treatment)

**References:**

Bock, O., Baetge, I., & Nicklisch, A. (2014) hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, *71*, 117-120.

Chen, D., Schonger, M., & Wickens, C. (2016) oTree - An open-source platform for laboratory, online and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88-97.

[67] Two teams' observations (six subjects) were omitted from the study. The first one was due to a subject's computer experiencing technical problems, meaning that their whole team could not participate in the task-solving phase, and the second was due to suspected cheating which was identified in the chat dialogue.

*A.1: Instructions: Slides for the Practice Phase (identical for the ENDO and EXO Treatments)*

At the onset of the experiment, the following instructions (PowerPoint file) were shown on Zoom and were made available on the subjects' experiment screens; they were read aloud to subjects by the researcher.

Slide 1:

Welcome

You are now taking part in a decision-making experiment. In the experiment, you will be able to earn money depending on your decisions and the decisions of other participants in some tasks. Please read the following instructions carefully.

During the experiment you are not allowed to communicate with other participants. If you have any questions, please write your questions in the Zoom chat box. We will answer your question in private. Your microphones will be automatically muted by us on Zoom. The camera of your computer should be turned off until you are put into your individual breakout rooms. In your breakout rooms, your camera should be turned on to ensure that you are still participating, but only the experimenter will be able to see you and you will not be recorded.

Slide 2:

There are **nine** participants in this experiment. You will not be made aware of any information about other participants' identity, such as their faces and names. At the onset of the experiment, each participant will be randomly assigned to **a team of three**. This means that there are three teams in the experiment. Team assignment is completely random. The three members in your team will then be randomly assigned ID numbers, 1, 2, or 3, so that each member receives a different number from the others. The player who is assigned number k in your team will become "player k."

This experiment consists of two phases. The team composition **stays the same for the two phases**. We will first explain the detail of Phase 1. We will explain Phase 2 once Phase 1 is over.

Slide 3:

**Phase 1:**

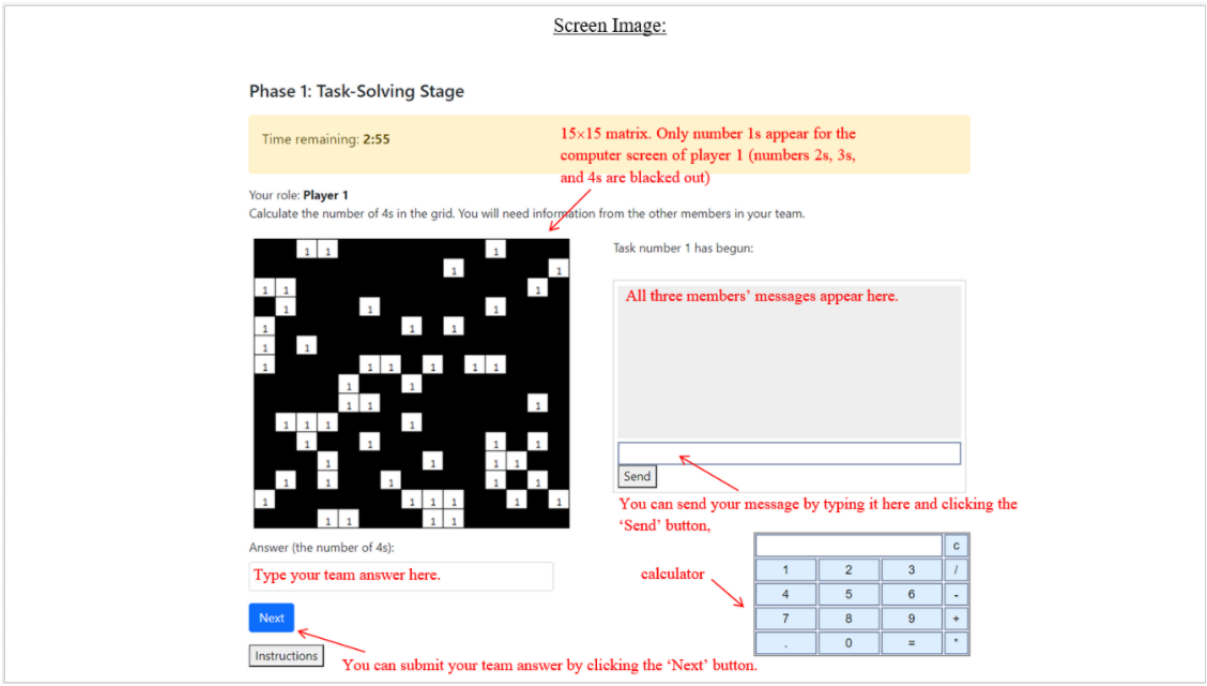
You and the other two members of your team will jointly solve **number counting tasks** for **3 minutes**. Your payment depends on the work performance of your team, not the performance of the other two teams. We will explain the nature of the number counting task, and how you are paid in this phase, in order.

**1. Number Counting Task:**

Your team will be asked to jointly solve as many number counting questions as possible. Each question has a table that consists of **225 randomly ordered 1s, 2s, 3s and 4s** (i.e., 15X15 matrix). Player k in your team will be shown a table on which only number ks appear while the other three numbers are blacked out, and they **will then count the number of ks**. For example, player 1 will count the number of 1s as in the following (next slide) computer screen image. You can see that on player 1's computer screen, the numbers of 2s, 3s, and 4s are blacked out.

Slide 4:

Screen Image:



Slide 5:

In order to submit a team answer, your team must jointly find **the number of 4s** included in the table. For example, if the numbers of 1s, 2s, and 3s are 43, 62, and 58, respectively, in a given table, you can find the number of 4s by the following calculation:

$$225 - 43 - 62 - 58 = 62.$$

Operationally, the three members in your team need to communicate with each other using an electronic chat window on the right side of the screen. You need to exchange information with each other regarding the numbers found on each of your own screens, although messages in the chat window are unrestricted. For example, player 1 can share the number of 1s with the two other members. The three members then jointly calculate the number of 4s through communication. A calculator is available also on each screen. In the communication stage, any kind of offensive language is prohibited and you should not share any information that compromises your anonymity.

Slide 6:

How to submit an answer: There is a blank space below the label "Answer" on each member's screen. All three members must fill the blank on their screens with the team's joint answer (the number of 4s) and click the "Submit" button. **Only when all three members submit the correct answer, will the team's answer count as correct**. The submit button **does not** need to be pressed at exactly the same time, but a new matrix will not be generated until all three members have submitted an answer for the current matrix. Hence, you need to communicate to submit the joint answer. Once an individual member has submitted an answer, they will move to a waiting room screen until both of the other two team members submit an answer. You will still be able to communicate with your team members while in the waiting room. You will not be given feedback regarding whether your team's answer is correct or not for each question. You will instead be informed of the total number of questions correctly answered at the end of the three-minute phase. Note that in the case of disagreement, each member can submit a different answer from the others. However, your team's answer is then counted as incorrect.

Once your team submits an answer, you will move on to the next number counting question and will then be shown **a new table with 1s, 2s, 3s and 4s randomly generated**. You can then solve the new question as before.

Slide 7:

**2. How to earn money in Phase 1:**

Payment is based on individual team performance. **Your team earns 180 pence for each correct team answer**. The payment is equally divided among the three members in your team, meaning that you and the other two members of your team each **earn 60 pence** for each question correctly answered.

Are there any questions? If you have any questions, please write your questions in the Zoom chat box now.

*A.2: Instructions: Slides for the Main Task-Solving Phase in the EXO Treatment*

Once Phase 1 was over, the following instructions were shown on Zoom and were made available on the subjects' experiment screens; they were read aloud to subjects by the researcher.

Slide 1:

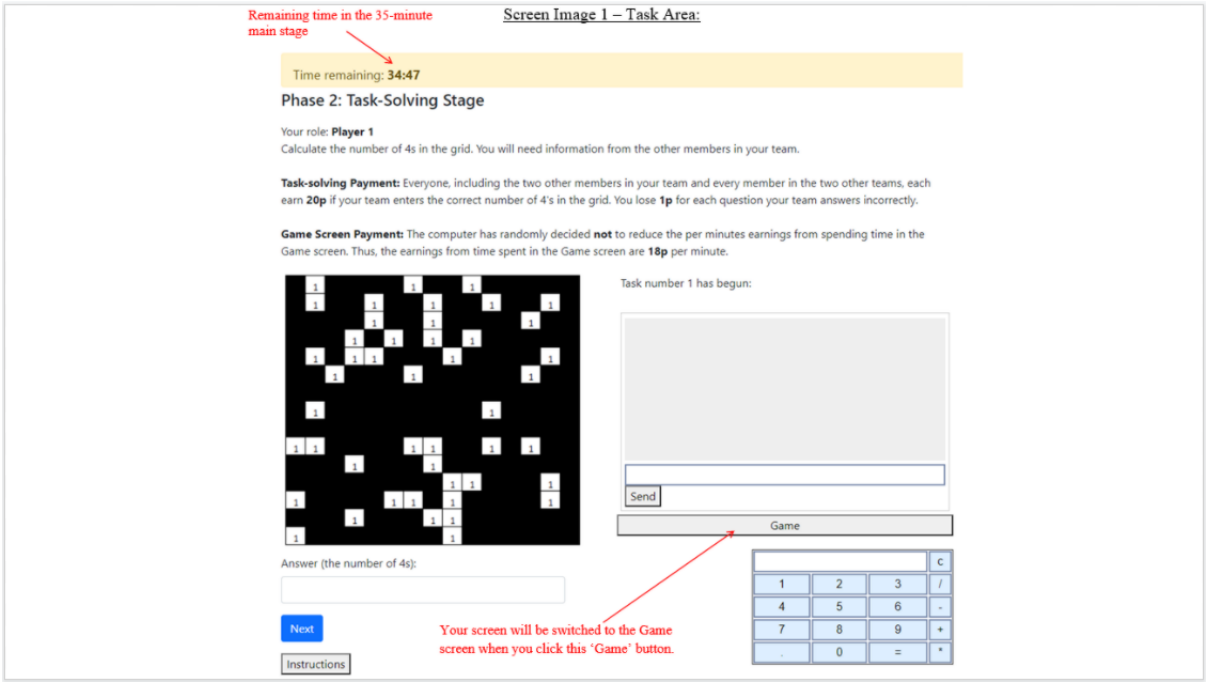
**Instructions for Phase 2**

As explained, you continue to be a member of the same team with two other participants in Phase 2.

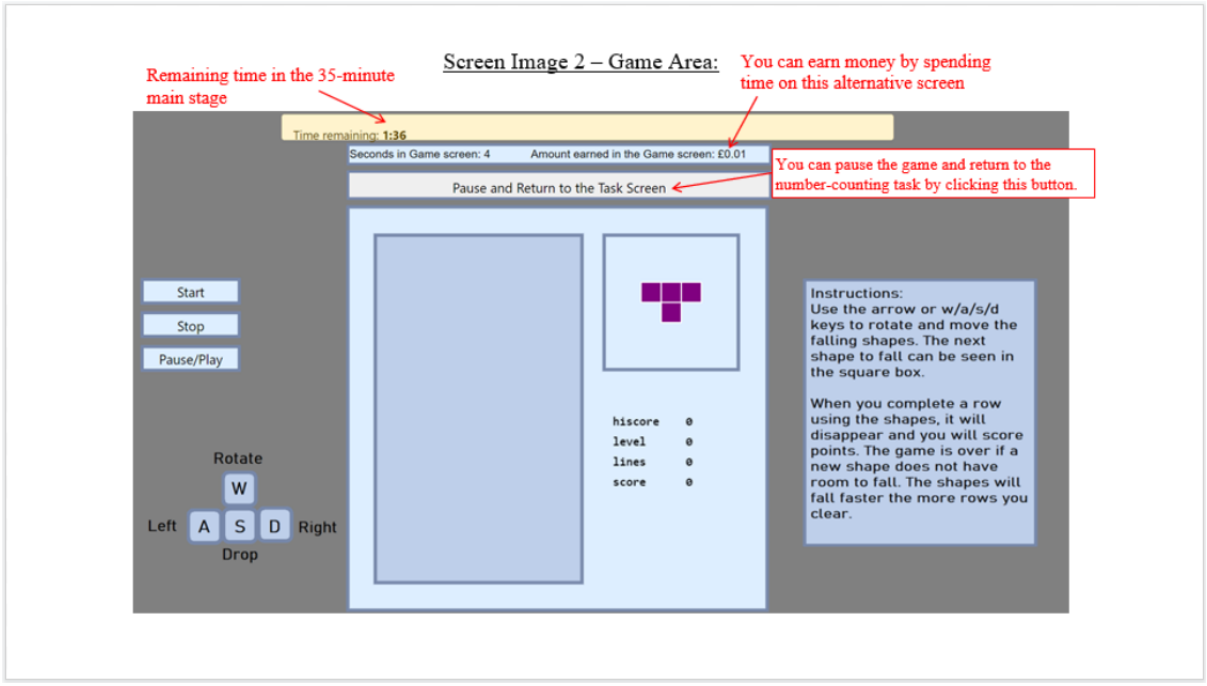
In Phase 2, you and the other two members of your team will jointly solve **number counting tasks** for **35 minutes**. Your payment depends on not only the work performance of your team **but also that of the other two teams**. The number counting task is exactly the same as the one you previously solved in Phase 1 except that an alternative activity is available.

An alternative activity – Tetris: In Phase 2, you can play Tetris, for example, to take a rest, whenever you like (see Screen Image 1 below). There is a **"Game"** button on the computer screen. You can choose to play Tetris at any point during the 35-minute task period by clicking the "Game" button. If clicked, your screen will change to one with a game of Tetris available (see Screen Image 2 below).

Slide 2:




Slide 3:



Slide 4:

In Tetris, you can use the arrow or w/a/s/d keys to rotate and move the falling shapes, with the aim of completing as many rows as possible. You can start or stop Tetris by clicking the 'Start' or 'Stop' button on the left side of the screen. You can also pause Tetris by clicking the 'Pause/Play' button while remaining on the Game screen (you can resume Tetris by clicking 'Pause/Play' button again).

Note that you cannot play Tetris while working on a counting task. You can return any time to the counting task screen and work on the same number counting question by clicking the "Pause and Return to the Task Screen" button on the Game screen, i.e., by pausing Tetris. You can also return to the Game screen and resume the game any time. No one, including the other two team members in your team, is informed about your decision to play Tetris unless you volunteer this information, for example in the team chat.

Slide 5:

**How to earn money in Phase 2:**

There are two sources for you to earn money:

(1) Earnings from the number counting task – you will be paid based on all of the participants' performances in this phase. After 35 minutes, the total number of correctly answered questions by each team are summed. Each correct answer is exchanged for **180 pence** as in Phase 1, which is equally shared among all the today's participants. There are three teams in the experiment, and each team receives **one-third of the total earnings from working on the number counting task**. This means that for each correct response, your team earns 180/3 = 60 pence, and each of the other two teams also earn 60 pence. Your team equally shares the team earnings among three members: you and the two other members in your team **will each receive 20 pence**. Note that you will also earn 20 pence for each question correctly answered by another team.

Unlike Phase 1, **there is also a penalty of 3 pence per incorrect answer** to your team; this penalty is only applied to your team's incorrect answers, and is shared equally between the three team members. This means that **you have to pay 1 penny per your team's incorrect answer**. You will only be told the total amount of correct and incorrect answers by your team, and the group, at the end of the 35-minute task-solving stage.

Slide 6:

(2) Earnings from staying in the Game screen – You will earn an amount of money also by switching to the Game screen instead of working on counting. Specifically, you will receive **18 pence** per minute spent in the Game screen. On this screen you can play Tetris. Note that the earnings from staying the Game screen will be added to your own earnings. They will not be shared with the other members of your team or with members of other teams, unlike working on the number counting task. Note that the score you get in Tetris does not affect your earnings; only the time you spend in the Game screen determines the size of payout here.

Slide 7:

**Remark:** Notice that if everyone stays in the Game screen for the entire 35-minute stage, each of you can earn 18 pence X 35 = 6.3 pounds in Phase 2. However, for example, if all three teams each answer two number counting questions correctly per 5 minutes, in this phase the group can correctly answer a total of 42 questions = 2 [correct answer per 5 minutes] X 7 [as the duration is 35 minutes] X 3 [teams]. This means that each participant can earn 20 pence X 42 = 8.4 pounds, which is larger than 6.3 pounds, in addition to whatever they earn from staying in the Game screen. If the three teams correctly answer more questions, it is possible for everyone to earn more than in this example. By contrast, if only your team works on the number counting task while other teams often switch to the Game screen, your team will earn less than the other teams, as revenue-sharing is used in this phase for the counting task only, not for spending time in the Game screen.

Are there any questions? If you have any questions, please write them in the Zoom chat box. Once all questions have been answered, we will explain the structure of Phase 2.

Slide 8:

**Structure of Phase 2**

Phase 2 proceeds with two steps:

Step 1: The computer's decision to randomly reduce the incentive to switch to the Game screen.

Before moving on to the main task-solving stage of 35 minutes, there is a possibility that the computer randomly enforces the change of payment rule from spending in the Game screen: this rule **reduces payment from 18 pence to 16 pence for each minute you spend in the Game screen**. If this policy change is made, then per minute return from being in the Game screen diminishes for every participant today.

Slide 9:

Step 2: Main Task-solving Stage

You will undertake the number counting task **for 35 minutes**. You and all other participants each get 20 pence for each question your team answers correctly, and lose 1 penny for each question your team answers incorrectly. You can also get 20 pence for each question answered correctly by another team. Instead of counting numbers, you can switch from the work screen to the Game screen, for example to play Tetris, as already discussed. You can earn points by spending time in the Game screen also – the per minute earnings differ dependent on the computer's random choice – either 16 or 18 pence per minute. The earnings from the Game screen will not be shared with anyone else in today's session.

Are there any questions? If you have any questions, please write them in the Zoom chat box

*A.3: Instructions: Slides for the Main Task-Solving Phase in the ENDO Treatment*

Once Phase 1 was over, the following instructions were shown on Zoom and were made available on the subjects' experiment screens; they were read aloud to subjects by the researcher.

Slide 1:

**Instructions for Phase 2**

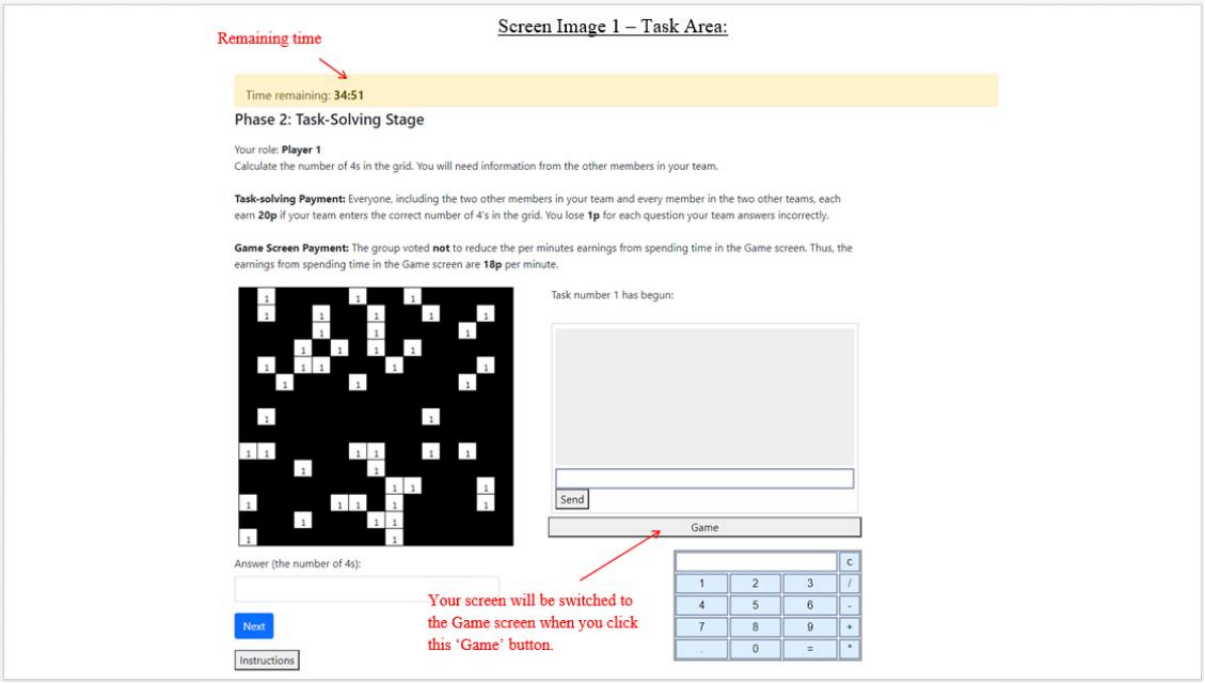
As explained, you continue to be a member of the same team with two other participants in Phase 2.

In Phase 2, you and the other two members of your team will jointly solve **number counting tasks** for **35 minutes**. Your payment depends on not only the work performance of your team **but also that of the other two teams**. The number counting task is exactly the same as the one you previously solved in Phase 1 except that an alternative activity is available.

