JENA ECONOMIC RESEARCH PAPERS

2012 – 030

Punishment Fosters Efficiency in the Minimum Effort Coordination Game

by

Fabrice Le Lec
Astrid Matthey
Ondřej Rydval

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2012 - 030

Punishment Fosters Efficiency in the Minimum Effort Coordination Game

Fabrice Le Lec

Astrid Matthey

and Ondřej Rydval*

June 19, 2012

Abstract: Using a laboratory experiment, we examine whether informal monetary sanctions can lead to better coordination in a repeated minimum effort coordination game. While most groups first experience inefficient coordination, the efficiency increases substantially after introducing an *ex post* sanctioning possibility. Namely, subjects can assign punishment points to other group members upon observing their efforts, which is costly for the punisher but twice as costly for the punished member. By contrast, introducing instead an *ex post* costless communication possibility fails to permanently increase efficiency. This suggests that decentralized monetary sanctions can play a major role as a coordination device in Pareto-ranked coordination settings, such as teamwork in firms and other organizational contexts.

Keywords: coordination, minimum effort, order-statistic game, punishment, sanction, weakest link
JEL classification: C72, C91, D01, D03

1 Introduction

Coordination issues arise routinely in economic circumstances. In microeconomics, the ubiquity of coordination problems within firms, organizations and

*Le Lec: Lille Economics and Management UMR CNRS 8179, Catholic University of Lille, Lille, France, e-mail: fabrice.lelec@icl-lille.fr, phone: +33 359 56 69 75. (corresponding author)
Matthey: Max Planck Institute of Economics, Jena, Germany, e-mail: matthey@econ.mpg.de, phone: +49 3641 686644
Rydval: Max Planck Institute of Economics, Jena, Germany, and CERGE-EI, Charles University Prague and Academy of Sciences of the Czech Republic, Prague, Czech Republic, e-mail: rydval@econ.mpg.de, phone: +49 3641 686641

even industrial branches has been widely acknowledged (e.g., Becker and Murphy, 1992). A general game theoretic description of such coordination issues is given by the minimum effort game, also known as the weakest-link game: a group member's payoff depends on her own effort (i.e., action) as well as the minimum effort in the group. The higher the minimum effort, the higher every member's payoff. Another well-known specification is the median effort game where the group's median effort is the relevant rank order statistic determining payoff. In contrast to social-dilemma games (e.g., public goods games), any common effort level chosen by all group members is an equilibrium, so it is in no-one's interest to deviate upward or downward from the common effort. Hence choosing the most efficient (i.e., payoff-dominant) equilibrium is a coordination rather than a cooperation problem. Many economic and organizational contexts feature such situations, where agents (e.g., group or team members) have to coordinate on a common action, with the group's success depending on the least favorable (successful, productive, etc.) action of a team member. Among canonical examples are teams of assembly-line workers whose overall productivity depends on the least productive member, teams of construction workers whose ability to proceed to the next construction step hinges on every member having completed a task, law firm cases that are solid only to the extent of its weakest part, or even co-authoring of scientific projects. Camerer and Knez (1994) have underlined how these simple weakest-link coordination games can account for within-firm interactions.

It is then an interesting empirical issue whether agents are able to collectively coordinate on efficient outcomes when facing such situations, whether such outcomes evolve over time, and if there exist external coordination devices that may enhance efficiency. As a consequence, such games have been extensively studied in laboratory experiments, starting with the seminal studies of Van Huyck, Battalio, and Beil (1990) and Van Huyck, Battalio, and Beil (1991). For minimum effort games in particular, ample evidence from various contexts has documented a widespread failure to coordinate on the most efficient or at least a highly efficient outcome on a long-term basis. This raises two issues, the first one of a normative nature: If Pareto-ranked coordination games lead robustly to low efficiency, it is of obvious interest to investigate what kind of coordination devices would improve the collective outcome, and if so, whether they could be applied in real, non-experimental contexts. The second, more descriptive problem is perhaps even more important: In numerous real-world situations of the kind evoked previously, there is little or no evidence of a robust dynamics towards less efficient coordination. Firms or organizations do not seem to get less efficient over time as would be suggested by experimental results. Put differently, the question is how and why organizations (firms or teams) are able to successfully coordinate in situations similar to the minimum effort game, when experimental groups seldom reach efficient coordination, at least after some history of play.