An alternative activity – Tetris: In Phase 2, you can play Tetris, for example, to take a rest, whenever you like (see Screen Image 1 below). There is a **"Game"** button on the computer screen. You can choose to play Tetris at any point during the 35-minute task period by clicking the "Game" button. If clicked, your screen will change to one with a game of Tetris available (see Screen Image 2 below).

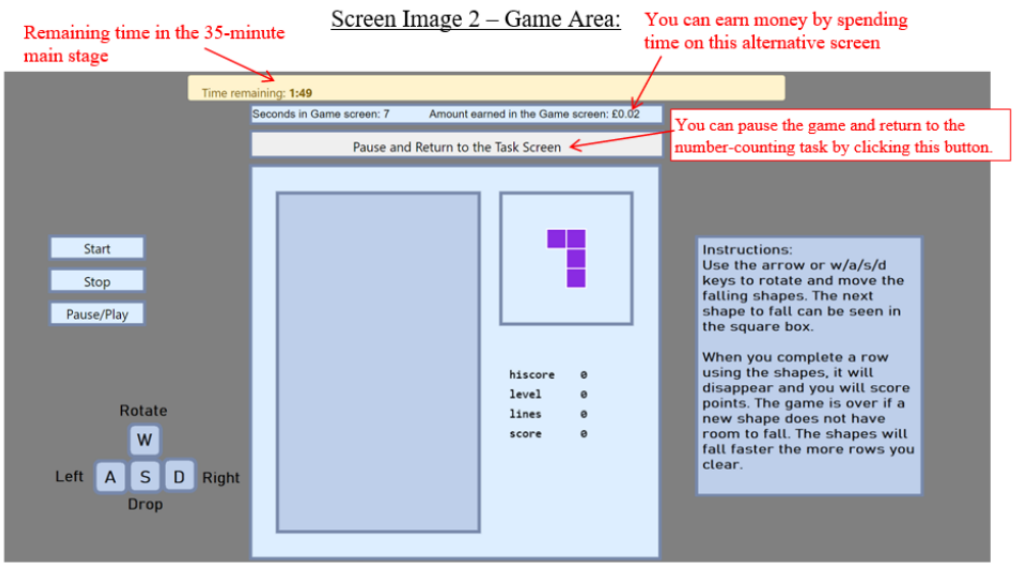
Slide 2:

Screen Image 1 – Task Area:




Slide 3:

Screen Image 2 – Game Area:




Slide 4:

In Tetris, you can use the arrow or w/a/s/d keys to rotate and move the falling shapes, with the aim of completing as many rows as possible. You can start or stop Tetris by clicking the 'Start' or 'Stop' button on the left side of the screen. You can also pause Tetris by clicking the 'Pause/Play' button while remaining on the Game screen (you can resume Tetris by clicking 'Pause/Play' button again).

Note that you cannot play Tetris while working on a counting task. You can return any time to the counting task screen and work on the same number counting question by clicking the "Pause and Return to the Task Screen" button on the Game screen, i.e., by pausing Tetris. You can also return to the Game screen and resume the game any time. No one, including the other two team members in your team, is informed about your decision to play Tetris unless you volunteer this information, for example in the team chat.

Slide 5:

**How to earn money in Phase 2:**

There are two sources for you to earn money:

(1) Earnings from the number counting task – you will be paid based on all of the participants' performances in this phase. After 35 minutes, the total number of correctly answered questions by each team are summed. Each correct answer is exchanged for **180 pence** as in Phase 1, which is equally shared among all the today's participants. There are three teams in the experiment, and each team receives **one-third of the total earnings from working on the number counting task**. This means that for each correct response, your team earns 180/3 = 60 pence, and each of the other two teams also earn 60 pence. Your team equally shares the team earnings among three members: you and the two other members in your team **will each receive 20 pence**. Note that you will also earn 20 pence for each question correctly answered by another team.

Unlike Phase 1, **there is also a penalty of 3 pence per incorrect answer** to your team; this penalty is only applied to your team's incorrect answers, and is shared equally between the three team members. This means that **you have to pay 1 penny per your team's incorrect answer**. You will only be told the total amount of correct and incorrect answers by your team, and the group, at the end of the 35-minute task-solving stage.

Slide 6:

(2) Earnings from staying in the Game screen – You will earn an amount of money also by switching to the Game screen instead of working on counting. Specifically, you will receive **18 pence** per minute spent in the Game screen. On this screen you can play Tetris. Note that the earnings from staying the Game screen will be added to your own earnings. They will not be shared with the other members of your team or with members of other teams, unlike working on the number counting task. Note that the score you get in Tetris does not affect your earnings; only the time you spend in the Game screen determines the size of payout here.

Slide 7:

**Remark:** Notice that if everyone stays in the Game screen for the entire 35-minute stage, each of you can earn 18 pence X 35 = 6.3 pounds in Phase 2. However, for example, if all three teams each answer two number counting questions correctly per 5 minutes, in this phase the group can correctly answer a total of 42 questions = 2 [correct answer per 5 minutes] X 7 [as the duration is 35 minutes] X 3 [teams]. This means that each participant can earn 20 pence X 42 = 8.4 pounds, which is larger than 6.3 pounds, in addition to whatever they earn from staying in the Game screen. If the three teams correctly answer more questions, it is possible for everyone to earn more than in this example. By contrast, if only your team works on the number counting task while other teams often switch to the Game screen, your team will earn less than the other teams, as revenue-sharing is used in this phase for the counting task only, not for spending time in the Game screen.

Are there any questions? If you have any questions, please write them in the Zoom chat box. Once all questions have been answered, we will explain the structure of Phase 2.

Slide 8:

**Structure of Phase 2**

Phase 2 proceeds with two steps:

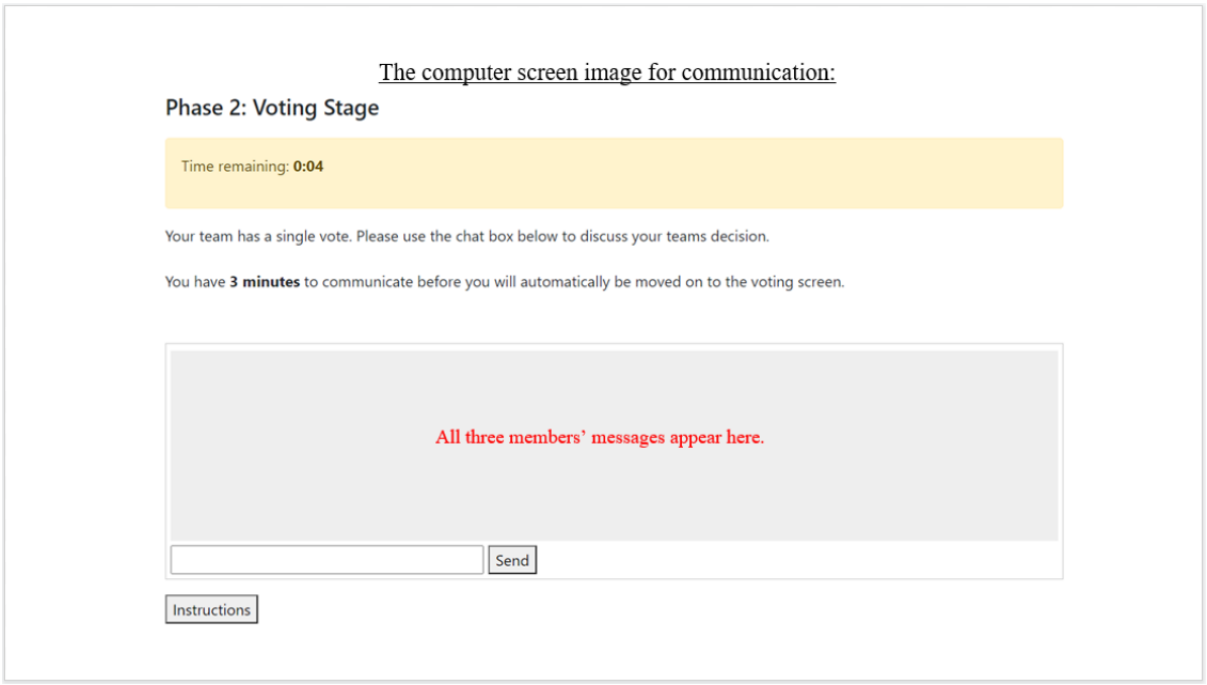
Step 1: Deciding whether to reduce the incentive to switch to the Game screen.

Before moving on to the main task-solving stage of 35 minutes, there is an option to **reduce the per minute earnings from spending time in the Game screen from 18 pence to 16 pence** in today's experiment. Each of the three teams independently vote on whether they wish to implement this reduction option. Each team is given three minutes to communicate using a chat window to decide a team vote. Each team has only one vote.

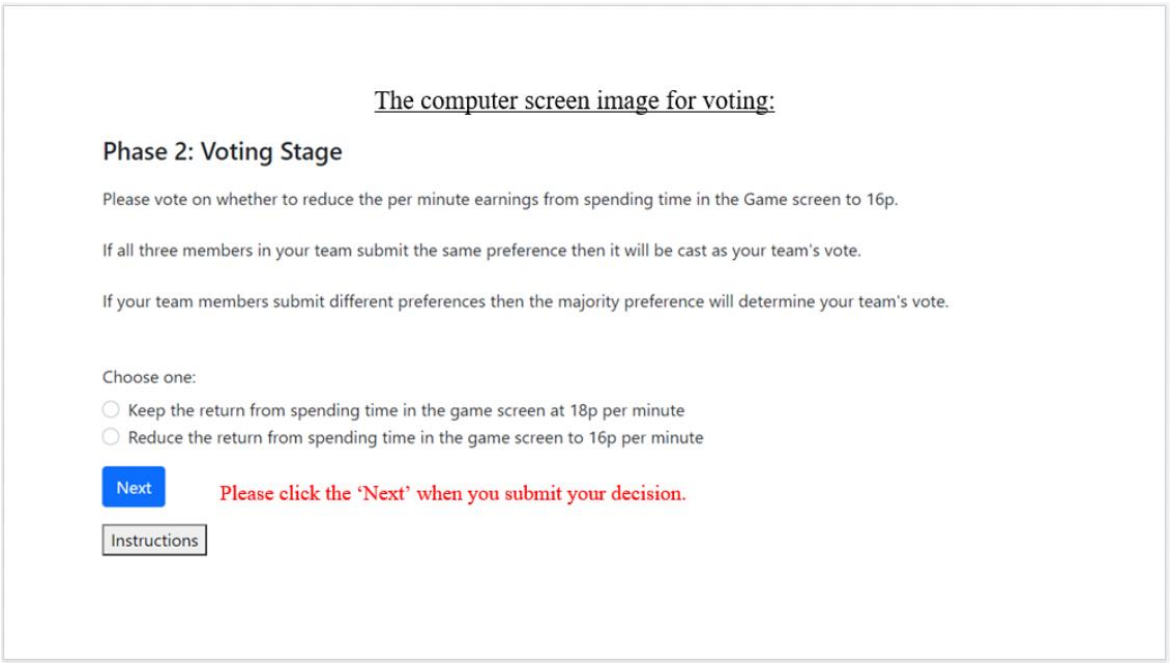
Once the 3 minutes of communication passes, the three members in your team each submit a preferred (agreed) decision. If the three members submit the same voting decision (either implement or not implement), then it becomes your team's joint voting decision. If the three members do not submit the same option, then a majority rule is applied. Whichever receives at least two members' support will be implemented as your team's vote.

After all three teams complete the voting decision, they are informed of the vote outcome. The reduction policy is implemented in today's experiment if **it receives a majority of the votes (2 or 3 votes)**. The reduction policy is not implemented if **it does not receive a majority of the votes**.

Slide 9:




Slide 10:



Slide 11:

Step 2: Main Task-solving Stage

You will undertake the number counting task **for 35 minutes**. You and all other participants each get 20 pence for each question your team answers correctly, and lose 1 penny for each question your team answers incorrectly. You can also get 20 pence for each question answered correctly by another team. Instead of counting numbers, you can switch from the work screen to the Game screen, for example to play Tetris as already discussed. You can earn points by spending time in the Game screen also – the per minute earnings differ dependent on the vote outcome – either 16 or 18 pence per minute. The earnings from the Game screen will not be shared with anyone else in today's sessions.

Are there any questions? If you have any questions, please write them in the Zoom chat box

## Appendix B.B: The Dividend of Democracy in a Theoretical Model

This part of the appendix illustrates how sacrifice helps improve effort provision. It also studies how democracy in decision-making helps improve productivity further. The analysis can be made using a similar framework to the one used in Kamei and Markussen (forthcoming) except changing the variables. In the present paper, a group consists of three teams, and each team consists of three individual members. As the likelihood to answer a collaborative counting task depends on a team's joint effort provision, it is reasonable to assume that the payoff a team receives depends on the team's degree of effort provision $e_i \in [0,1]$. For example, if all three members put their highest effort without any shirking, $e_i = 1$. However, *e* would be considerably smaller if just one member puts in very little effort, as then the number of 4s cannot be answered accurately. On the other hand, *e* may be at an adequate level when all three execute adequate effort with some shirking. Each team's decision can be expressed as below:

$$\max_{e_i \in [0,1]} \{\pi_i(e_i|e_{-i}) = \sum_{n=1}^{3}(s_n + \mu D) \cdot e_n + r_{k\in\{\mathrm{S,NS}\}} \cdot g_i - f(e_i)\}, \text{ where}$$

$$f(e_i) = (c_i - \delta D){e_i}^2 \text{ [cost function]; and}$$

$$e_i + g_i = 1 \text{ [allocation of effort and shirking activities].} \quad \text{(B1)}$$

Here, $g_i$ is team *i*'s average shirking level in the phase 2 task-solving stage. Note that $e_i + g_i = 1$ because there are only two possibilities for simplicity: work or shirk. $s_i$ is the marginal return of effort provision by team *i*, and it is assumed to be a constant (reflecting the number of tasks answered, the likelihood to answer correctly, and the earnings from correct answers and from mistakes), and $r_k$ is marginal return of shirking which depends on their group's sacrifice decisions, i.e., *k* = S (Sacrifice) or NS (Not Sacrifice), and $r_{\mathrm{NS}} > r_{\mathrm{S}}$.

As in Kamei and Markussen (forthcoming), it can be assumed that being involved in democratic decision-making eases workers' effort cost ($\delta > 0$) such that $\delta < c_i$, due to either enhanced intrinsic motivation or signaling. It can also be assumed that democracy boosts team *i*'s work productivity, defined as per effort productivity, from $s_i$ to $s_i + \mu$, where $\mu > 0$. $D$ is an indicator variable which equals 1 in the ENDO treatment.[68]

[68] It can further be assumed that $\delta$ and $\mu$ depend on the policy outcome such that these parameters are larger when the policy is selected than is not selected: $\delta|_{\text{imposed}} > \delta|_{\text{not imposed}} > 0$ and $\mu|_{\text{imposed}} > \mu|_{\text{not imposed}} > 0$. This Appendix provides the theoretical result when these effects of democratic decision-making do not depend on the outcome for simplicity as the theoretical implication is similar regardless of the assumption.

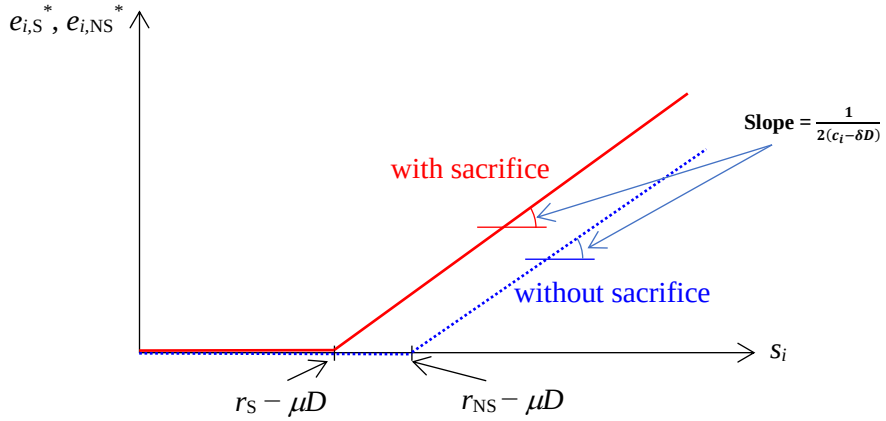
Team $i$'s optimal effort level can be derived merely using the first-order condition for (B1) as the cost function is quadratic. The optimality condition is summarized in Equation (B2) below.

$$e_i^* = \frac{s_i+\mu-r_k}{2(c_i-\delta D)} \text{ if } s_i > r_k - \mu; \text{ and } e_i^* = 0 \text{ otherwise, for } k \in \{\text{S}, \text{NS}\}. \quad \text{(B2)}$$

It is clear from (B2) and the figure on the next page that sacrifice has a positive impact on teams' effort provision since $r_{\text{NS}} > r_{\text{S}}$:

$$e_{i,\text{S}}{}^* = \frac{s_i+\mu-r_{\text{S}}}{2(c_i-\delta D)} \geq e_{i,\text{NS}}{}^* = \frac{s_i+\mu-r_{\text{NS}}}{2(c_i-\delta D)}. \quad \text{(B3)}$$

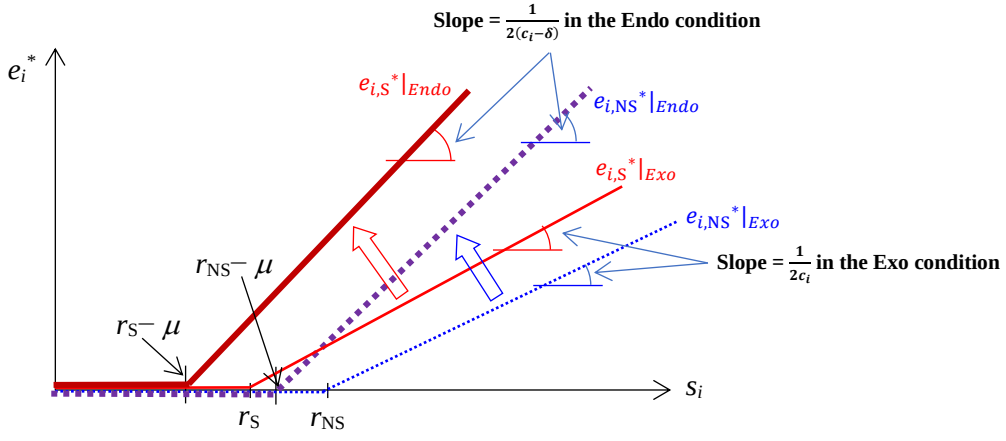
**Figure B.1:** *The Reduction Policy and Worker's Optimal Effort Schedule*




Further, as $\delta > 0$ and $\mu > 0$, teams in the ENDO treatment work harder than those in the EXO treatment – see the figure below:

$$e_{i,k}{}^*|_{Endo} = \frac{s_i+\mu-r_k}{2(c_i-D)} \geq e_{i,k}{}^*|_{Exo} = \frac{s_i-r_k}{2c_i} \text{ for given } k \in \{\text{S,NS}\}. \quad \text{(B4)}$$

**Figure B.2:** *Effects of Democracy on Worker's Optimal Effort Schedule*




These analyses can be summarized as in Summary 1:

**Summary 1**: *(a) Teams work harder with than without the reduction policy. (b) Democracy induces the workers to work harder.*

Does the sacrifice benefit teams? This question may not be obvious because a rise in effort provision means not only an increased return from work but also a rise in the effort cost. To make the further analysis simple, assume the homogeneity in skills ($s_i = s$ for all *i*) and the following condition:

**Assumption 1:** $s > c_i$.

Assumption 1 means that the unit effort cost is not too large compared to the material return from task-solving. A calculation suggests that under Assumption 1, teams earn more when the reduction policy is imposed in their group than otherwise if $s > r_{k\in\{\mathrm{S,NS}\}}$, irrespective of whether they are in the ENDO or EXO treatment. However, if $s < r_{k\in\{\mathrm{S,NS}\}}$, sacrifice is purely harmful to welfare.

**Summary 2**: *Suppose Assumption 1 holds. Regardless of whether they are in the ENDO or EXO treatment, if the material benefit from working is high (low) enough that* $s > r_{k\in\{\mathrm{S,NS}\}}$ $(s < r_{k\in\{\mathrm{S,NS}\}})$*, teams earn more (less) when the reduction policy is imposed than is not imposed.*

Proof: Suppose first that $s > r_{k\in\{\mathrm{S,NS}\}}$, i.e., the situations are characterized by interior solutions. Then, we can show the beneficial effect of sacrifice by simply calculating the difference in the payoff between the two conditions.

$$\pi_i{}^{\mathrm{S}} = \pi_i\left(e_{i,\mathrm{S}}{}^{*}|e_{i,\mathrm{S}}{}^{*}\right) = 3s\frac{s+\mu D-r_{\mathrm{S}}}{2(c_i-\delta D)} + r_{\mathrm{S}}\cdot\left(1-\frac{s+\mu D-r_{\mathrm{S}}}{2(c_i-\delta D)}\right) - \frac{(s+\mu D-r_{\mathrm{S}})^2}{4(c_i-\delta D)}. \quad \text{(B5)}$$

$$\pi_i{}^{\mathrm{NS}} = \pi_i\left(e_{i,\mathrm{NS}}{}^{*}|e_{i,\mathrm{NS}}{}^{*}\right) = 3s\frac{s+\mu D-r_{\mathrm{NS}}}{2(c_i-\delta D)} + r_{\mathrm{NS}}\cdot\left(1-\frac{s+\mu D-r_{\mathrm{NS}}}{2(c_i-\delta D)}\right) - \frac{(s+\mu D-r_{\mathrm{NS}})^2}{4(c_i-\delta D)}. \quad \text{(B6)}$$

Then, $\pi_i{}^{\mathrm{S}} - \pi_i{}^{\mathrm{NS}} = 3s\frac{r_{\mathrm{NS}}-r_{\mathrm{S}}}{2(c_i-\delta D)} + r_{\mathrm{S}}\cdot\left(1-\frac{s+\mu D-r_{\mathrm{S}}}{2(c_i-\delta D)}\right) - r_{\mathrm{NS}}\cdot\left(1-\frac{s+\mu D-r_{\mathrm{NS}}}{2(c_i-\delta D)}\right) - \frac{(s+\mu D-r_{\mathrm{S}})^2}{4(c_i-\delta D)} + \frac{(s+\mu D-r_{\mathrm{NS}})^2}{4(c_i-\delta D)}$

$$= \frac{r_{\mathrm{NS}}-r_{\mathrm{S}}}{4(c_i-\delta D)}\{6s - 4c_i - (r_{\mathrm{NS}}+r_{\mathrm{S}}) + 4\delta D\} > 0. \quad \text{(B7)}$$

Consider next the case of corner solutions both with and without sacrifice (i.e., $s < r_{k\in\{\mathrm{S,NS}\}}$). In this case, $e_{i,\mathrm{S}}{}^{*} = e_{i,NS}{}^{*} = 0$; written differently, $g_{i,\mathrm{S}}{}^{*} = g_{i,NS}{}^{*} = 1$. Then, from Equation (B1), $\pi_i{}^{\mathrm{S}} - \pi_i{}^{\mathrm{NS}} < 0$.

The remaining case is the situation with an interior solution under sacrifice but a corner solution without sacrifice ($r_{\mathrm{NS}} > s > r_{\mathrm{S}}$,). Then, $e_{i,\mathrm{S}}{}^{*} = \frac{s_i-r_{\mathrm{S}}}{2(c_i-\delta D)}$, and $e_{i,NS}{}^{*} = 0$), and we have the following:

$$\pi_i{}^{\mathrm{S}} = \pi_i\left(e_{i,\mathrm{S}}{}^{*}|e_{i,\mathrm{S}}{}^{*}\right) = 3s\frac{s+\mu D-r_{\mathrm{S}}}{2(c_i-\delta D)} + r_{\mathrm{S}}\cdot\left(1-\frac{s+\mu D-r_{\mathrm{S}}}{2(c_i-\delta D)}\right) - \frac{(s+\mu D-r_{\mathrm{S}})^2}{4(c_i-\delta D)}.$$

$$\pi_i{}^{\mathrm{NS}} = r_{\mathrm{NS}}.$$

In this case, it is not obvious which is larger, $\pi_i{}^{\mathrm{S}}$ or $\pi_i{}^{\mathrm{NS}}$.

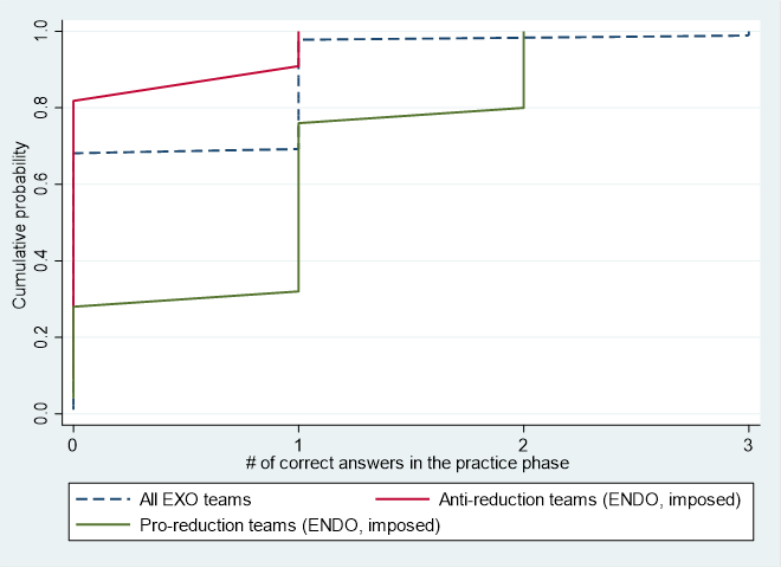
Notice that $\pi_i{}^{\mathrm{S}} - \pi_i{}^{\mathrm{NS}}$ in Equation (B7) is decreasing in $c_i$. This means that teams who are better at solving the collaborative counting task (i.e., teams with smaller $c_i$) have larger gains from the reduction policy through its strong positive impact on their effort provision.

It should be emphasized here that the beneficial effects of the reduction policy emerge when teams have sufficiently low effort costs as expressed by Assumption 1. However, introducing the policy is oppositely harmful if they are not skilled and therefore incur large costs from effort provision. This theoretical implication suggests that teams who are skilled at solving the collaborative counting task vote in favor of the reduction policy in phase 2.

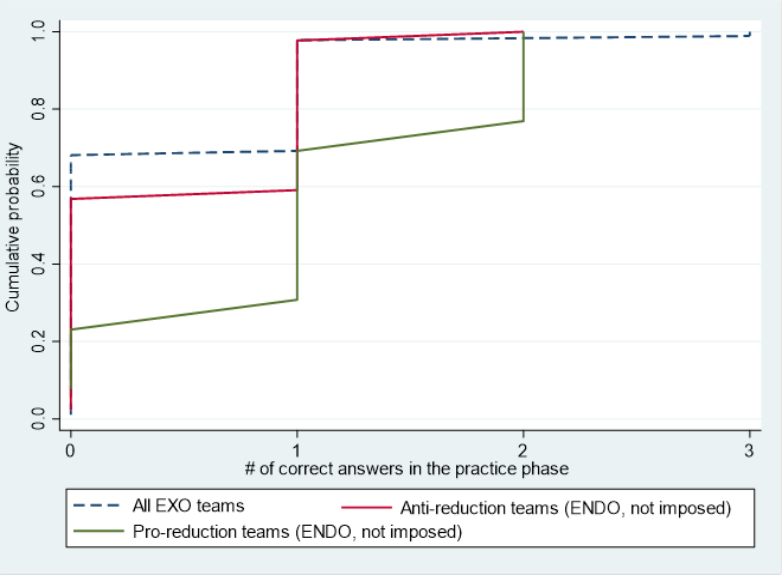
**Summary 3:** *Those who are better at solving the collaborative counting task in phase 1 are more likely to vote for the reduction policy in phase 2.*

## Appendix B.C: Additional Figure and Tables

**Figure C.1.** *Cumulative Distribution of Performance by Voting in the Three-minutes Practice Phase*




(A) Pro- and anti-reduction teams in the groups where the policy was endogenously imposed




(B) Pro- and anti-reduction teams in the groups where the policy was not endogenously imposed

*Note*: The cumulative distribution of teams' performance in the EXO treatment was also drawn as a reference.

**Table C.1**: *Privately and Socially Optimal Choices between Task-Solving and Gaming*

**[1. Which activity is <u>privately optimal</u>, task-solving or gaming?]**

**A. ENDO treatment**

| | A1: Under the reduction policy | | A2: Without reduction policy | |
|---|---|---|---|---|
| | a. Task-solving is a privately optimal | b. Gaming is a privately optimal | c. Task-solving is a privately optimal | d. Gaming is a privately optimal |
| i. # of team | 8 | 28 | 0 | 57 |
| ii. Avg # of correct answer per minute | 0.97 | 0.38 | --- | 0.33 |
| iii. Avg work time (min) | 33.33 | 26.42 | --- | 18.69 |

| | A3: All data in the ENDO | |
|---|---|---|
| | e. Task-solving is a privately optimal | f. Gaming is a privately optimal |
| i. # of team | 8 | 85 |
| ii. Avg # of correct answer per minute | 0.97 | 0.35 |
| iii. Avg work time (min) | 33.33 | 21.24 |

**B. EXO treatment**

| | B1: With reduction policy | | B2: Without reduction policy | |
|---|---|---|---|---|
| | a. Task-solving is a privately optimal | b. Gaming is a privately optimal | c. Task-solving is a privately optimal | d. Gaming is a privately optimal |
| i. # of team | 3 | 38 | 1 | 49 |
| ii. Avg # of correct answer per minute | 0.94 | 0.37 | 1 | 0.30 |
| iii. Avg work time (min) | 30.74 | 25.86 | 35.00 | 21.00 |

| | B3: All data in the EXO | |
|---|---|---|
| | e. Task-solving is a privately optimal | f. Gaming is a privately optimal |
| i. # of team | 4 | 87 |
| ii. Avg # of correct answer per minute | 0.96 | 0.33 |
| iii. Avg work time (min) | 31.80 | 23.12 |

**[2. Which activity is <u>socially optimal</u>, task-solving or gaming?]**

**A. ENDO treatment**

| | A1: Under the reduction policy | | A2: Without reduction policy | |
|---|---|---|---|---|
| | a. Task-solving is a socially optimal | b. Gaming is a socially optimal | c. Task-solving is a socially optimal | d. Gaming is a socially optimal |
| i. # of team | 28 | 8 | 29 | 28 |
| ii. Avg # of correct answer per minute | 0.64 | 0.08 | 0.58 | 0.07 |
| iii. Avg work time (min) | 32.10 | 13.43 | 26.98 | 10.11 |

| | A3: All data in the ENDO | |
|---|---|---|
| | e. Task-solving is a socially optimal | f. Gaming is a socially optimal |
| i. # of team | 57 | 36 |
| ii. Avg # of correct answer per minute | 0.61 | 0.07 |

| | | |
|---|---|---|
| iii. Avg work time (min) | 29.50 | 10.85 |

**B. EXO treatment**

| | B1: With reduction policy | | B2: Without reduction policy | |
|---|---|---|---|---|
| | a. Task-solving is a socially optimal | b. Gaming is a socially optimal | c. Task-solving is a socially optimal | d. Gaming is a socially optimal |
| i. # of team | 32 | 9 | 24 | 26 |
| ii. Avg # of correct answer per minute | 0.51 | 0.07 | 0.58 | 0.07 |
| iii. Avg work time (min) | 28.50 | 18.09 | 28.81 | 14.33 |

| | B3: All data in the EXO | |
|---|---|---|
| | e. Task-solving is a socially optimal | f. Gaming is a socially optimal |
| i. # of team | 56 | 35 |
| ii. Avg # of correct answer per minute | 0.54 | 0.07 |
| iii. Avg work time (min) | 28.63 | 15.29 |

## Appendix B.D: Coding Procedure and Analysis Results for the Communication Contents

*D.1. Coding Procedure*

Two coders were hired to judge each team's communication content for both the 3-minute communication segment prior to voting in the ENDO treatment and the 35-minute communication segments in the main task-solving phase for both the ENDO and EXO treatments, by assigning the relevant codes (summarized in section D.2). The treatments were presented as 'Treatment A' and Treatment B' to the coders, which alternated as coders completed the treatments in different orders. The coders were provided with a copy of the experiment instructions. Each coder was provided with three Excel files, termed "Coding Sheet – Treatment XY," where X indicates either "A" or "B" to designate the treatment and Y indicates the communication length, either 3 or 35 minutes (e.g., Coding Sheet – TreatmentA35). Each file had separate sections for each code type and only the relevant codes for that communication type were available. In the columns, the files contained a list of the team numbers in ascending order, which corroborated with "Segment" numbers found in the "Communication Files."

Six Communication files were also provided, comprising a sample set of ten communication segments and the remaining set of communication segments for each of the three communication combinations. Coders were instructed to first read the entire communication segment of a team and then assign as many codes as deemed appropriate in the Coding Sheet in a given teams column.

The files consisted of data from 93 teams in the ENDO treatment, each having one 3-minute and one 35-minute dialogue segment, and 91 teams from the EXO treatment, with just a 35-minute dialogue segment each, resulting in 277 dialogue segments to be coded. Coding was conducted by treatment and communication type, and further broken into four blocks as detailed below. While the coders were aware that there were two coders, they were kept anonymous from each other for the entire process and so were unable to communicate with each other.

<u>The first block (first nine days):</u>

The coding sheet, experiment instructions, and 10-segment sample communication file for the 3-minute communication in Treatment A (35-minute communication in Treatment B for the other coder) were provided on the first day. A meeting was scheduled for the same day, separately for each coder, to allow one of the researchers to explain the coding process and treatment in more detail. Coders were not made aware of the purpose of the research, subject details, or any of the analysis/results throughout the coding process.

After the sample set had been coded, a researcher met with each coder to discuss any problems or difficulties. This initial practice and feedback process took two days. After that, the

researchers sent the communication file with the remaining 83 (81) dialogue segments for that block to be completed over the following seven days.

Once all 93 or 91 dependent on the coder (including the sample set) of the dialogue segments had been coded, the Coding Sheet was returned to the researchers and no further changes could be made (unless there had been some misunderstanding about the coding practice). Feedback was not given to the coders regarding their coding practice.

The second block (next nine days):

Once the first block was completed, the coders were given the coding sheet and 10-segment sample communication for the 35-minute communication in Treatment A (3-minute communication in Treatment A for the other coder, along with the instructions for Treatment A), and a meeting was scheduled for that day, separately for each coder, to go through the instructions and codes. The remaining procedure is the same as in the first block. Two days were given to complete the sample set, after which there was a meeting to clarify any questions. After that, seven days were given to code the remaining 83 dialogue segments. As before, no feedback was given to the coders regarding their coding practice.

The third block (next nine days):

As in the first and second block, the coders were provided with the instructions, coding sheet and 10-segment sample communication for the 35-minute communication in Treatment B (35-minute communication in Treatment A for the other coder), and a meeting was held to discuss the instructions and codes. Two days were given to complete the 10-segment sample set, before a further meeting was held to clarify any questions. The coders were allowed a further seven days to code the remaining 81 (83) dialogue segments. No feedback was given regarding the coders' coding practices.

The fourth block (final seven days):

The coding results were compared for discrepancies between the two coders' coding results. The discrepancies were then highlighted in the Excel spreadsheets and a copy was given to each coder. The coders were given a further seven days to re-evaluate these discrepancies, with the additional knowledge of one another's codes, and to either confirm or alter their initial findings. Each coder was informed that their codes would be sent to the other coder, and that they would simultaneously re-evaluate the discrepancies. Coder identity remained anonymous throughout the process (and also after the coding work) and no communication was permitted.

*D.2. Full List of Codes*

(a) ENDO 3-minute communication immediately before voting

| Code | Description | Interpretation (not shown to coders) |
|---|---|---|
| Note: Coders may assign as many codes as appropriate, including assigning no codes, to a dialogue segment. | | |
| **Codes related to voting decision** | | |
| A1 | Agreement among the three team members explicitly to vote for the policy | Consensus |
| A2 | Agreement among the three team members explicitly to vote against the policy | Consensus |
| A3 | The team's majority favored decision changes over the course of the discussion | Learning |
| A4 | The team's majority favored decision does not change over the course of the discussion | Learning |
| A5 | Disagreement on what to vote for at the beginning of communication which is then resolved | Disagreement |
| A6 | One or more teammates decide to cast their own preference (leaving the majority rule to decide the team vote) | Disagreement |
| A7 | There is an unresolved split in opinion about whether to vote for or against the policy due to strong preferences on both sides | Disagreement |
| A8 | Teammates do not reach a consensus by the end of the 3-minute communication period for reasons other than A6 or A7 | Lack of time |
| A9 | Discuss how the other two teams may vote | Strategic |
| A10 | Confusion about the voting rule | Confusion |
| **Codes related to deciding what to do during the task-solving phase** | | |
| B1 | Agree/Imply to count as primary behavior | Cooperative |
| B2 | Agree/Imply to game as primary behavior | Uncooperative/free-riding |
| B3 | Agree to hybrid behavior e.g. so many tasks/minutes before switching to the game screen | Somewhat cooperative |
| B4 | Agree to discuss, decide and/or alter behavior during the counting task later (35-minute phase) based on performance/needs in Phase 2 | Flexible strategy |
| B5 | Suggest altering and/or discussing their behavior for the task-solving phase depending on the vote outcome | Rational |
| B6 | Confusion about the rules in phase 2 (e.g., the revenue-sharing rule) | Confusion |
| **Codes related to why they are pro/anti the policy** | | |
| C1 | Pro-policy to deter others from switching to the game screen by reducing the return (monetary deterrence) | Monetary incentive/punishment |
| C2 | Pro-policy to signal intention to complete the tasks to other teams | Signaling/Information Effects |
| C3 | Pro-policy for a normative reason e.g. it is the right thing to do, it is desirable socially for their group, it is morally good | Normative reasoning |
| C4 | Pro-policy as the policy is perceived as fair, i.e., reduces income inequality among subjects in a given group | Fairness preference |
| C5 | Pro-policy out of spite or enjoyment of punishment | Spitefulness |
| C6 | Pro-policy out of anticipated anger should other teams not complete tasks | Emotive reasoning/anger |
| C7 | Pro-policy for strategic reasons (e.g., induce other teams to complete the task while they themselves do not complete the task) | Strategic |
| C8 | Anti-policy as they intend to game for at least some of the task-solving period | Selfish |
| C9 | Anti-policy as they like unfair distribution of income | Fairness preference (reverse) |
| C10 | Anti-policy as they dislike punishment philosophically (e.g., dislike coercive punishment), and/or do not perceive other teams' gaming as negative | Dislike of punishment on private activity |
| C11 | Anti-policy as they are uncertain about whether they will want to access the game screen | Uncertainty |
| C12 | Express that the policy is not strong enough to deter others switching to the game screen (monetary) | Punishment strength |
| C13 | Express that the policy is unlikely to affect other teams' working behavior | Policy insensitivity |

| | | |
|---|---|---|
| **Codes related to why they chose a certain behavior** | | |
| D1 | Discuss ability to complete X tasks in Y minutes | Skills/Ability |
| D2 | Believe they as a team make the most money from counting | Monetary incentive |
| D3 | Believe they as a team make the most money from gaming | Monetary incentive |
| D4 | Discuss their performance or comfort in Phase 1 (strong/positive) | Incentives/Rationality |
| D5 | Discuss their performance or comfort in Phase 1 (weak/negative) | Incentives/Rationality |
| D6 | Discuss behavior in terms of guaranteed pay (game screen) versus uncertain pay (tasks) | Uncertainty |
| D7 | Discuss use of the game screen as a break | Team behavior: Strategizing |
| D8 | Discuss uncertainty surrounding other teams' work choices or abilities | Uncertainty |
| D9 | Suggest distrust of other teams e.g. expect them to take advantage | Group behavior: Distrust |
| D10 | Discuss expected fatigue from performing the task in Phase 2 based on their experience in Phase 1 | |
| D11 | Discuss how total earnings for the whole group are maximized if all 3 teams work on the counting task | Pareto Efficiency |
| D12 | Discuss how their vote affects the vote outcome | Pivotal voting |
| **Codes related to team behavior** | | |
| E1 | Negativity towards teammates e.g. being critical of others or discouraging | Team behavior: Negative communication |
| E2 | Positivity towards teammates e.g. attempts to encourage others or being supportive | Team behavior: Positive communication |
| E3 | Discusses or suggests task-related behavior e.g. double-counting, waiting for each other, a specific method, etc. | Team behavior: Strategizing |
| E4 | No communication of the entire team | Team behavior: No communication |
| E5 | No communication from just 1 or 2 team members | Team behavior: Limited/poor communication |
| E6 | Chat about topics unrelated to the experiment | Team behavior: team identity |
| E7 | Resolve some confusion about the experiment through communication | Team behavior: Communication |