Various efficiency-enhancing features have been tested experimentally and several of those were found to partly achieve this goal, such as smaller groups, higher incentives, finer action space, communication opportunities, and more homogeneous socio-demographic group composition (see section 2 for a detailed account

of the literature). Yet, except for very specific settings, there appears to be a gradual and pronounced failure to coordinate on the payoff-maximizing equilibrium, even with partner matching. We turn to an alternative efficiency-enhancing device, namely voluntary sanctions inflicted on group members deviating from efficient coordination. Such a mechanism has been found to be a powerful force to foster cooperation in public goods games (e.g., Fehr and Gaechter, 2000), suggesting that decentralized, informal sanctions might explain successful cooperation in the field. A similar mechanism may be at work in coordination contexts. For instance, in team projects similar to the examples above, workers may have many opportunities to retaliate against low-effort individuals (e.g., by lack of sharing of strategic information, future refusal of help, etc.). The sociological literature has long put forth that conventions and norms are often, if not always, enforced by individuals, most of the time in an informal, decentralized and voluntary manner (Horwitz, 1990). The possibility of sanctions could thus have a strong effect on coordination dynamics as well as on its efficiency, potentially explaining high levels of efficient coordination in specific real-world settings.

To examine this hypothesis, we set up an experiment based on the minimum effort game whose purpose is to test whether the possibility of *ex post* punishment of fellow group members can foster efficient coordination. Following the original design of Van Huyck, Battalio, and Beil (1990), at the beginning of each round subjects (in groups of eight) choose an effort level between 1 and 7. Then subjects receive (anonymous) feedback on the effort choices of their fellow group members, and, depending on the treatment, can assign points to them: In the *Disapproval* treatment, these points simply act as a communication device signaling disapproval, with no monetary consequence. In the *Punishment* treatment, assigning the points imposes a *fine* on the punished group member, but also comes at a *fee* to the punisher, with the fine being twice as large as the fee. The details of the design are laid out in section 3. The point of interest here is to establish whether the possibility of punishment can lead to an increase in efficiency, just as it does in public goods games (e.g., Fehr and Gaechter, 2000), and to compare it with the effect of disapproval communication as in Masclet, Noussair, Tucker, and Villeval (2003), again in the context of a cooperation game.

To provide an even stronger test, subjects in all treatments first complete eight rounds of play in the baseline design without punishment, likely creating a history of low efficiency that then has to be overcome in the next eight rounds with disapproval or punishment opportunities. A similar setup with a baseline phase has been used, for instance, by Brandts and Cooper (2006) to study the effect of *ex ante* communication, Romero (2011) to examine variation in effort cost, and Fatas, Neugebauer, and Perote (2006) to assess the magnitude of a pure "restart" effect between two succesive identical baseline stages. Based on these studies, we expect to find strong path-dependence and a mild restart effect, providing a strong test of the viability of *ex post* monetary punishment and cheap-talk disapproval as a coordination devices. This initial baseline phase distinguishes our *Disapproval* treatment from a similar disapproval treatment conducted in Dugar (2010).

Our results show that, even after a history of coordination on inefficient equilibria, the possibility to punish individuals in the minimum effort game brings groups to (or very close to) Pareto-optimality in about a third of cases and considerably improves efficiency in another third of cases, even without much punishment being implemented. By contrast, only temporary efficiency improvements are observed in the payoff-neutral disapproval treatment, and only a very limited restart effect takes place in a baseline treatment without any communication or sanctioning device. This suggests that even after a history of inefficient coordination, punishment provides a powerful coordination device, similar to its effect in public goods games, and superior to the effect of an *ex post* communication device alone.

The remainder of the paper is organized as follows. The next section provides an overview on the experimental literature on coordination games and coordination failure, and section 3 presents the experimental design. Results are then described in detail in section 4, and discussed with some concluding remarks in the last one.

2 Background

Since Van Huyck, Battalio, and Beil (1990, 1991), laboratory experiments have shown widespread failure to coordinate on the efficient equilibrium[1] (or a close-to-efficient equilibrium) in order-statistic games in general, and the minimum effort game in particular. This result induced many consecutive attempts to increase the efficiency of coordination through changing certain features of the minimum effort game. For example, changing the payoff structure or the action space of the game have been found to increase efficiency. Brandts and Cooper (2007) show that an increase in incentives (the size of the bonus for coordinating on the efficient outcome) can strongly improve efficiency. Goeree and Holt (2005) find a similar effect for lower effort costs. Van Huyck, Battalio, and Rankin (2007) demonstrate that efficiency improves if participants have a finer action space to choose from and hence (upward) exploration is less costly. In Cachon and Camerer (1996), low-efficiency equilibria lead to negative payoffs, inducing participants to coordinate on more efficient equilibria. Interestingly, Engelmann and Normann (2010) show that even the socio-demographic (or cultural) composition of a group can affect efficiency (in their case, efficiency increases with the share of Danish participants). An extensive overview by Devetag and Ortmann (2007) provides a more comprehensive account of the effects of various changes to the baseline design of van Huyck et al., such as the number of repetitions and group size.