(b) ENDO 35-minute communication during the main task-solving phase

| **Code** | **Description** | **Interpretation (<u>not</u> shown to coders)** |
|---|---|---|
| Note: Coders may assign as many codes as appropriate, including assigning no codes, to a dialogue segment. | | |
| **Codes related to voting outcome** | | |
| F1 | Express negative emotions (e.g., upset, anger) about the outcome of the vote | Group identity (negative) |
| F2 | Express positive emotions (e.g., happiness) about the outcome of the vote | Group identity (positive) |
| F3 | Agree/Imply to count as primary behavior | Cooperative |
| F4 | Agree/Imply to game as primary behavior | Uncooperative/free-riding |
| F5 | Agree to hybrid behavior e.g. so many tasks/minutes before switching to the game screen | Somewhat cooperative |
| F6 | Agree to discuss, decide and/or alter behavior during the counting task later (35-minute phase) based on performance/needs in Phase 2 | Flexible strategy |
| F7 | Express belief/hope that other teams will complete tasks following the vote | Positive expectations |
| F8 | Express belief that teams will not complete tasks following the vote | Negative expectations |
| F9 | Discuss the distribution of votes and predict how each team may respond to one another | Signaling |
| F10 | Belief on other teams' responses: pro-policy teams will work hard | Belief on voter type |
| F11 | Belief on other teams' responses: anti-policy teams will work hard | Belief on voter type |
| F12 | Belief on other teams' responses: pro-policy teams will work little | Belief on voter type |

| | | |
|---|---|---|
| F13 | Belief on other teams' responses: anti-policy teams will work little | Belief on voter type |
| F14 | Discuss some confusion (e.g., they voted based on some misunderstanding of the experiment) | Confusion |
| F15 | Discuss whether to change behavior based on the vote outcome | Conditional cooperation/rational |
| **Codes related to task performance** | | |
| G1 | Discuss wanting to switch to the game screen some time during the task-solving phase | Fatigue/inability |
| G2 | Discuss difficulty/unpleasantness of task e.g. being slow, tired, bored, etc. | Fatigue/inability |
| G3 | Enact hybrid behavior e.g. set a given number of tasks/minutes before switching to the game screen and back | |
| G4 | Expression of strong negative emotion e.g. frustration, anger, disappointment | Emotive: Negative |
| G5 | Expression of strong positive emotion e.g. enjoyment, things are going well | Emotive: Positive |
| **Codes related to why they chose/are choosing a certain behavior** | | |
| D1 | Discuss ability to complete X tasks in Y minutes | Rationality |
| D2 | Believe they as a team make the most money from counting | Monetary Incentive |
| D3 | Believe they as a team make the most money from gaming | Monetary Incentive |
| D4 | Discuss their performance or comfort in Phase 1 and/or so far in Phase 2 (strong/positive) | Rationality |
| D5 | Discuss their performance or comfort in Phase 1 and/or so far in Phase 2 (weak/negative) | Rationality |
| D6 | Discuss behavior in terms of guaranteed pay (game screen) versus uncertain pay (tasks) | Uncertainty |
| D7 | Discuss use of the game screen as a break | Team behavior: Strategizing |
| D8 | Discuss uncertainty surrounding other teams' work choices or abilities | Uncertainty |
| D9 | Suggest distrust of other teams e.g. expect them to take advantage | Group behavior: Distrust |
| D10 | Compare their estimated earnings so far from task-solving and forgone earnings from not staying in the Game screen | Rationality |
| D11 | Discuss how total earnings for the whole group are maximized if all 3 teams work on the counting task | Pareto Efficiency |
| **Codes related to team behavior** | | |
| H1 | Evident mismatch in behavior e.g. one teammate switches to the game screen against others' wishes | Team behavior: miscommunication/selfishness |
| H2 | 1 or more players discuss being trapped in the waiting screen | Team behavior: miscommunication/poor planning |
| H3 | Positivity towards teammates e.g. attempts to encourage others or being supportive | Team behavior: Negative communication |
| H4 | Negativity towards teammates e.g. being critical of others or discouraging | Team behavior: Positive communication |
| H5 | Discusses or suggests task-related behavior e.g. double-counting, waiting for each other, a specific method, etc. | Team behavior: Strategizing |
| H6 | Checking whether their teammates refrain from using the Game screen | Team behavior: Monitoring |
| H7 | Communicate about their Tetris score or enjoying playing Tetris | Team behavior: Enjoy shirking |
| H8 | Disagreement on what to do (count or gaming) at the beginning of communication | Team behavior: Disagreement |
| H9 | No communication of the entire team | Team behavior: No communication |
| H10 | No communication from just 1 or 2 team members; or ignore messages from their teammates | Team behavior: Limited/poor communication |

| | | |
|---|---|---|
| H11 | A team exclusively communicates numbers (for the counting task) throughout the 35-minute phase | Team behavior |
| H12 | A player/s states that they have not or do not spend any time in the game screen | Strategic/Lying |
| H13 | Chat about topics unrelated to the experiment | Team behavior: team identity |
| H14 | Resolve some confusion about the experiment through communication | Team behavior: Communication |

(c) EXO 35-minute communication during the main task-solving phase

| **Code** | **Description** | **Interpretation (<u>not</u> shown to coders)** |
|---|---|---|
| Note: Coders may assign as many codes as appropriate, including assigning no codes, to a dialogue segment. | | |
| **Codes related to policy outcome** | | |
| I1 | Express negative emotions (e.g., upset, anger) about the policy outcome | Group identity (negative) |
| I2 | Express positive emotions (e.g., happiness) about the policy outcome | Group identity (positive) |
| I3 | Agree/Imply to count as primary behavior | Cooperative |
| I4 | Agree/Imply to game as primary behavior | Uncooperative/free-riding |
| I5 | Agree to hybrid behavior e.g. so many tasks/minutes before switching to the game screen | Somewhat cooperative |
| I6 | Agree to discuss, decide and/or alter behavior during the counting task later (35-minute phase) based on performance/needs in Phase 2 | Flexible strategy |
| I7 | Express belief/hope that other teams will complete tasks following the policy outcome | Positive expectations |
| I8 | Express belief that teams will not complete tasks following the policy outcome | Negative expectations |
| I9 | Discuss whether the method that the computer randomly decides whether the policy is implemented is fair | |
| I10 | Discuss whether the method (computer's random choice) is accurate as described by the instructions (e.g., the computer's choice may not be random) | |
| I11 | Discuss some confusion | |
| **Codes related to task performance** | | |
| G1 | Discuss wanting to switch to the game screen some time during the task-solving phase | Fatigue/inability |
| G2 | Discuss difficulty/unpleasantness of task e.g. being slow, tired, bored, etc. | Fatigue/inability |
| G3 | Enact hybrid behavior e.g. set a given number of tasks/minutes before switching to the game screen and back | Team behavior: Strategizing |
| G4 | Expression of strong negative emotion e.g. frustration, anger, disappointment | Emotive: Negative |
| G5 | Expression of strong positive emotion e.g. enjoyment, things are going well | Emotive: Positive |
| **Codes related to why they chose/are choosing a certain behavior** | | |
| D1 | Discuss ability to complete X tasks in Y minutes | Rationality |
| D2 | Believe they as a team make the most money from counting | Rationality |
| D3 | Believe they as a team make the most money from gaming | Rationality |
| D4 | Discuss their performance or comfort in Phase 1 and/or so far in Phase 2 (strong/positive) | Rationality |
| D5 | Discuss their performance or comfort in Phase 1 and/or so far in Phase 2 (weak/negative) | Rationality |
| D6 | Discuss behavior in terms of guaranteed pay (game screen) versus uncertain pay (tasks) | Uncertainty |
| D7 | Discuss use of the game screen as a break | Team behavior: Strategizing |
| D8 | Discuss uncertainty surrounding other teams' work choices or abilities | Uncertainty |
| D9 | Suggest distrust of other teams e.g. expect them to take advantage | Group behavior: Distrust |
| D10 | Compare their estimated earnings so far from task-solving and forgone earnings from not staying in the Game screen | Rationality |
| D11 | Discuss how total earnings for the whole group are maximized if all 3 teams work on the counting task | Pareto Efficiency |
| **Codes related to team behavior** | | |

| | | |
|---|---|---|
| H1 | Evident Mismatch in behavior e.g. one teammate switches to the game screen against others' wishes | Team behavior: miscommunication or selfishness |
| H2 | 1 or more players discuss being trapped in the waiting screen | Team behavior: miscommunication/poor planning |
| H3 | Positivity towards teammates e.g. attempts to encourage others or being supportive | Team behavior: Negative communication |
| H4 | Negativity towards teammates e.g. being critical of others or discouraging | Team behavior: Positive communication |
| H5 | Discusses or suggests task-related behavior e.g. double-counting, waiting for each other, a specific method, etc. | Team behavior: Strategizing |
| H6 | Checking whether their teammates refrain from using the Game screen | Team behavior: Monitoring |
| H7 | Communicate about their Tetris score or enjoying playing Tetris | Team behavior: Enjoy shirking |
| H8 | Disagreement on what to do (count or gaming) at the beginning of communication | Team behavior: Disagreement |
| H9 | No communication of the entire team | Team behavior: No communication |
| H10 | No communication from just 1 or 2 team members; or ignore messages from their teammates | Team behavior: Limited/poor communication |
| H11 | A team exclusively communicates numbers (for the counting task) throughout the 35-minute phase | Team behavior |
| H12 | A player/s states that they have not or do not spend any time in the game screen | Strategic/Lying |
| H13 | Chat about topics unrelated to the experiment | Team behavior: team identity |
| H14 | Resolve some confusion about the experiment through communication | Team behavior: Communication |

*D.3. Agreement rates and Kappas*

The average Cohen's Kappas for the initial coding were 0.67, 0.60, and 0.45 for the ENDO 3-minute dialogue segments, ENDO 35-minute dialogue segments, and EXO 35-minute dialogue segments, respectively. The reconsideration step improved the Kappas. After the independent reconsideration process, the Kappas became 0.87, 0.87, and 0.78 for the ENDO 3-minute dialogue segments, ENDO 35-minute dialogue segments, and EXO 35-minute dialogue segments, respectively.

Remark: The overall agreement rates of coding between the two coders after (before) the reconsideration process were 96.3% (90.5%), 97.5% (93.0%), and 94.9% (88.5%) for the ENDO 3-minute dialogue segments, ENDO 35-minute dialogue segments, and EXO 35-minute dialogue segments, respectively.

The following summarizes the agreement rates and the Kappas before and after the reconsideration step for each code:

(a) ENDO 3-Minute Dialogue Segments

[Agreement Rate:]

| Agreement rate | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 96.8% | 97.8% | 93.5% | 80.6% | 90.3% | 91.4% | 92.5% | 96.8% | 88.2% | 93.5% |
| After reconsideration | 98.9% | 100.0% | 100.0% | 95.7% | 96.8% | 95.7% | 95.7% | 100.0% | 93.5% | 100.0% |

| Agreement rate | B1 | B2 | B3 | B4 | B5 | B6 |
|---|---|---|---|---|---|---|
| Before reconsideration | 86.0% | 93.5% | 89.2% | 88.2% | 96.8% | 90.3% |
| After reconsideration | 95.7% | 100.0% | 95.7% | 98.9% | 98.9% | 96.8% |

| Agreement rate | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 91.4% | 59.1% | 97.8% | 96.8% | 98.9% | 96.8% | 95.7% | 92.5% | 100.0% | 100.0% | 87.1% | 94.6% | 94.6% |
| After reconsideration | 98.9% | 67.7% | 97.8% | 100.0% | 100.0% | 100.0% | 97.8% | 97.8% | 100.0% | 100.0% | 97.8% | 97.8% | 98.9% |

| Agreement rate | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 94.6% | 77.4% | 91.4% | 96.8% | 95.7% | 87.1% | 89.2% | 79.6% | 86.0% | 91.4% | 72.0% | 82.8% |
| After reconsideration | 97.8% | 90.3% | 95.7% | 98.9% | 96.8% | 92.5% | 97.8% | 91.4% | 92.5% | 96.8% | 81.7% | 91.4% |

| Agreement rate | E1 | E2 | E3 | E4 | E5 | E6 | E7 |
|---|---|---|---|---|---|---|---|
| Before reconsideration | 93.5% | 65.6% | 81.7% | 100.0% | 97.8% | 96.8% | 92.5% |
| After reconsideration | 97.8% | 86.0% | 100.0% | 100.0% | 98.9% | 98.9% | 98.9% |

[Cohen's Kappa:]

| Kappa | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.93 | 0.96 | -0.03 | 0.00 | 0.27 | 0.29 | 0.49 | 0.39 | 0.36 | -0.02 |
| After reconsideration | 0.98 | 1.00 | 1.00 | 0.64 | 0.82 | 0.58 | 0.73 | 1.00 | 0.69 | 1.00 |

| Kappa | B1 | B2 | B3 | B4 | B5 | B6 |
|---|---|---|---|---|---|---|
| Before reconsideration | 0.69 | 0.80 | 0.75 | 0.61 | 0.65 | 0.00 |
| After reconsideration | 0.90 | 1.00 | 0.90 | 0.97 | 0.90 | 0.00 |

| Kappa | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.83 | 0.00 | 0.49 | -0.01 | 0.66 | -0.01 | 0.58 | 0.85 | n.a. | 1.00 | 0.28 | 0.42 | 0.00 |
| After reconsideration | 0.98 | 0.22 | 0.49 | 1.00 | 1.00 | 1.00 | 0.82 | 0.96 | n.a. | 1.00 | 0.91 | 0.79 | 0.85 |

| Kappa | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.85 | 0.55 | 0.65 | 0.82 | 0.86 | 0.63 | 0.58 | 0.53 | 0.31 | 0.51 | 0.31 | 0.04 |
| After reconsideration | 0.94 | 0.81 | 0.82 | 0.94 | 0.90 | 0.80 | 0.92 | 0.81 | 0.59 | 0.85 | 0.57 | 0.59 |

| Kappa | E1 | E2 | E3 | E4 | E5 | E6 | E7 |
|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.37 | 0.22 | 0.59 | n.a. | 0.49 | 0.71 | 0.43 |
| After reconsideration | 0.74 | 0.57 | 1.00 | n.a. | 0.66 | 0.92 | 0.90 |

(b) ENDO 35-Minute Dialogue Segments

[Agreement Rate:]

| Agreement rate | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 94.6% | 97.8% | 87.1% | 80.6% | 86.0% | 87.1% | 97.8% | 97.8% | 97.8% | 97.8% | 100.0% | 97.8% | 97.8% | 97.8% | 96.8% |
| After reconsideration | 98.9% | 100.0% | 81.7% | 96.8% | 95.7% | 93.5% | 97.8% | 98.9% | 100.0% | 98.9% | 100.0% | 100.0% | 97.8% | 98.9% | 100.0% |

| Agreement rate | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| Before reconsideration | 88.2% | 92.5% | 83.9% | 87.1% | 88.2% |
| After reconsideration | 100.0% | 97.8% | 100.0% | 92.5% | 96.8% |

| Agreement rate | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 98.9% | 96.8% | 96.8% | 92.5% | 92.5% | 97.8% | 86.0% | 96.8% | 96.8% | 98.9% | 94.6% |
| After reconsideration | 98.9% | 100.0% | 100.0% | 97.8% | 97.8% | 100.0% | 96.8% | 100.0% | 100.0% | 100.0% | 97.8% |

| Agreement rate | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 93.5% | 93.5% | 77.4% | 83.9% | 89.2% | 95.7% | 97.8% | 89.2% | 100.0% | 84.9% | 88.2% | 97.8% | 96.8% | 94.6% |
| After reconsideration | 97.8% | 97.8% | 91.4% | 92.5% | 95.7% | 98.9% | 98.9% | 94.6% | 100.0% | 95.7% | 93.5% | 98.9% | 97.8% | 98.9% |

[Cohen's Kappa:]

| Kappa | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.43 | 0.49 | 0.67 | 0.21 | 0.24 | -0.07 | 0.66 | 0.74 | 0.00 | 0.49 | n.a. | 0.00 | 0.49 | 0.00 | 0.39 |
| After reconsideration | 0.88 | 1.00 | 0.60 | 0.88 | 0.81 | 0.54 | 0.66 | 0.88 | 1.00 | 0.66 | n.a. | n.a. | 0.74 | 0.00 | 1.00 |

| Kappa | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| Before reconsideration | 0.76 | 0.62 | 0.57 | 0.43 | 0.12 |
| After reconsideration | 1.00 | 0.90 | 1.00 | 0.71 | 0.75 |

| Kappa | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.66 | 0.56 | 0.00 | 0.00 | -0.02 | 0.66 | 0.40 | 0.65 | -0.01 | 0.79 | -0.02 |
| After reconsideration | 0.66 | 1.00 | 1.00 | 0.82 | 0.79 | 1.00 | 0.86 | 1.00 | 1.00 | 1.00 | 0.74 |

| Kappa | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.80 | 0.64 | 0.28 | 0.15 | 0.62 | 0.48 | 0.00 | 0.40 | 1.00 | 0.54 | 0.66 | 0.49 | 0.71 | 0.59 |
| After reconsideration | 0.93 | 0.88 | 0.75 | 0.59 | 0.86 | 0.88 | 0.66 | 0.75 | 1.00 | 0.88 | 0.81 | 0.79 | 0.82 | 0.92 |

(c) EXO 35-Minute Dialogue Segments

[Agreement Rate:]

| Agreement rate | I1 | I2 | I3 | I4 | I5 | I6 | I7 | I8 | I9 | I10 | I11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 100.0% | 100.0% | 5.4% | 72.0% | 87.1% | 90.3% | 95.7% | 100.0% | 100.0% | 100.0% | 92.5% |
| After reconsideration | 100.0% | 100.0% | 41.9% | 84.9% | 97.8% | 97.8% | 96.8% | 100.0% | 100.0% | 100.0% | 98.9% |

| Agreement rate | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| Before reconsideration | 89.2% | 92.5% | 72.0% | 82.8% | 89.2% |
| After reconsideration | 98.9% | 97.8% | 81.7% | 93.5% | 95.7% |

| Agreement rate | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 92.5% | 88.2% | 89.2% | 93.5% | 92.5% | 97.8% | 86.0% | 98.9% | 98.9% | 91.4% | 96.8% |
| After reconsideration | 96.8% | 95.7% | 96.8% | 94.6% | 96.8% | 98.9% | 94.6% | 98.9% | 98.9% | 97.8% | 97.8% |

| Agreement rate | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 92.5% | 89.2% | 76.3% | 83.9% | 68.8% | 81.7% | 97.8% | 86.0% | 100.0% | 89.2% | 89.2% | 97.8% | 95.7% | 84.9% |
| After reconsideration | 97.8% | 94.6% | 89.2% | 92.5% | 93.5% | 88.2% | 97.8% | 95.7% | 100.0% | 98.9% | 96.8% | 98.9% | 98.9% | 94.6% |

[Cohen's Kappa:]

| Kappa | I1 | I2 | I3 | I4 | I5 | I6 | I7 | I8 | I9 | I10 | I11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 1.00 | n.a. | 0.00 | 0.05 | 0.13 | 0.00 | 0.00 | n.a. | n.a. | n.a. | 0.00 |
| After reconsideration | 1.00 | n.a. | 0.01 | 0.48 | 0.90 | 0.86 | 0.39 | n.a. | n.a. | n.a. | 0.90 |

| Kappa | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| Before reconsideration | 0.79 | 0.70 | 0.41 | 0.34 | 0.12 |
| After reconsideration | 0.98 | 0.92 | 0.63 | 0.75 | 0.64 |

| Kappa | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.50 | 0.59 | 0.44 | 0.22 | 0.49 | 0.88 | 0.54 | 0.88 | 0.00 | -0.03 | 0.56 |
| After reconsideration | 0.75 | 0.83 | 0.81 | 0.27 | 0.78 | 0.94 | 0.78 | 0.88 | 0.00 | 0.66 | 0.74 |

| Kappa | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Before reconsideration | 0.82 | 0.39 | 0.41 | 0.10 | 0.28 | 0.07 | 0.49 | 0.47 | n.a. | 0.71 | 0.71 | 0.49 | 0.48 | 0.39 |
| After reconsideration | 0.95 | 0.68 | 0.73 | 0.55 | 0.87 | 0.51 | 0.49 | 0.87 | n.a. | 0.97 | 0.91 | 0.66 | 0.85 | 0.75 |

*Note*: The Cohen's Kappa cannot be calculated where a code that is not used by either coder; these are marked with "n.a."

### *D.4. Regression Analysis*

This section reports a regression analysis to explore subjects' reasoning behind their voting decisions regarding the reduction policy and their task-solving behavior, utilizing the classified codes (see Section D.2 for the full list of codes). Following the convention in the experimental literature on team decision-making, the codes with Kappa values greater than 0.4 were used in each model.

As listed in Section D.2, five kinds of coding categories (Code As, Bs, Cs, Ds and Es) were used to classify the subjects' reasoning behind voting (the ENDO 3 Minutes Dialogue). Four kinds of coding categories (Code Fs, Gs, Ds, and Hs for the ENDO treatment; Code Is, Gs, Ds, and Hs for the EXO treatment) were used to classify their reasoning in task-solving. As each coding category classifies the same behavior just from a different angle, having all codes altogether in a regression leads to serious collinearity. The codes of one coding category are therefore included as independent variables in each model in the following analyses. Note that as shown in Section D.3, the Kappa values of almost all classified codes are more than 0.4.

(a) Voting whether to implement the reduction policy (An analysis for the ENDO 3-Minute Dialogue Segments)

The following table reports results when using Code Bs, Cs, Ds, and Es as independent variables in Models 1, 2, 3 and 4, respectively. Code As (Codes related to the voting decision), such as "A1: Agreement among the three team members explicitly to vote for the policy," have almost the same information as the teams' voting decisions. Unsurprisingly, collinearity is strong and no meaningful results are obtained when using Code As as independent variables. The results are omitted to conserve space.