[1] Various terminologies have been used in the literature. We use the term *coordination* to denote homogeneous effort choices within a group regardless of the effort level. Hence, *coordination* on effort level 1 is just as possible as on level 7, as long as (almost) all group members make the same choice. By contrast, *efficient* and *inefficient* effort choices will denote high and low effort levels, respectively. Accordingly, all members of a group choosing effort level 1 is called *coordination on the least efficient equilibrium*, while all choosing effort level 7 is called *coordination on the most efficient equilibrium*. More generally, we will use the terms *more efficient* and *less efficient* to denote higher and lower effort choices, respectively.

In another strand of the literature, introducing *pre-play* communication has been found to increase efficiency by about as much as changing the payoff structure of the game. In Weber, Camerer, Rottenstreich, and Knez (2001), after two rounds of play, a "leader" makes a statement emphasizing the payoff-dominance of the efficient equilibrium. With small groups of only two members, the statement seems to increase efficiency in the remaining six rounds of the game. For large groups of 10 members, the results are less clear. Brandts and Cooper's (2007) subjects first play 10 rounds in the baseline setup. After establishing a history of inefficient effort choices, a "manager" who is not a member of the group can send messages to everyone, trying to induce higher effort choices. This type of centralized communication is shown to raise efforts more than higher financial incentives alone, especially two-way communication (group members being able to reply to the manager). In Blume and Ortmann (2007), prior to actually choosing their effort, subjects can communicate their *intended* effort choice to the other group members. Although actual choices do not always follow announced intentions, overall efficiency significantly increases relative to the comparable baseline treatment without cheap talk. Chaudhuri, Schotter, and Sopher (2009) have a design where players receive pre-play advice from the previous generation of players who have just completed 10 rounds of the game. The advice increases efficiency if it is common knowledge and (almost) unanimously advocates the payoff-dominant equilibrium. Similar efficiency-enhancing effects of pre-play communication have been found for stag hunt games (Charness, 2000; Cooper, DeJong, Forsythe, and Ross, 1992) and median effort games (e.g., Blume and Ortmann, 2007).

Part of our study focuses on the effect of *ex post* communication on coordination and efficiency. To our knowledge, the only study that has analyzed this effect in a coordination game is Dugar (2010). In his design, participants have the opportunity to assign "disapproval points" to their fellow group members after each round of effort choices. In contrast to a mirror treatment where subjects can assign "approval points," disapproval is found to increase coordination on the efficient equilibrium although it has absolutely no monetary consequence. One of our treatments replicates certain features of this design, giving subjects the opportunity to assign disapproval points.

However, all our subjects first play several rounds in the baseline design without the chance to express disapproval, such that most groups will have already had a history of coordination on low efforts when the disapproval opportunity is introduced in a later stage of the experiment. In line with earlier studies that also find an influence of past behavior, we expect this history of inefficiency to influence coordination in later rounds. For example, Romero (2011) analyzes whether behavior adapts after changes in the effort cost. He shows that groups which move *down* to a given cost, i.e., those with a history of high costs, coordinate on lower efforts compared to groups which move *up* to that cost, i.e., those with a history of low costs. Brandts and Cooper (2006) also find path-dependence: While most groups fail to coordinate on high efforts in the first (baseline) stage, those not coordinating on the least efficient equilibrium also choose higher efforts in the second stage. Furthermore, Fatas et al. (2006) find that after restarting the

game, subjects on average choose higher efforts than prior to the restart, although - as in our baseline treatment - the game restarts with exactly the same setup and parameters. Yet the restart effect cannot generally compensate for the history of inefficient coordination since effort levels are on average much lower when restarting than at the very beginning of play. This suggests that it may be more difficult to recover from a history of inefficient coordination than to start afresh with no such history. Path-dependence and restart are interesting effects, since many real coordination situations occur in a repeated context and are divided into sub-stages that allow some form of fresh start, offering a chance to move closer to the efficient outcome. For example, teams that work together in many different projects can try to improve on past performance when entering a new project. We will consider both effects in detail in section 4.