**Table D.1:** *Codes and their Impact on Voting for the Reduction Policy*

Dependent variable: A dummy which equals 1 if team *i* voted for (against) the reduction policy

| **Model 1**: Using codes related to deciding what to do during task-solving as independent variables | | **Model 2**: Using codes related to why they are pro/anti the policy as independent variables | | **Model 3**: Using codes related to why they chose a certain behavior as independent variables | | **Model 4**: Using codes related to team behavior as independent variables | |
|---|---|---|---|---|---|---|---|
| Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates |
| Code B1 | 0.31*** (0.11) | Code C1 | 0.57*** (0.10) | Code D1 | 0.07 (0.11) | Code E1 | -0.21 (0.28) |
| Code B2 | -0.26* (0.13) | Code C3 | 0.06 (0.05) | Code D2 | 0.41*** (0.11) | Code E2 | 0.30* (0.15) |
| Code B3 | -0.27** (0.11) | Code C4 | 0.06 (0.05) | Code D3 | 0.04 (0.13) | Code E3 | -0.11 (0.11) |
| Code B4 | -0.27** (0.12) | Code C5 | -0.21 (0.16) | Code D4 | 0.28 (0.17) | Code E5 | -0.31*** (0.08) |
| Code B5 | 0.02 (0.11) | Code C6 | 0.26 (0.17) | Code D5 | -0.26** (0.10) | Code E6 | 0.19 (0.24) |
| Constant | 0.40*** (0.12) | Code C7 | 0.13 (0.11) | Code D6 | -0.21 (0.13) | Code E7 | -0.18 (0.21) |
| # of obs. | 93 | Code C8 | -0.41*** (0.09) | Code D7 | -0.16 (0.15) | Constant | 0.42*** (0.08) |
| R-squared | 0.34 | Code C10 | -0.10 (0.13) | Code D8 | -0.03 (0.12) | # of obs. | 93 |
| | | Code C11 | -0.27** (0.13) | Code D9 | 0.19 (0.19) | R-squared | 0.07 |
| | | Code C12 | -0.30 (0.23) | Code D10 | 0.14 (0.12) | | |
| | | Code C13 | 0.05 (0.14) | Code D11 | 0.05 (0.14) | | |
| | | Constant | 0.37*** (0.09) | Code D12 | 0.19 (0.17) | | |
| | | # of obs. | 93 | Constant | 0.22*** (0.08) | | |
| | | R-squared | 0.72 | # of obs. | 93 | | |
| | | | | R-squared | 0.40 | | |

*Notes*: Linear regressions with robust standard errors. The numbers in parentheses are standard errors. Codes whose Kappa values are equal to or above 0.4 are used as independent variables. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

(b) Task-solving (An analysis for the ENDO/EXO 35-Minute Dialogue Segments)

Subjects' behaviors in the main task-solving phase can be characterized as (i) their work time (minutes), i.e., the duration in which they work on counting, rather than staying in the Game screen, and (ii) work productivity, i.e., the number of correct answers per minute of work time. The regression analysis below reports two versions for each coding category: one with the work time as the dependent variable, and the other with the work productivity as the dependent variable.

Subsections b1, b2, b3, and b4 below each include regression results of the ENDO treatment (Models 1 and 2) and of the EXO treatment (Models 3 and 4), side by side, to make comparison easier.

**Table D.2:** *Codes and their Impact on Task-Solving Behavior*

b1: Codes related to voting outcome (Code Fs) or policy outcome (Code Is))

| ENDO (Code Fs)[#1] | | | | EXO (Code Is) | | | |
|---|---|---|---|---|---|---|---|
| Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | | Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | |
| Model 1 | | Model 2 | | Model 3 | | Model 4 | |
| Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates |
| Code F1 | -14.77*** (1.71) | Code F1 | -0.11 (0.17) | Code I1 | -8.94*** (2.98) | Code I1 | -0.24*** (0.07) |
| Code F2 | -3.01 (6.47) | Code F2 | -0.03 (0.21) | Code I4 | -21.71*** (2.60) | Code I4 | -0.41*** (0.05) |
| Code F3 | 11.27*** (2.70) | Code F3 | 0.24*** (0.07) | Code I5 | -9.28*** (3.09) | Code I5 | -0.19** (0.08) |
| Code F4 | -16.15*** (2.90) | Code F4 | -0.25*** (0.09) | Code I6 | -7.47* (3.92) | Code I6 | -0.11 (0.07) |
| Code F5 | -3.98** (1.85) | Code F5 | -0.09* (0.05) | Code I11 | 2.08 (1.85) | Code I11 | 0.07 (0.06) |
| Code F6 | -3.93 (6.04) | Code F6 | -0.28* (0.15) | Cons. | 27.33*** (0.98) | Cons. | 0.43*** (0.04) |
| Code F7 | 6.83*** (0.53) | Code F7 | 0.24** (0.12) | | | | |
| Code F8 | 5.25*** (1.03) | Code F8 | 0.28*** (0.04) | | | | |
| Code F9 | 13.64*** (0.48) | Code F9 | 0.17 (0.10) | | | | |
| Code F13 | 2.46 (2.99) | Code F13 | -0.41*** (0.15) | | | | |
| Code F15 | 4.80*** (1.15) | Code F15 | 0.07 (0.17) | | | | |
| Cons. | 18.47*** (2.73) | Cons. | 0.32*** (0.06) | | | | |
| # of obs. | 93 | # of obs. | 93 | | | | |
| R-squared | 0.46 | R-squared | 0.31 | | | | |

b2: Codes related to task performance (Code Gs)

| ENDO (Code Gs) | | | | EXO (Code Gs) | | | |
|---|---|---|---|---|---|---|---|
| Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | | Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | |
| Model 1 | | Model 2 | | Model 3 | | Model 4 | |
| Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates |
| Code G1 | -1.80 (3.43) | Code G1 | -0.11 (0.08) | Code G1 | -11.08*** (2.64) | Code G1 | -0.16** (0.07) |
| Code G2 | 5.32* (3.01) | Code G2 | -0.01 (0.07) | Code G2 | 5.93** (2.41) | Code G2 | 0.16*** (0.06) |
| Code G3 | 2.13 (3.07) | Code G3 | 0.06 (0.07) | Code G3 | 6.66** (2.67) | Code G3 | 0.14** (0.07) |
| Code G4 | -1.64 (3.05) | Code G4 | -0.16 (0.08) | Code G4 | -3.86 (2.91) | Code G4 | -0.22*** (0.06) |
| Code G5 | 5.60 (5.62) | Code G5 | 0.03 (0.11) | Code G5 | 7.66*** (2.63) | Code G5 | 0.30** (0.13) |
| Cons. | 21.86*** (2.42) | Cons. | 0.46*** (0.06) | Cons. | 26.76*** (1.43) | Cons. | 0.39*** (0.04) |
| # of obs. | 93 | # of obs. | 93 | # of obs. | 91 | # of obs. | 91 |
| R-squared | 0.03 | R-squared | 0.06 | R-squared | 0.32 | R-squared | 0.29 |

b3: Codes related to why they chose/are choosing a certain behavior (Code Ds)

| ENDO (Code Ds) | | | | EXO (Code Ds) | | | |
|---|---|---|---|---|---|---|---|
| Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | | Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | |
| Model 1 | | Model 2 | | Model 3 | | Model 4 | |
| Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates |
| Code D1 | 19.24<br>(11.86) | Code D1 | 0.64**<br>(0.31) | Code D1 | 4.44*<br>(2.32) | Code D1 | 0.09<br>(0.10) |
| Code D2 | 4.13<br>(2.63) | Code D2 | -0.28***<br>(0.06) | Code D2 | 9.16***<br>(2.07) | Code D2 | 0.23***<br>(0.08) |
| Code D3 | -20.20***<br>(1.69) | Code D3 | -0.42***<br>(0.04) | Code D3 | -8.45**<br>(3.72) | Code D3 | -0.11<br>(0.09) |
| Code D4 | 7.95***<br>(2.42) | Code D4 | 0.27***<br>(0.06) | Code D5 | -11.07**<br>(4.71) | Code D5 | -0.13<br>(0.10) |
| Code D5 | -18.53***<br>(4.44) | Code D5 | -0.43***<br>(0.10) | Code D6 | -5.78<br>(4.24) | Code D6 | -0.12<br>(-0.12) |
| Code D6 | 3.98**<br>(1.69) | Code D6 | 0.07<br>(0.04) | Code D7 | -0.97<br>(2.24) | Code D7 | 0.06<br>(0.07) |
| Code D7 | 3.13<br>(2.18) | Code D7 | -0.01<br>(0.08) | Code D8 | -14.89***<br>(4.36) | Code D8 | -0.23**<br>(0.09) |
| Code D8 | 0.59<br>(7.80) | Code D8 | 0.03<br>(0.21) | Code D10 | -1.72<br>(2.74) | Code D10 | -0.22**<br>(0.09) |
| Code D9 | -19.21***<br>(4.76) | Code D9 | -0.37***<br>(0.12) | Code D11 | 9.25<br>(6.54) | Code D11 | 0.15<br>(0.17) |
| Code D10 | 11.06<br>(11.34) | Code D10 | 0.19<br>(0.29) | Cons | 24.51***<br>(1.28) | Cons | 0.36***<br>(0.04) |
| Code D11 | -7.74<br>(5.09) | Code D11 | -0.35***<br>(0.12) | # of obs. | 91 | # of obs. | 91 |
| Cons | 22.42***<br>(1.69) | Cons | 0.42***<br>(0.04) | R-squared | 0.29 | R-squared | 0.19 |
| # of obs. | 93 | # of obs. | 93 | | | | |
| R-squared | 0.14 | R-squared | 0.11 | | | | |

b4: Codes related to team behavior (Code Hs)

| ENDO (Code Hs) | | | | EXO (Code Hs) | | | |
|---|---|---|---|---|---|---|---|
| Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | | Dependent Var.: Work time (minutes) | | Dependent Var.: Work productivity | |
| Model 1 | | Model 2 | | Model 3 | | Model 4 | |
| Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates | Independent variable | Coefficient estimates |
| Code H1 | -5.16<br>(5.89) | Code H1 | -0.15<br>(0.10) | Code H1 | -1.87<br>(5.98) | Code H1 | -0.20**<br>(0.08) |
| Code H2 | 10.79**<br>(4.70) | Code H2 | 0.17*<br>(0.10) | Code H2 | 5.63<br>(6.58) | Code H2 | 0.00<br>(0.07) |
| Code H3 | 3.01<br>(3.38) | Code H3 | 0.02<br>(0.09) | Code H3 | 2.62<br>(2.09) | Code H3 | 0.03<br>(0.07) |
| Code H4 | 0.26<br>(6.48) | Code H4 | -0.02<br>(0.13) | Code H4 | -2.12<br>(4.41) | Code H4 | -0.12<br>(0.08) |
| Code H5 | 6.70**<br>(2.72) | Code H5 | 0.04<br>(0.09) | Code H5 | 3.64*<br>(2.03) | Code H5 | 0.09<br>(0.06) |
| Code H6 | 3.90<br>(5.88) | Code H6 | 0.25**<br>(0.12) | Code H6 | 2.76<br>(2.50) | Code H6 | 0.04<br>(0.10) |
| Code H7 | -5.09<br>(7.91) | Code H7 | -0.30*<br>(0.16) | Code H7 | -2.57<br>(1.60) | Code H7 | -0.06<br>(0.07) |
| Code H8 | -7.07<br>(4.98) | Code H8 | -0.13<br>(0.09) | Code H8 | -5.23<br>(3.32) | Code H8 | -0.09<br>(0.06) |
| Code H9 | -22.78***<br>(1.91) | Code H9 | -0.46***<br>(0.05) | Code H10 | -8.60*<br>(4.80) | Code H10 | -0.13*<br>(0.08) |
| Code H10 | -7.97<br>(4.91) | Code H10 | -0.14<br>(0.09) | Code H11 | 8.17***<br>(2.62) | Code H11 | 0.07<br>(0.07) |
| Code H11 | 10.19***<br>(1.95) | Code H11 | 0.26***<br>(0.08) | Code H12 | 4.67***<br>(1.62) | Code H12 | -0.17***<br>(0.06) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Code H12 | 10.95***<br>(1.99) | Code H12 | 0.30***<br>(0.05) | Code H13 | -2.85<br>(3.44) | Code H13 | -0.06<br>(0.08) |
| Code H13 | 2.07<br>(5.42) | Code H13 | -0.27**<br>(0.10) | Code H14 | 0.13<br>(2.36) | Code H14 | -0.10<br>(0.08) |
| Code H14 | -3.98<br>(3.08) | Code H14 | -0.17*<br>(0.10) | Cons. | 22.97***<br>(2.56) | Cons. | 0.43***<br>(0.06) |
| Cons. | 23.27***<br>(1.90) | Cons. | 0.46***<br>(0.05) | # of obs. | 91 | # of obs. | 91 |
| # of obs. | 93 | # of obs. | 93 | R-squared | 0.57 | R-squared | 0.50 |
| R-squared | 0.68 | R-squared | 0.56 | | | | |

*Notes*: Linear regressions with robust standard errors. The numbers in parentheses are standard errors. Work productivity is calculated as the number of correct answers divided by the duration to stay on the work site (minutes). Codes whose Kappa values are equal to or above 0.4 are used as independent variables. [#1] Code F10 was omitted despite the Kappa values being above 0.4, because only one chat dialogue was categorized as this code (making estimating its coefficient estimate impossible). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

# Appendix C: Appendix for Chapter 4

## Appendix C.A: Other Figures and Tables

**Table A.1:** *Socially and Privately Optimal Behaviour in Part 2*

| Civic Tasks (Total) | R% | Points Earned when $p_i$=8 (Ind.) | Marginal Benefit-Marginal Cost (Ind.) | Socially Optimal $p_i$* | Points Earned (Group) | Marginal Benefit-Marginal Cost (Group) |
|---|---|---|---|---|---|---|
| 0 | 50.0 | 154.3 | -0.3 | 8 | 1851 | - |
| 1 | 44.8 | 163.9 | -1.4 | 9 | 1968 | 106.9 |
| 2 | 40.4 | 172.5 | -2.4 | 10 | 2083 | 104.7 |
| 3 | 36.6 | 180.1 | -3.2 | 10 | 2196 | 102.8 |
| 4 | 33.3 | 187.0 | -3.8 | 11 | 2303 | 97.0 |
| 5 | 30.4 | 193.1 | -4.4 | 11 | 2401 | 88.0 |
| 6 | 27.7 | 198.7 | -4.9 | 11 | 2489 | 77.8 |
| 7 | 25.4 | 203.8 | -5.4 | 10 | 2546 | 47.5 |
| 8 | 23.3 | 208.4 | -5.8 | 10 | 2614 | 57.7 |
| 9 | 21.4 | 212.6 | -6.1 | 10 | 2675 | 51.5 |
| 10 | 19.7 | 216.5 | -6.4 | 10 | 2725 | 40.0 |
| 11 | 18.1 | 220.1 | -6.7 | 10 | 2748 | 12.8 |
| 12 | 16.6 | 223.4 | -6.9 | 10 | 2769 | 10.8 |
| 13 | 15.3 | 226.5 | -7.2 | 9 | 2808 | 28.9 |
| 14 | 14.1 | 229.4 | -7.3 | 9 | 2846 | 27.7 |
| 15 | 12.9 | 232.0 | -7.5 | 9 | 2881 | 25.0 |
| 16 | 11.9 | 234.5 | -7.7 | 9 | 2913 | 22.5 |
| 17 | 10.9 | 236.8 | -7.8 | 9 | 2939 | 16.3 |
| 18 | 10.0 | 239.0 | -8.0 | 9 | 2951 | 1.9 |
| 19 | 9.1 | 241.0 | -8.1 | 9 | 2962 | 1.1 |
| 20 | 8.3 | 242.9 | -8.2 | 9 | 2973 | 0.4 |
| 21 | 7.5 | 244.7 | -8.3 | 9 | 2983 | -0.2 |
| 22 | 6.8 | 246.4 | -8.4 | 9 | 2992 | -0.8 |
| 23 | 6.2 | 248.0 | -8.5 | 9 | 3001 | -1.3 |
| 24 | 5.5 | 249.5 | -8.6 | 9 | 3009 | -1.8 |
| 25 | 4.9 | 250.9 | -8.6 | 9 | 3016 | -2.3 |
| 26 | 4.4 | 252.3 | -8.7 | 8 | 3027 | 0.9 |
| 27 | 3.8 | 253.6 | -8.8 | 8 | 3043 | 5.5 |
| 28 | 3.3 | 254.8 | -8.8 | 8 | 3058 | 4.7 |
| 29 | 2.8 | 256.0 | -8.9 | 8 | 3072 | 4.0 |
| 30 | 2.3 | 257.1 | -8.9 | 8 | 3085 | 3.3 |
| 31 | 1.9 | 258.1 | -9.0 | 8 | 3098 | 2.7 |
| 32 | 1.5 | 259.2 | -9.0 | 8 | 3110 | 2.2 |
| 33 | 1.1 | 260.1 | -9.1 | 8 | 3121 | 1.6 |
| 34 | 0.7 | 261.0 | -9.1 | 8 | 3133 | 1.1 |
| 35 | 0.3 | 261.9 | -9.2 | 8 | 3143 | 0.7 |
| 36 | 0.0 | 262.7 | -10.0 | 8 | 3152 | -0.8 |
| 37 | 0.0 | 262.7 | -10.0 | 8 | 3152 | -10.0 |

*Notes for Table A.1:* This table shows how the total number of civic tasks completed (column 1) impacts the level of loss experienced in the mainstage (column 2). Columns 3 and 4 demonstrate that it is never privately optimal for a subject to complete a civic task. Column 3 shows the points earned for an individual that contributes 8 tokens (the privately optimal level in Part 2) to the public sector for a given level of loss, R%. Column 4 shows the difference in earnings from an individual completing one civic task minus the opportunity cost of completing a private task (10 points), for the respective number of civic task completion by others in their group. As shown, it is never privately

optimal to complete a civic task, regardless as to other group members' civic task completion, as the marginal benefit of doing so is never positive.

The values in columns 5, 6, and 7 may be used to identify the socially optimal level of civic task completion by the group. The socially optimal level of contribution, p*, is dictated by the amount that would maximise group earnings for a given level of loss. Column 6 shows the total points earned by a group, assuming each group member contributes the socially optimal level of tokens to the public sector (Column 5). As shown, the group's total payoff from the mainstage increases as loss is reduced until percentage loss reaches 0.0%, where total payoff remains stable at 3152 points. Column 7 shows the difference in total group earnings by completing one more civic task (the marginal benefit) minus the opportunity cost of completing a private task (the marginal cost of 10 points). The cost of completing another task exceeds the benefit between 20 and 24 civic tasks, and from 35 civic tasks onwards. Given the social optimum is concerned with maximising group earnings, and earnings are greater at 35 civic tasks than at any point between 20 and 24 civic tasks, the socially optimal number of civic tasks is 35(or 2.92 tasks on average); this is the point at which a group can maximise their total earnings when accounting for the opportunity cost of forgoing a private task. Owing to the shape of the D(P) and V(P) curves, given in the experimental design section, 8 is the privately and socially optimal contribution, for both the privately and socially optimal level of civic task completion.

**Table A.2:** *Allocations to the Public Sector in Part 1 and Part 2*

Dependent variable: Allocations to the public sector in period t

| | Part 1 | | | Part 2 | | |
|---|---|---|---|---|---|---|
| | (1a) | (1b) | (1c) | (2a) | (2b) | (2c) |
| (a) Neutral-With Feedback dummy | 0.50<br>(0.51) | 0.50<br>(0.42) | -0.19<br>(0.76) | 0.07*<br>(0.04) | 0.07<br>(0.05) | 0.13<br>(0.08) |
| (b) Political-No Feedback dummy | -0.51<br>(0.57) | -0.51<br>(0.51) | -1.92**<br>(0.79) | -0.01<br>(0.09) | -0.01<br>(0.08) | 0.04<br>(0.11) |
| (c) Political-With Feedback dummy | -0.25<br>(0.49) | -0.23<br>(0.53) | -0.98<br>(0.82) | 0.03<br>(0.04) | 0.03<br>(0.04) | 0.13<br>(0.12) |
| (d) Period {=1 to 4}{=5-19) | --- | -1.75***<br>(0.09) | -2.04***<br>(0.19) | --- | -0.02***<br>(0.00) | -0.01**<br>(0.00) |
| (e) Interaction: (a) × Period | --- | --- | 0.29<br>(0.26) | --- | --- | -0.01<br>(0.01) |
| (f) Interaction: (b) × Period | --- | --- | 0.59**<br>(0.29) | --- | --- | -0.01<br>(0.01) |
| (g) Interaction: (c) × Period | --- | --- | 0.31<br>(0.23) | --- | --- | -0.01<br>(0.01) |
| Constant | 2.94***<br>(0.35) | 7.31***<br>(0.36) | 8.02***<br>(0.56) | 8.01***<br>(0.03) | 8.14***<br>(0.04) | 8.09***<br>(0.05) |
| # of observations | 1,536 | 1,536 | 1,536 | 5,760 | 5,760 | 5,760 |
| Control | Yes | Yes | Yes | Yes | Yes | Yes |
| Wald $\chi 2$ | 11.18 | 395.07 | 351.85 | -7403.11 | 43.44 | 42.31 |
| Prob > Wald $\chi 2$ | 0.025 | 0.000 | 0.000 | 0.109 | 0.000 | 0.000 |
| $H_0$: (a) = (b) | 0.089* | 0.082* | 0.030** | 0.350 | 0.320 | 0.516 |
| $H_0$: (a) = (c) | 0.185 | 0.224 | 0.331 | 0.333 | 0.375 | 0.933 |
| $H_0$: (b) = (c) | 0.627 | 0.644 | 0.241 | 0.636 | 0.656 | 0.553 |

Note: Individual-level random effect Tobit regressions. Number in parenthesis are bootstrapped standard errors. Group average data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table A.3:** *Average Allocations to the Public Sector in Part 2 by Half*

| | Avg. Allocation to the Public Sector | | | | |
|---|---|---|---|---|---|
| Treatment | (i) Periods 1-7 (Part 2) | (ii) Periods 8-15 (Part 2) | $H_0$: (i) = 8 | $H_0$: (ii) = 8 | $H_0$: (i) = (ii) |
| [Individual treatments:] | | | | | |
| (a) Neutral-No Feedback | 8.09 | 8.03 | 0.0015*** | 0.1173 | 0.0499** |
| (b) Neutral-With Feedback | 8.18 | 8.09 | 0.0089*** | 0.1460 | 0.0499** |
| (c) Political-No Feedback | 8.12 | 7.98 | 0.2223 | 0.8717 | 0.0251** |
| (d) Political-With Feedback | 8.16 | 8.02 | 0.0818* | 0.5410 | 0.0797* |
| [Across-treatment comparisons:] | | | | | |
| *p* for $H_0$: (a) = (b) | 0.1703 | 0.6657 | --- | --- | --- |
| *p* for $H_0$: (a) = (c) | 0.1706 | 0.3138 | --- | --- | --- |
| *p* for $H_0$: (a) = (d) | 0.7518 | 0.5562 | --- | --- | --- |
| *p* for $H_0$: (b) = (c) | 1.0000 | 0.7513 | --- | --- | --- |
| *p* for $H_0$: (b) = (d) | 0.5615 | 0.3108 | --- | --- | --- |
| *p* for $H_0$: (c) = (d) | 0.5982 | 0.2680 | --- | --- | --- |

*Notes*: All *p*-values are based on two-sided tests. Wilcoxon signed rank (Mann-Whitney) tests were conducted for within(across)-treatments comparisons in columns 1, 2, and 4, and single-sample t-tests were used in columns 2 and 3, using group means of contributions to the public sector. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table A.4:** *Average Payments by Treatment*

| Treatment | Av. Payment |
|---|---|
| [Individual treatments:] | |
| (a) Neutral-No Feedback | £21.12 |
| (b) Neutral-With Feedback | £21.81 |
| (c) Political-No Feedback | £21.00 |
| (d) Political-With Feedback | £21.46 |
| [Across-treatment comparisons:] | |
| *p* for $H_0$: (a) = (b) | 0.0156** |
| *p* for $H_0$: (a) = (c) | 0.7525 |
| *p* for $H_0$: (a) = (d) | 0.1719 |
| *p* for $H_0$: (b) = (c) | 0.0519* |
| *p* for $H_0$: (b) = (d) | 0.2076 |
| *p* for $H_0$: (c) = (d) | 0.3446 |

*Notes*: All *p*-values are based on two-sided tests. Mann-Whitney tests were conducted for across-treatment comparisons, using group means of payment received. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table A.5:** *Average Civic Task Completion per Person per Period, by Experiment Half*

Dependent variable: The average number of civic tasks correctly completed per person per period

| | Periods 1-7 of Part 2 | | | Periods 8-15 of Part 2 | | |
|---|---|---|---|---|---|---|
| | (1a) | (2a) | (3a) | (1b) | (2b) | (3b) |
| (a) Neutral-With Feedback dummy | 0.64*** (0.17) | 0.64*** (0.17) | 0.35 (0.22) | 0.84*** (0.15) | 0.84*** (0.15) | 1.07*** (0.21) |
| (b) Political-No Feedback dummy | 0.15 (0.18) | 0.15 (0.18) | 0.10 (0.23) | 0.20 (0.13) | 0.20 (0.13) | 0.22 (0.30) |
| (c) Political-With Feedback dummy | 0.44*** (0.19) | 0.44** (0.19) | 0.06 (0.23) | 0.70*** (0.15) | 0.70*** (0.15) | 0.84*** (0.20) |
| (d) Period {=1 to 15} | --- | 0.04*** (0.01) | -0.01 (0.02) | --- | -0.04*** (0.01) | -0.03** (0.01) |
| (e) Interaction: (a) × Period | --- | --- | 0.07** (0.03) | --- | --- | -0.02 (0.02) |
| (f) Interaction: (b) × Period | --- | --- | 0.01 (0.03) | --- | --- | 0.00 (0.02) |
| (g) Interaction: (c) × Period | --- | --- | 0.09*** (0.03) | --- | --- | -0.01 (0.02) |
| Constant | 1.37*** (0.16) | 1.21*** (0.18) | 1.39*** (0.20) | 1.05*** (0.11) | 1.49*** (0.13) | 1.39*** (0.19) |
| # of observations | 224 | 224 | 224 | 256 | 256 | 256 |
| Control | Yes | Yes | Yes | Yes | Yes | Yes |
| Wald $\chi2$ | 26.96 | 32.71 | 54.01 | 49.34 | 101.32 | 146.41 |
| Prob > Wald $\chi2$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| [P-Value for Wald $\chi2$ tests of coefficient differences]: | | | | | | |
| $H_0$: (a) = (b) | 0.001*** | 0.001*** | 0.1416 | 0.000*** | 0.000*** | 0.002*** |
| $H_0$: (a) = (c) | 0.214 | 0.215 | 0.096* | 0.425 | 0.426 | 0.150 |
| $H_0$: (b) = (c) | 0.090* | 0.091* | 0.8456 | 0.001*** | 0.001*** | 0.019** |
| $H_0$: (d) + (e) = 0 | --- | --- | 0.001*** | --- | --- | 0.000*** |
| $H_0$: (d) + (f) = 0 | --- | --- | 0.765 | --- | --- | 0.108 |
| $H_0$: (d) + (g) = 0 | --- | --- | 0.000*** | --- | --- | 0.001*** |

Note: Group-level random effect linear regressions. Number in parenthesis are robust standard errors. Group average data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table A.6:** *All Task Completion in Part 2*

Dependent variable: The average number of civic and private tasks correctly completed per person

| | Periods 1 to 15 of Part 2 | | |
|---|---|---|---|
| | (1a) | (1b) | (1c) |
| (a) Neutral-With Feedback dummy | -0.01<br>(0.12) | -0.01<br>(0.12) | 0.06<br>(0.21) |
| (b) Political-No Feedback dummy | -0.13<br>(0.12) | -0.13<br>(0.13) | -0.10<br>(0.21) |
| (c) Political-With Feedback dummy | -0.06<br>(0.12) | -0.06<br>(0.12) | -0.05<br>(0.21) |
| (d) Period {=1-15} | --- | 0.11***<br>(0.00) | 0.11***<br>(0.01) |
| (e) Interaction: (a) × Period | --- | --- | -0.01<br>(0.01) |
| (f) Interaction: (b) × Period | --- | --- | 0.00<br>(0.01) |
| (g) Interaction: (c) × Period | --- | --- | 0.00<br>(0.02) |
| Constant | 4.02***<br>(0.09) | 3.14***<br>(0.11) | 3.11***<br>(0.18) |
| # of observations | 480 | 480 | 480 |
| Control | Yes | Yes | Yes |
| Wald $\chi 2$ | 3.51 | 783.83 | 1204.57 |
| Prob > Wald $\chi 2$ | 0.476 | 0.000 | 0.000 |
| $H_0$: (a) = (b) | 0.190 | 0.190 | 0.241 |
| $H_0$: (a) = (c) | 0.560 | 0.560 | 0.399 |
| $H_0$: (b) = (c) | 0.434 | 0.434 | 0.764 |

Note: Group-level random effect linear regressions. Number in parenthesis are robust standard errors. Group average data are used. The reference treatment is the neutrally-framed no-feedback treatment (Neutral-No Feedback). 'Control' indicates the inclusion of a control dummy for laboratory location. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table A.7:** *Two-Sample Kolmogorov-Smirnov Tests for Average Correctly Completed Civic Tasks*

| Av. Civic Tasks Correctly Completed per Person | Neutral-No Feedback | Neutral-With Feedback | Political-No Feedback | Political-With Feedback |
|---|---|---|---|---|
| Neutral-No Feedback | --- | --- | --- | --- |
| Neutral-With Feedback | 0.0000*** | --- | --- | --- |
| Political-No Feedback | 0.2590 | 0.0030*** | --- | --- |
| Political-With Feedback | 0.0030*** | 0.4410 | 0.0080*** | --- |

Note: Values are P-Values for subject-level two-way Kolmogorov-Smirnov tests. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

**Table A.8:** *Relative Feedback Received as a Function of Relative Civic Tasks Completed*

Dependent variable: The relative feedback received by subject i in period t

| | (1) |
|---|---|
| (a) Relative Civic Task Completion $\{t_{c,i,t} - t_{c,sc\text{-}i}\}$ | 0.54*** |
| | (0.03) |
| (b) Constant | 1.18*** |
| | (0.00) |
| # of observations | 2,813 |
| F | 236.71 |
| Prob > F | 0.000 |

Note: Individual-level fixed effect linear regression. Numbers in parenthesis are robust standard errors clustered by group. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

## Appendix C.B: Sample Instructions

**[Instructions for Part 1 (Politically Framed)]**

**INSTRUCTIONS FOR PART 1**

This experiment involves a set of decisions by 12 participants, yourself included, in which others' decisions can affect your earnings, and your decisions can affect their earnings. Whenever you are shown feedback on the decisions of others, their real identities will be kept anonymous, but please be assured that reported decisions are those of the same actual participants (group composition does not change) and never fictitious participants simulated by a computer program or members of the experimenter team.

No communication between participants will be permitted during the experiment. You are also not permitted to use your phone, tablet computer, or programs other than the designated experiment software. Members of the experiment team will check that this rule is adhered to. You will have an opportunity to ask questions before the experiment begins. We ask that you devote your full attention to the experiment while it is in progress.

In the experiment, we'll be using two different currencies. The first currency, called **tokens**, is something you are given each period to allocate as you wish in order to earn the second currency, called **points**. Throughout the experiment, you can try to accumulate points. At the end of the experiment, your points will be converted to money (pounds) at a rate of 260 points to £1. You will receive your payment in cash at the end of the experiment. As you'll see below, while the value of a point is small, your total earnings can still be substantial. Please listen carefully to the instructions and ask questions if something is unclear.

Decisions and earnings

The main decision to be made, and the main way in which you can earn points, involves the allocation of your tokens between a private income-generating activity and a public sector. Allocating tokens to your private activity is always beneficial to you, but the size of the benefit is larger when the public sector is well funded. The amount jointly allocated to the public sector also determines a direct benefit evenly distributed across each participant, regardless of what they allocated to the public sector individually, similar to the benefits in everyday life from having safe roads, law and order, and clean air. Each participant has a private activity of their own, whereas there is only one public sector for the whole group. We will now provide further details about the allocation decision between the public sector and private activity.

More about the main allocation problem

In each period, you and every other participant will be endowed with 20 tokens that you must decide how to allocate between two accounts, your private activity, and the public sector. As mentioned above, each participant has their own private activity, while there is a single public sector for all 12 participants in a group. In a period, you can assign any integer number of tokens (including zero) to the public sector, assigning the rest of that period's 20 tokens to your private activity. Examples include 0 to the public sector and 20 to your private activity; 7 to the public sector and 13 to your private activity; 14 to the public sector and 6 to the private activity, and so on. These are among the twenty-one possible ways you can allocate your twenty tokens. Each of you makes an allocation decision with your own 20 tokens separately and simultaneously, learning of the others' decisions afterwards.

The number of points you earn from tokens allocated to your private activity depends on the number of tokens put into the public sector in that period by you and the other 11 participants. Call the

number of tokens you put into the private activity *b* (for "business") and the number you allocate to the group account *p* (for "public"). Since you always start with 20 tokens, $b + p = 20$. We'll call the sum of *p*'s allocated to the public sector by all 12 participants *P*.

The points you get from each token you allocate to your private activity—i.e., *b* —depends on *P*. Each token of *b* increases your earnings by 6 points when $P = 0$, and by a larger number of points, rising to a maximum of 18 points per token when $P = 96$ or more. See Table 1 and Figure 1.

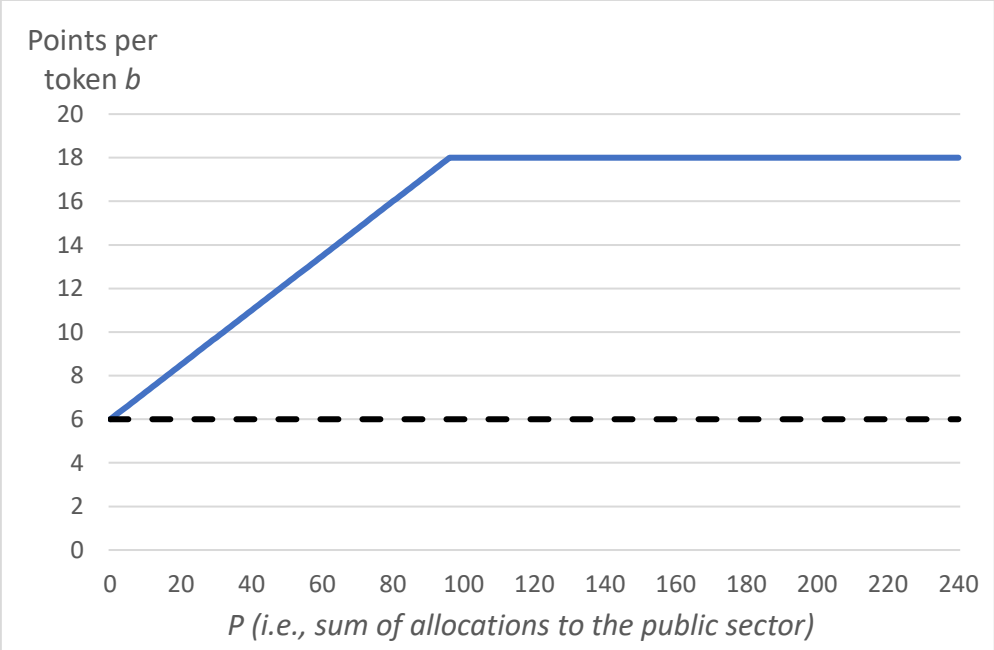



Figure 1. Points earned per token of *b* as a function of *P*

| *P* | 0 | 8 | 16 | 24 | 32 | 40 | 48 | 56 | 64 | 72 | 80 | 88 | 96 | Above 96 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| income from *b* | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 18 |

Table 1: income per token of *b* as a function of *P*

In addition to *P*'s effect on your earnings by influencing the income from tokens assigned to your private activity, *P* also affects your earnings in a direct way which is the same for all participants. Each participant in the experiment receives a number of points that rises as *P* does, and that goes equally to participants regardless of their individual choices of *b* and *p*. We will call this the "General Benefit". This general benefit of *P* rises as *P* increases, continuing to rise, although more slowly, even when $P > 96$, as shown in the figure below.

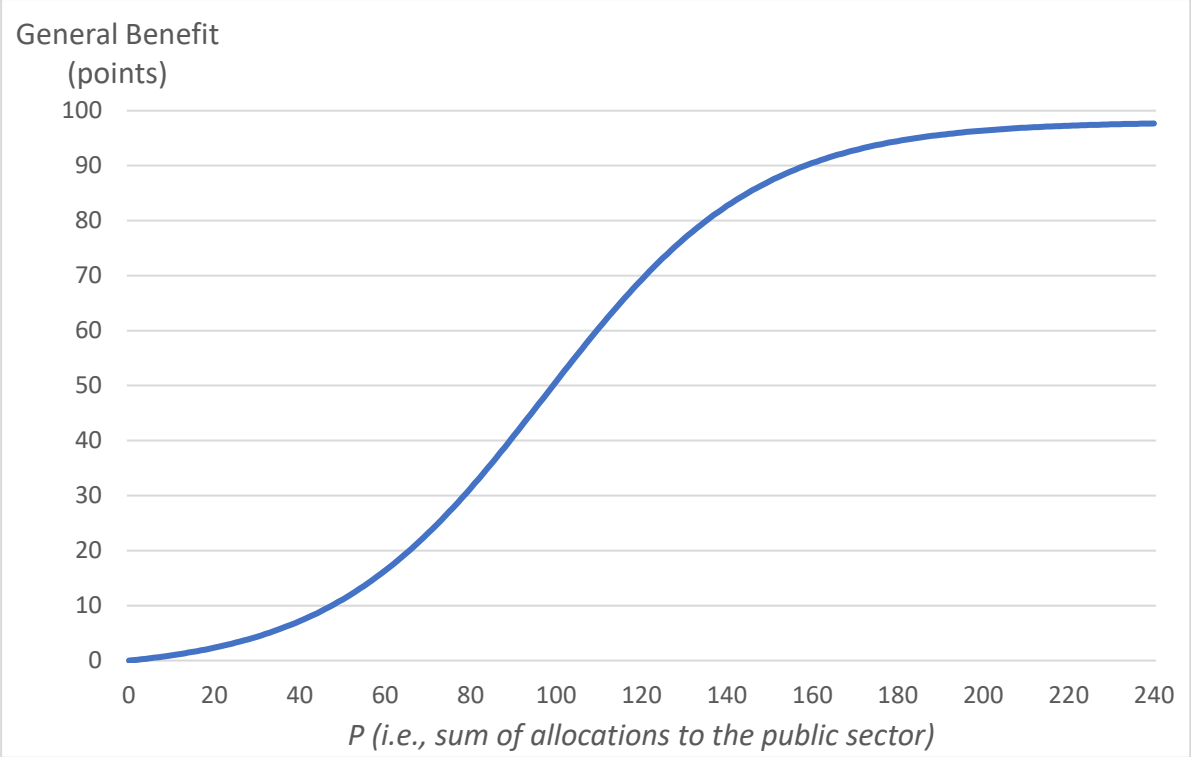



Figure 2. shows the benefit from the public sector *P* that is given to each participant, regardless of their *b* and *p* token allocation.

The two ways in which allocations to the public account affect earnings—partly through increasing the returns to any token allocated to one's private activity, and partly by yielding an equal amount for all participants—are summarized in Table 2. The columns correspond to different allocations of tokens to the public sector by you, and the rows correspond to different average allocations of tokens to the public sector by the other 11 participants. To make the presentation more compact, the table shows only one's own and others' average allocations that are divisible by four.

| *Average allocation of 11 others* | *Own allocation to the public sector (p)* | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 4 | 8 | 12 | 16 | 20 |
| 0 | 120 | 104 | 85 | 61 | 34 | 2 |
| 4 | 239 | 202 | 162 | 118 | 70 | 19 |
| 8 | 379 | 323 | 263 | 195 | 127 | 59 |
| 12 | 438 | 368 | 299 | 229 | 158 | 88 |
| 16 | 454 | 382 | 311 | 239 | 168 | 96 |
| 20 | 457 | 385 | 313 | 242 | 170 | 98 |

Table 2: Earnings as a function of your allocation to the public sector (*p*) and the average allocation *p* of the other 11 participants to the public sector

We've shaded the diagonal entries of the table, which represent situations in which you and the others in your group happen to allocate the same number of tokens (or for the others, the same number on average) to the public sector. For example, the entry *202* (second row from top, second column from left) is the total amount that you would earn if you allocated 4 of your 20 tokens to the public sector and 16 of your tokens to your private activity, while the other 11 participants allocate an average of 4 tokens each to the public sector. Notice that among these shaded diagonal cells, your earnings would be highest when you and the others on average allocate 8 tokens to the public sector, giving you 263

points. That's more than double your earnings if all participants put 0 into the public sector, and the fact that it occurs when all allocate 8 tokens to the public sector is consistent with the fact that the return from allocating a token to your private activity reaches its maximum value when $P = 96$ (= 12 x 8) (see Figure 1), and that the General Benefit of $P$ (shown in Figure 2) increases at a slower rate after $P = 96$. Table 2 is available on your screen during the allocation stages of the experiment by double-clicking the 'Payment Table' button. You can also open (and close) an expanded table showing outcomes for all integer combinations of allocations by yourself and others by clicking the 'Full Table' button, which becomes available when the smaller table is open.

Two further things to note are the following. First, your earnings are not sensitive to *how* others' allocations add up to a given average; any combination of choices by others that generates a given average has the same impact on your earnings. Second, what you earn does change if your own allocation varies, taking the average allocation of the others as given. For example, suppose that the others allocate an average of 8 tokens to the public sector. You earn more by allocating less than 8 yourself, as shown by the cells to the left of the one with the shaded value of 260. The largest number in the table, 457, is what you would earn if others assigned all their tokens to the public sector, while you allocate all of yours to your private activity.

In summary, there will be four periods in Part 1 of the experiment followed by a break for further instructions. Operationally, each of the 4 periods in Part 1 will unfold as follows:

- You'll initially see a screen where you'll be asked to decide how many (if any) of the 20 tokens you wish to allocate to the public sector (the rest automatically go to your private activity).
- When everyone has submitted their decisions, you'll see a screen showing your overall results for the period.
- When you click "Next", you'll see a screen showing the amount that you and each of the other 11 participants assigned to the public sector in this period, plus the points that each of you earned. These results will be anonymous; you will only see the tokens allocated and the corresponding points earned.
- You can take a moment to absorb this information, then click "Next" to begin the next period.

**[Instructions for Part 2 (Politically Framed)]**

**INSTRUCTIONS FOR PART 2**

The remaining fifteen periods of the experiment have a core structure identical to those of the first four periods. In what we'll now call the "main stage" of each period, you and the other 11 participants each have 20 tokens to allocate between your private activity and the public sector. However, whereas the allocation decision was strictly voluntary in Part 1, there will now be **a government** that makes allocating a minimum number of tokens to the public sector a requirement, subject to a penalty if not fulfilled. The allocation to the public sector that is required to avoid a penalty will be 8 of your 20 tokens, which, as you will recall, was the allocation (among those in which all allocated equally) at which total earnings of participants were maximized in Part 1. For each token less than 8 that you allocate to the public sector, you will be penalized 35 points. The size of the penalty is large so you will definitely earn less if you allocate less than 8 tokens to the public sector (see Table 3 below, where the struck through amounts indicate points earned before the penalty has been applied).

| Average *p* of 11 others | Own allocation to the public sector | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 4 | 8 | 12 | 16 | 20 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | ~~120~~<br>-160 | ~~104~~<br>-36 | 85 | 61 | 34 | 2 |
| 4 | ~~239~~<br>-41 | ~~202~~<br>62 | 162 | 118 | 70 | 19 |
| 8 | ~~379~~<br>99 | ~~323~~<br>183 | 263 | 195 | 127 | 59 |
| 12 | ~~438~~<br>158 | ~~368~~<br>228 | 299 | 229 | 158 | 88 |
| 16 | ~~454~~<br>174 | ~~382~~<br>242 | 311 | 239 | 168 | 96 |
| 20 | ~~457~~<br>177 | ~~385~~<br>245 | 313 | 242 | 170 | 98 |

Table 3: Earnings as a function of your allocation to the public sector (*p*) and the average allocation *p* of the other 11 participants to the public sector when there is a **minimum required allocation** of 8 tokens

In addition to having a minimum required allocation, a further change may also affect the total amount allocated to the public sector, *P*, in Part 2. Although having a government to enforce a penalty scheme can increase the amounts citizens allocate to the public sector, potentially increasing earnings, real-world governments sometimes have leaders and officials that don't act fully in the public's interest. Indeed, some government revenue can be lost to lax oversight, negligence, or corruption by government officials. To capture this point, the tokens in *P* may be reduced by a percentage, which we will call *R*%, that varies depending on your own and others' actions. Tokens that are removed from the public sector by this reduction process will not be used in the calculation of the general benefit received by everyone and won't help to increase your return from allocating tokens to your private activity. Given this, *P* can now be re-defined as the total amount of tokens allocated to the public sector minus any reductions due to corruption or waste by officials. We will explain how the percentage that *P* is reduced by is determined next.

In what follows, we assume that governments exhibit less corruption when citizens engage more in public affairs. Examples of **civic engagement** in the real world include paying attention to the news, voting in elections, participating in a campaign or rally, signing a petition, or other actions that may hold a government to account.

Each of the fifteen periods remaining will include an extra stage before the main stage—we'll call it the "pre-stage"—during which you'll have the opportunity to perform two types of tasks. The first type of task, called a **"Civic Task",** decreases the amount *P* is reduced by in the period's main stage. Put differently, the more civic tasks that are completed in a period's pre-stage, the smaller the percentage (*R*%) by which *P* gets reduced. The way in which *R*% decreases as you and others increase the number of civic tasks completed overall is shown in the graph below. *R*% starts at 50% when no civic tasks are completed; this means that the value of *P* is reduced by 50% before the general benefit and your private return are calculated in the main stage. Completing civic tasks reduces *R*%, for example, if an average of two civic tasks are completed by you and the other participants (a total of 24 civic tasks), then *R*% falls from 50% to 5.2%. This means that only 5.2% of tokens are removed from the total amount allocated to the public sector before your earnings are calculated. If 36 or more civic tasks are completed, no tokens are removed from the total put into the public sector —i.e., *R*% = 0.

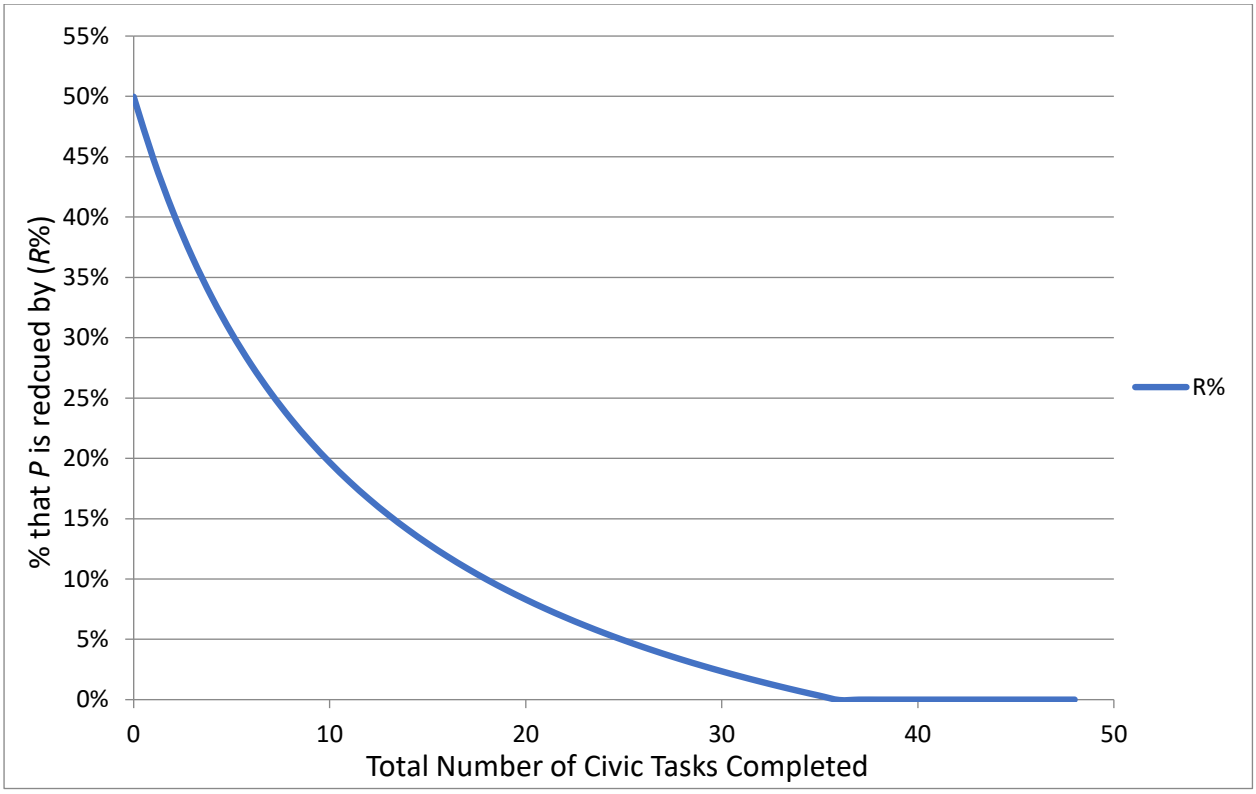



Figure 3. shows the percentage (*R*%) that *P* is reduced by, due to corruption or waste, for a given level of civic tasks completed in total.

To give you an idea how the percentage reduction (*R*%) affects your earnings, the payment table available on your screen has been updated with a slider. You can adjust the slider for hypothetical numbers of civic tasks completed by all 12 participants and see the corresponding *R*% and payment table (which is read in the same way as Table 2 in Part 1). The table also accounts for the penalty which is applied if fewer than 8 tokens are allocated to the public sector. As in Part 1, you can view an expanded table showing outcomes for all integer combinations of allocations by yourself and others by clicking the 'Full Table' button. Please take a moment to open the table and use the slider to see how *R*% affects the number of points you earn depending on the tokens you and the other 11 participants allocate to the public sector.

The second type of activity available during the pre-stage is **"Private Tasks"**. Completing a private task correctly adds 10 points directly to your earnings and has no effect on *R*%. Tasks of both types take about 10 - 15 seconds to complete, and a total of 40 seconds will be available each period for the task portion of the pre-stage. Any points you earn in the pre-stage are added to your overall accumulation and they convert to real money at the same rate as other points at the end of the experiment. The potential to earn points in a period's pre-stage does not affect what allocations you can make in its main stage. You will have 20 tokens available to allocate to the public sector and your private activity in the period's main stage, regardless of how many tasks you complete.

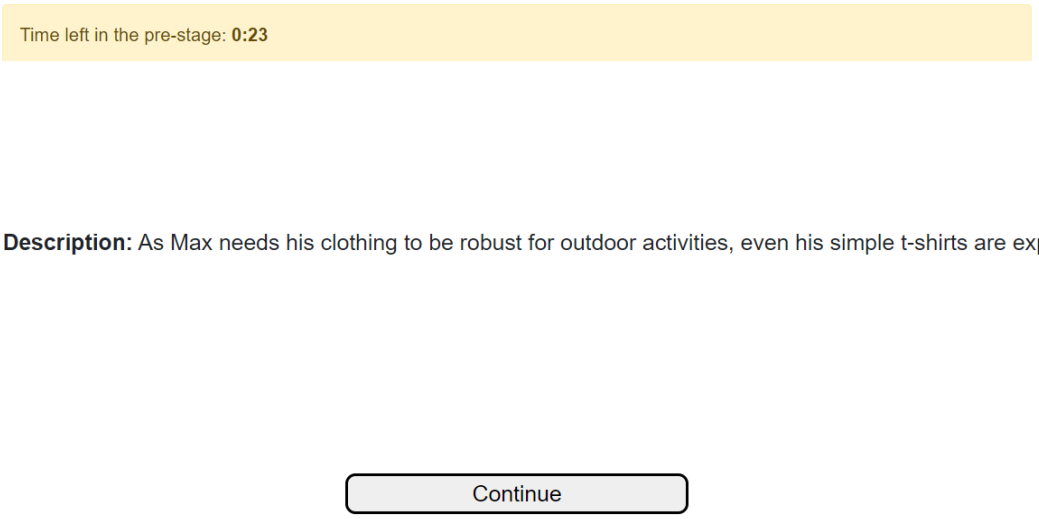
Information sharing and feedback. {Treatments with Feedback Only}

In the real world, you might wish to share with others the fact that you registered to vote, went to the polls, read up on candidates' positions, or took part in some other civic activity. Sharing with others information about your completion of civic tasks is also possible in the experiment. At the end of each period's pre-stage, information about the number of civic tasks that you and 3 other randomly chosen participants have completed will be displayed, along with their identification letter (A, B, C, or D). The composition of this set of four participants remains fixed for the remainder of Part 2, and will

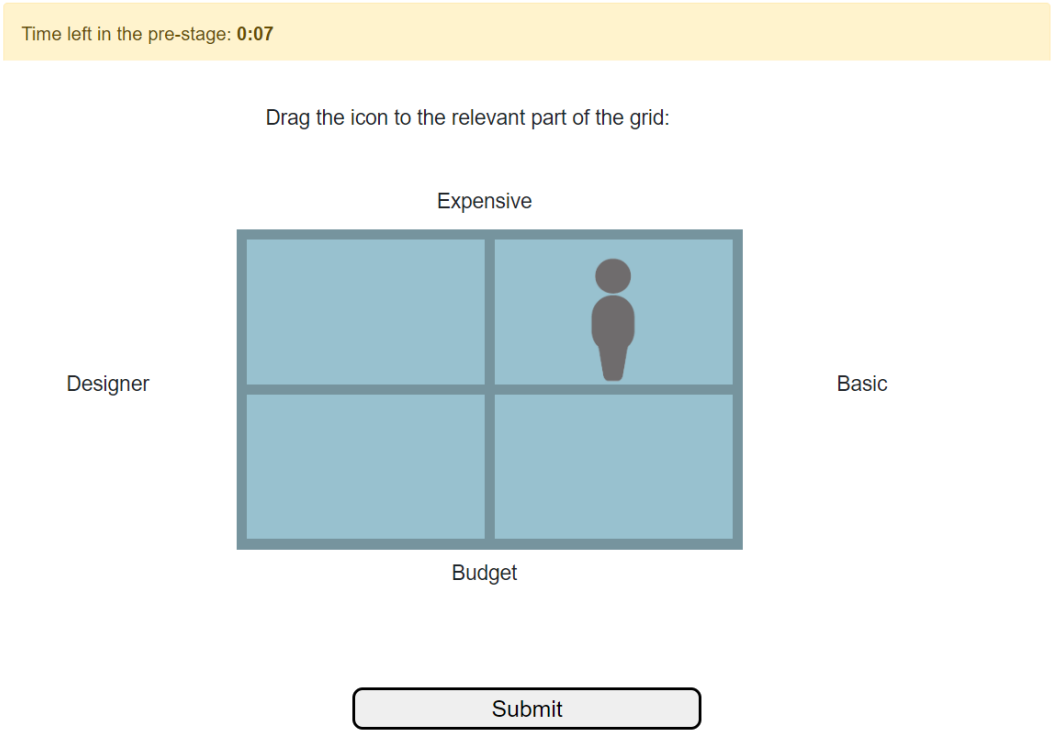
be referred to as your **'social circle'**. The pre-stage of each period will end with an opportunity to provide feedback to the others in your social circle, and for them to do the same to you anonymously. Specifically, you can give a smiley face (☺) to any or none of them. On the final pre-stage screen, you'll be shown the feedback other social circle members submitted about you (in total), as well as the feedback that the other social circle members received. {/}

More about pre-stage tasks.

When a Part 2 period begins, always with its pre-stage, you'll see a screen on which you select whether the first task you want to do will be a civic or private task. Once you click on your choice, you'll begin that task. The tasks are identical in nature, only how they impact the main stage differs (as described above). Each task begins with a description of a person differing in two dimensions or characteristics (see screen image 1.a below), for example, what type of food they like and whether they prefer to cook or eat in a restaurant. After reading the description and clicking continue, you'll see a two-dimensional grid (screen image 1.b). There, you'll click and drag a person-shaped icon to whichever of the four quadrants corresponds to the description, drop it in place, and submit that answer by clicking the "Submit" button. Note that you cannot go back from the grid screen to view the description, although you are free to take notes to help you remember it. Once you have submitted an answer, you will be told whether it was correct or not, and then click 'Return' to select the next task type.




screen image 1.a




screen image 1.b

On the screen showing the description of the individual, the experiment software requires you to spend a minimum of 3 seconds before you can continue to the screen showing the four-quadrant grid, the screen with the grid further requires you to spend a minimum of 2 seconds before submitting your answer. This time requirement is to encourage you to pay attention to the tasks, rather than engage in random clicking.

Summary

As mentioned, the task part of the pre-stage will last for a total of 40 seconds. In that time, you may complete civic tasks, which reduce *R*% (representing loss from corruption or waste), or private tasks, which add to your personal earnings without reducing *R*%. When that time runs out, you'll be informed of the total number of civic tasks completed (from all 12 participants, combined) and of the resulting percentage reduction (*R*%) that will exist in the main stage as a result of this, as well as any earnings from private tasks.

{Treatments with Feedback Only} On the next screen, you will be shown the number of civic tasks completed by each of the other 3 members in your 4-person social circle, as well as their identification letters. You can then assign feedback (or not) to each of the other members in your social circle by clicking the checkbox in their row. On the screen after, you will see your own feedback from the other social circle members, as well as the feedback each other member received. You will not see who you or the other members received feedback from in your social circle. {/}

When you click continue, you'll go to the main stage (where you can allocate 20 tokens between the public sector or your private activity). The main stage will work as in Part 1 except that there is a penalty if you put less than 8 tokens into the public sector and the total amount allocated to the public sector may be reduced by *R*%, which varies depending on the number of civic tasks completed by all 12 participants in the pre-stage. The more civic tasks completed in the pre-stage, the less the amount in the public sector will be reduced.

**[Instructions for Part 1 (Neutrally Framed)]**

## INSTRUCTIONS FOR PART 1

This experiment involves a set of decisions by 12 participants, yourself included, in which others' decisions can affect your earnings, and your decisions can affect their earnings. Whenever you are shown feedback on the decisions of others, their real identities will be kept anonymous, but please be assured that reported decisions are those of the same actual participants (group composition does not change) and never fictitious participants simulated by a computer program or members of the experimenter team.

No communication between participants will be permitted during the experiment. You are also not permitted to use your phone, tablet computer, or programs other than the designated experiment software. Members of the experiment team will check that this rule is adhered to. You will have an opportunity to ask questions before the experiment begins. We ask that you devote your full attention to the experiment while it is in progress.

In the experiment, we'll be using two different currencies. The first currency, called **tokens**, is something you are given each period to allocate as you wish in order to earn the second currency, called **points**. Throughout the experiment, you can try to accumulate points. At the end of the experiment, your points will be converted to money (pounds) at a rate of 260 points to £1. You will receive your payment in cash at the end of the experiment. As you'll see below, while the value of a point is small, your total earnings can still be substantial. Please listen carefully to the instructions and ask questions if something is unclear.

Decisions and earnings

The main decision to be made, and the main way in which you can earn points, involves the allocation of your tokens between your private account and a group account. Allocating tokens to your private account is always beneficial to you, but the size of the benefit is larger when the group account is well funded. The amount jointly allocated to the group account also determines a direct payment evenly distributed across each participant, regardless of what they allocated to the group account individually. Each participant has a private account of their own, whereas there is only one group account for the whole group. We will now provide further details about the allocation decision between the group and private account.

More about the main allocation problem

In each period, you and every other participant will be endowed with 20 tokens that you must decide how to allocate between two accounts, your private account, and the group account. As mentioned above, each participant has their own private account, while there is a single group account for all 12 participants in a group. In a period, you can assign any integer number of tokens (including zero) to the group account, assigning the rest of that period's 20 tokens to your private account. Examples include 0 to the group account and 20 to your private account; 7 to the group account and 13 to your private account; 14 to the group account and 6 to the private account, and so on. These are among the twenty-one possible ways you can allocate your twenty tokens. Each of you makes an allocation decision with your own 20 tokens separately and simultaneously, learning of the others' decisions afterwards.

The number of points you earn from tokens allocated to your private account depends on the number of tokens put into the group account in that period by you and the other 11 participants. Call the number of tokens you put into the private account $p$ (for "private") and the number you allocate to the group account $g$ (for "group"). Since you always start with 20 tokens, $p + g = 20$. We'll call the sum of $g$'s allocated to the group account by all 12 participants $G$.

The points you get from each token you allocate to your private account—i.e., *p* —depends on *G*. Each token of *p* increases your earnings by 6 points when *G* = 0, and by a larger number of points, rising to a maximum of 18 points per token when *G* = 96 or more. See Table 1 and Figure 1.

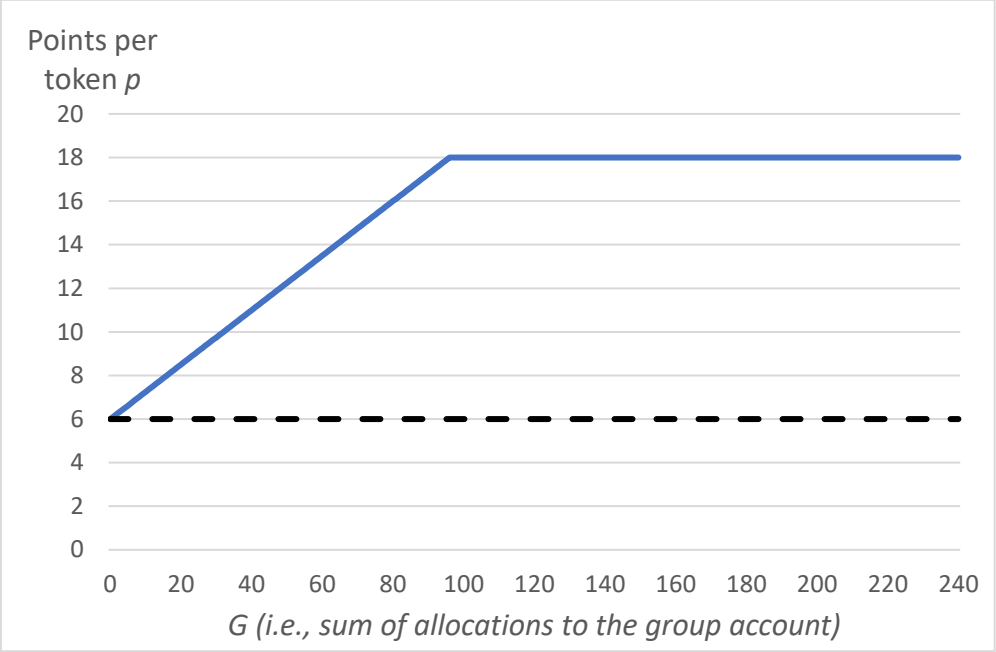



Figure 1. Points earned per token of *p* as a function of *G*

| *G* | 0 | 8 | 16 | 24 | 32 | 40 | 48 | 56 | 64 | 72 | 80 | 88 | 96 | Above 96 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| income from *p* | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 18 |

Table 1: income per token of *p* as a function of *G*

In addition to *G*'s effect on your earnings by influencing the income from tokens assigned to your private account, *G* also affects your earnings in a direct way which is the same for all participants. Each participant in the experiment receives a number of points that rises as *G* does, and that goes equally to participants regardless of their individual choices of *p* and *g*. We will call this the "General Benefit". This general benefit of *G* rises as *G* increases, continuing to rise, although more slowly, even when *G* > 96, as shown in the figure below.

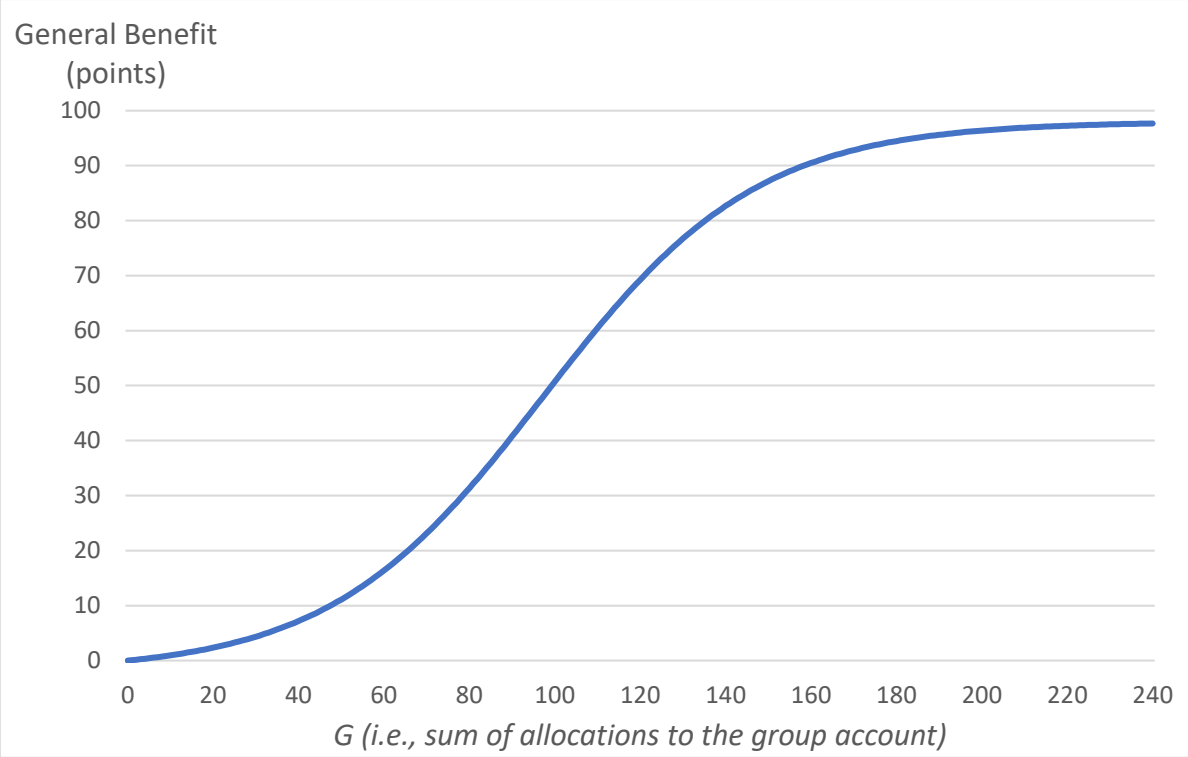



Figure 2. shows the benefit from the group account (*G*) that is given to each participant, regardless of their *p* and *g* token allocation.

The two ways in which allocations to the group account affect earnings—partly through increasing the returns to any token allocated to one's private account, and partly by yielding an equal amount for all participants—are summarized in Table 2. The columns correspond to different allocations of tokens to the group account by you, and the rows correspond to different average allocations of tokens to the group account by the other 11 participants. To make the presentation more compact, the table shows only one's own and others' average allocations that are divisible by four.

| *Average allocation of 11 others* | *Own allocation to the group account (g)* | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 4 | 8 | 12 | 16 | 20 |
| 0 | 120 | 104 | 85 | 61 | 34 | 2 |
| 4 | 239 | 202 | 162 | 118 | 70 | 19 |
| 8 | 379 | 323 | 263 | 195 | 127 | 59 |
| 12 | 438 | 368 | 299 | 229 | 158 | 88 |
| 16 | 454 | 382 | 311 | 239 | 168 | 96 |
| 20 | 457 | 385 | 313 | 242 | 170 | 98 |

Table 2: Earnings as a function of your allocation to the group account (*g*) and the average allocation *g* of the other 11 participants to the group account

We've shaded the diagonal entries of the table, which represent situations in which you and the others in your group happen to allocate the same number of tokens (or for the others, the same number on average) to the group account. For example, the entry *202* (second row from top, second column from left) is the total amount that you would earn if you allocated 4 of your 20 tokens to the group account and 16 of your tokens to your private account, while the other 11 participants allocate an average of 4 tokens each to the group account. Notice that among these shaded diagonal cells, your earnings would be highest when you and the others on average allocate 8 tokens to the group account,

giving you 263 points. That's more than double your earnings if all participants put 0 into the group account, and the fact that it occurs when all allocate 8 tokens to the group account is consistent with the fact that the return from allocating a token to your private account reaches its maximum value when $G$ = 96 (= 12 x 8) (see Figure 1), and that the General Benefit of $G$ (shown in Figure 2) increases at a slower rate after $G$ = 96. Table 2 is available on your screen during the allocation stages of the experiment by double-clicking the 'Payment Table' button. You can also open (and close) an expanded table showing outcomes for all integer combinations of allocations by yourself and others by clicking the 'Full Table' button, which becomes available when the smaller table is open.

Two further things to note are the following. First, your earnings are not sensitive to *how* others' allocations add up to a given average; any combination of choices by others that generates a given average has the same impact on your earnings. Second, what you earn <u>does</u> change if your own allocation varies, taking the average allocation of the others as given. For example, suppose that the others allocate an average of 8 tokens to the group account. You earn more by allocating less than 8 yourself, as shown by the cells to the left of the one with the shaded value of 263. The largest number in the table, 457, is what you would earn if others assigned all their tokens to the group account, while you allocate all of yours to your private account.

In summary, there will be four periods in Part 1 of the experiment followed by a break for further instructions. Operationally, each of the 4 periods in Part 1 will unfold as follows:

- You'll initially see a screen where you'll be asked to decide how many (if any) of the 20 tokens you wish to allocate to the group account (the rest automatically go to your private account).
- When everyone has submitted their decisions, you'll see a screen showing your overall results for the period.
- When you click "Next", you'll see a screen showing the amount that you and each of the other 11 participants assigned to the group account in this period, plus the points that each of you earned. These results will be anonymous; you will only see the tokens allocated and the corresponding points earned.
- You can take a moment to absorb this information, then click "Next" to begin the next period.

**[Instructions for Part 2 (Neutrally Framed – With Feedback)]**

**INSTRUCTIONS FOR PART 2**

The remaining fifteen periods of the experiment have a core structure identical to those of the first four periods. In what we'll now call the "main stage" of each period, you and the other 11 participants each have 20 tokens to allocate between your private account and the group account. However, whereas the allocation decision was strictly voluntary in Part 1, there will now be **a minimum required allocation** to the group account, subject to a penalty if not fulfilled. The allocation to the group account that is required to avoid a penalty will be 8 of your 20 tokens, which, as you will recall, was the allocation (among those in which all allocated equally) at which total earnings of participants were maximized in Part 1. For each token less than 8 that you allocate to the group account, you will be penalized 35 points. The size of the penalty is large so you will definitely earn less if you allocate less than 8 tokens to the group account (see Table 3 below, where the struck through amounts indicate points earned before the penalty has been applied).

| Average $g$ of 11 others | Own allocation to the group account | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 4 | 8 | 12 | 16 | 20 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | ~~120~~<br>-160 | ~~104~~<br>-36 | 85 | 61 | 34 | 2 |
| 4 | ~~239~~<br>-41 | ~~202~~<br>62 | 162 | 118 | 70 | 19 |
| 8 | ~~379~~<br>99 | ~~323~~<br>183 | 263 | 195 | 127 | 59 |
| 12 | ~~438~~<br>158 | ~~368~~<br>228 | 299 | 229 | 158 | 88 |
| 16 | ~~454~~<br>174 | ~~382~~<br>242 | 311 | 239 | 168 | 96 |
| 20 | ~~457~~<br>177 | ~~385~~<br>245 | 313 | 242 | 170 | 98 |

Table 3: Earnings as a function of your allocation to the group account (*g*) and the average allocation *g* of the other 11 participants to the group account when there is a **minimum required allocation** of 8 tokens

In addition to having a minimum required allocation, a further change may also affect the total amount allocated to the group account, *G*, in Part 2. Specifically, the tokens in *G* may be reduced by a percentage, which we will call *R*%, that varies depending on your own and others' actions. Tokens that are removed from the group account by this reduction process will not be used in the calculation of the general benefit received by everyone and won't help to increase your return from allocating tokens to your private account. Given this, *G* can now be re-defined as the total amount of tokens allocated to the group account minus any reductions. We will explain how the percentage that *G* is reduced by is determined next.

Each of the fifteen periods remaining will include an extra stage before the main stage—we'll call it the "pre-stage"—during which you'll have the opportunity to perform two types of tasks. The first type of task, called a **"Type A Task"**, decreases the amount *G* is reduced by in the period's main stage. Put differently, the more Type A tasks that are completed in a period's pre-stage, the smaller the percentage (*R*%) by which *G* gets reduced. The way in which *R*% decreases as you and others increase the number of Type A tasks completed overall is shown in the graph below. *R*% starts at 50% when no Type A tasks are completed; this means that the value of *G* is reduced by 50% before the general benefit and your private return are calculated in the main stage. Completing Type A tasks reduces *R*%, for example, if an average of two Type A tasks are completed by you and the other participants (a total of 24 Type A tasks), then *R*% falls from 50% to 5.5%. This means that only 5.5% of tokens are removed from the total amount allocated to the group account before your earnings are calculated. If 36 or more Type A tasks are completed, no tokens are removed from the total put into the group account—i.e., *R*% = 0.

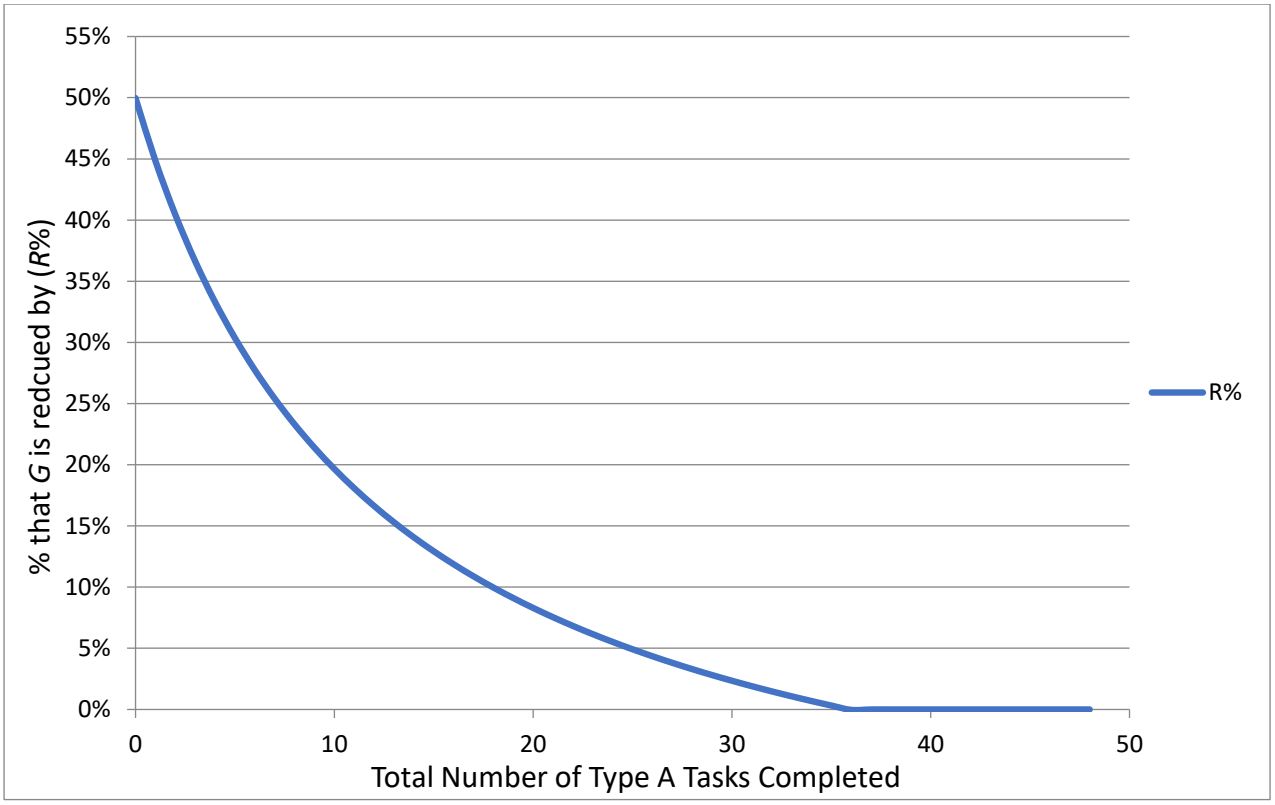



Figure 3. shows the percentage (*R*%) that *G* is reduced by for a given level of Type A tasks completed in total.

To give you an idea of how the percentage reduction (*R*%) affects your earnings, the payment table available on your screen has been updated with a slider. You can adjust the slider for hypothetical numbers of Type A tasks completed by all 12 participants and see the corresponding *R*% and payment table (which is read in the same way as Table 3). The table also accounts for the penalty which is applied if fewer than 8 tokens are allocated to the group account. As in Part 1, you can view an expanded table showing outcomes for all integer combinations of allocations by yourself and others by clicking the 'Full Table' button. Please take a moment to open the table and use the slider to see how *R*% affects the number of points you earn depending on the tokens you and the other 11 participants allocate to the group account.

The second type of activity available during the pre-stage is **"Type B Tasks"**. Completing a Type B task correctly adds 10 points directly to your earnings and has no effect on *R*%. Tasks of both types take about 10 - 15 seconds to complete, and a total of 40 seconds will be available each period for the task portion of the pre-stage. Any points you earn in the pre-stage are added to your overall accumulation and they convert to real money at the same rate as other points at the end of the experiment. The potential to earn points in a period's pre-stage does not affect what allocations you can make in its main stage. You will have 20 tokens available to allocate to the group account and your private account in the period's main stage, regardless of how many tasks you complete.

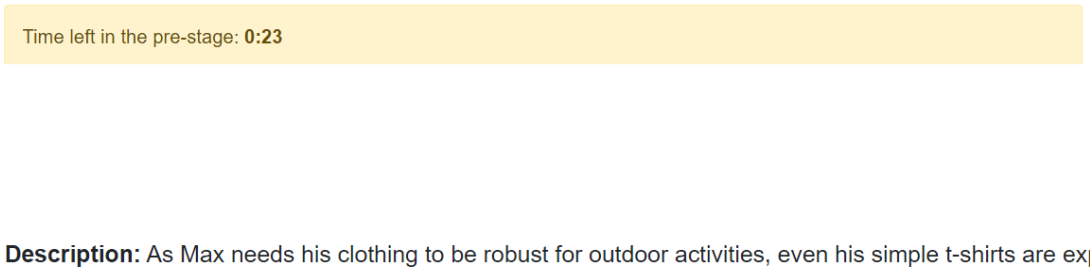
Information sharing and feedback. {Treatments with Feedback Only}

At the end of each period's pre-stage, information about the number of Type A tasks that you and 3 other randomly chosen participants have completed will be displayed, along with their identification letter (A, B, C, or D). The composition of this set of four participants remains fixed for the remainder of Part 2, and will be referred to as your **'subgroup'**. The pre-stage of each period will end with an opportunity to provide feedback to the others in your subgroup, and for them to do the same to you anonymously. Specifically, you can give a smiley face (☺) to any or none of them. On the final pre-

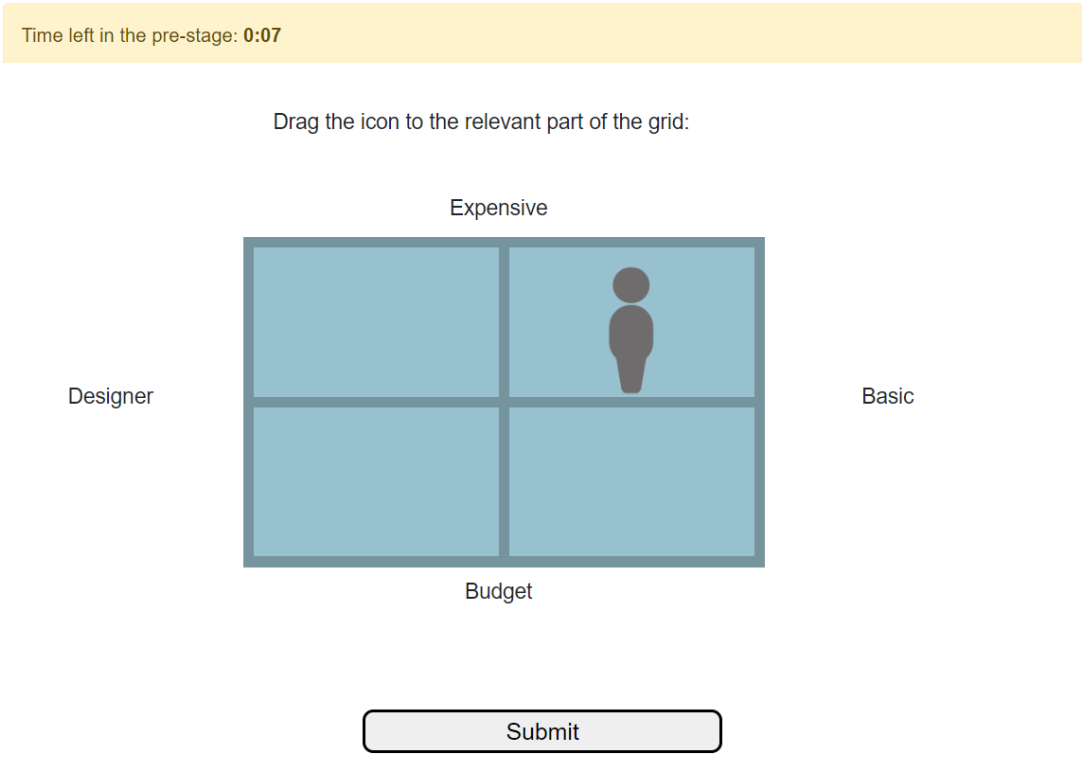
stage screen, you'll be shown the feedback other subgroup members submitted about you (in total), as well as the feedback that the other subgroup members received. {/}

More about pre-stage tasks.

When a Part 2 period begins, always with its pre-stage, you'll see a screen on which you select whether the first task you want to do will be a Type A or Type B task. Once you click on your choice, you'll begin that task. The tasks are identical in nature, only how they impact the main stage differs (as described above). Each task begins with a description of a person differing in two dimensions or characteristics (see screen image 1.a below), for example, what type of food they like and whether they prefer to cook or eat in a restaurant. After reading the description and clicking continue, you'll see a two-dimensional grid (screen image 1.b). There, you'll click and drag a person-shaped icon to whichever of the four quadrants corresponds to the description, drop it in place, and submit that answer by clicking the "Submit" button. Note that you cannot go back from the grid screen to view the description, although you are free to take notes to help you remember it. Once you have submitted an answer, you will be told whether it was correct or not, and then click 'Return' to select the next task type.




screen image 1.a




screen image 1.b

On the screen showing the description of the individual, the experiment software requires you to spend a minimum of 3 seconds before you can continue to the screen showing the four-quadrant grid, the screen with the grid further requires you to spend a minimum of 2 seconds before submitting your answer. This time requirement is to encourage you to pay attention to the tasks, rather than engage in random clicking.

Summary

As mentioned, the task part of the pre-stage will last for a total of 40 seconds. In that time, you may complete Type A tasks, which reduce *R*%, or Type B tasks, which add to your personal earnings without reducing *R*%. When that time runs out, you'll be informed of the total number of Type A tasks completed (from all 12 participants, combined) and of the resulting percentage reduction (*R*%) that will exist in the main stage as a result of this, as well as any earnings from Type B tasks.

{Treatments with Feedback Only} On the next screen, you will be shown the number of Type A tasks completed by each of the other 3 members in your 4-person subgroup, as well as their identification letters. You can then assign feedback (or not) to each of the other members in your subgroup by clicking the checkbox in their row. On the screen after, you will see your own feedback from the other subgroup members, as well as the feedback each other member received. You will not see who you or the other members received feedback from in your subgroup. {/}

When you click continue, you'll go to the main stage (where you can allocate 20 tokens between the group or private account). The main stage will work as in Part 1 except that there is a penalty if you put less than 8 tokens into the group account and the total amount allocated to the group account may be reduced by *R*%, which varies depending on the number of Type A tasks completed by all 12 participants in the pre-stage. The more Type A tasks completed in the pre-stage, the less the amount in the group account will be reduced.