We complement and compare the treatment with *ex post* communication - which can be thought of as non-monetary sanctions - with a parallel treatment where we introduce monetary sanctions. Sanctioning fellow group members has been shown to increase contributions in cooperation games, for instance in public goods games by Fehr and Gaechter (2000), Anderson and Putterman (2006), Carpenter (2007) and others. Despite the difference in the strategic nature of coordination and cooperation games, they share a number of important features, for instance a conflict between socially desirable outcomes and individual decisions. To compare the two effects (disapproval and punishment), we adopt the sanctioning scheme developed in cooperation games, where after each round of contributions subjects can fine their fellow players at a certain cost to themselves (fee). Following the discussion in Casari (2005), we choose a fixed fee-to-fine ratio (see section 3 for details), so that sanctioning other players is equally costly regardless of the punished player's payoff as well as her decision at the coordination stage. Hence changes in the sanctioning behavior can be attributed to changes in behavior or beliefs, rather than changes in costs. Indeed, Galbiati, Schlag, and van der Weele (2009) show for a two-by-two minimum effort game (i.e., stag hunt) game that monetary sanctions can alter beliefs and behavior. However, in their experiment, the effect holds only if sanctions are implemented by the experimenter rather than a third-party player, since due to the limited information subjects receive in their design, sanctioning by the latter carries mixed signals regarding previous choices of other players. Although the design and underlying research question of Galbiati et al. (2009) are very different from ours, their results suggest that monetary sanctions may influence efficiency in coordination settings.

To summarize, the experimental literature tends to show that coordination on efficient outcomes is hard to reach, even though it can be partly enhanced by design variations on the coordination game itself or by efficiency-enhancing devices. Yet, two things should be noted: First, both the design variations and devices only lead to a limited long-run efficiency improvement, and second, especially the design variations tend to change the nature of the coordination game. The aim of our study is then to determine, in a manner similar to Masclet, Noussair, Tucker, and Villeval's (2003) for public goods games, whether punishment opportunities can robustly enhance efficiency in the weakest-link coordination setting, and whether

their effect goes beyond the effect of communicating disapproval.

3 Experiment

3.1 Experimental Protocol

The participants (henceforth called subjects or players) played 16 rounds of the minimum effort game, split into two stages of eight rounds each. At the beginning of the experiment, we handed out the instructions for the first stage and announced that there would be a second stage of unspecified nature (experimental instructions are available in the on-line Additional Material). Subjects also knew that only one of the two stages would be chosen at random to be paid.

In *Stage 1*, all the treatments featured a baseline design closely resembling the seminal one of Van Huyck et al. (1990). Groups consisting of eight players were formed randomly prior to the first round and remained the same for the entire experiment. In each round, players simultaneously chose an integer effort level between 1 and 7. Each player's payoff depended on her effort choice and the lowest effort choice in her group. In particular, let $N = \{1, 2, 3, ..., 8\}$ be the group of players and $E = \{1, 2, .., 7\}$ be the set of effort levels, each player choosing effort $e_i \in E$. With $s = (e_i)_{i \in N}$ being the strategy profile of all players in the group, the payoff (in euros) of player i in a given round is

$$\pi_i(e_i) = 0.4 \times \min_{j \in N}(e_j) - 0.2e_i + 1.2 \tag{1}$$

Table 1 shows the corresponding payoff matrix. This payoff matrix with seven Pareto-ranked equilibria along the main diagonal was used by Van Huyck et al. (1990), Blume and Ortmann (2010), Dugar (2010) and many others.

		minimum effort choice in the group						
		7	**6**	**5**	**4**	**3**	**2**	**1**
	7	2.60	2.20	1.80	1.40	1.00	0.60	0.20
	6		2.40	2.00	1.60	1.20	0.80	0.40
own	**5**			2.20	1.80	1.40	1.00	0.60
choice	**4**				2.00	1.60	1.20	0.80
	3					1.80	1.40	1.00
	2						1.60	1.20
	1							1.40

Table 1: *Payoff matrix of the minimum effort game (in euros)*

After all group members had made their effort choices, the feedback screen displayed the player's effort choice and payoff for the current round as well as her cumulative payoff for Stage 1. The same screen showed table with the current effort choices and payoffs of the other group members, ordered from the lowest to the highest effort. This feedback format is similar to the one used by Engelmann and Normann (2010) and Dugar (2010), and it matches the feedback format required for the subsequent Stage 2 (i.e., for the treatments with monetary and non-monetary sanctions). A player's payoff for Stage 1 consisted of the sum of her round payoffs plus an initial endowment of 4 euros (for reasons explained below).

After Stage 1, subjects received the instructions for *Stage 2* in which the design differed across treatments. In the *Baseline* treatment, Stage 2 was identical to Stage 1. In the *Punishment* treatment, after receiving the feedback on effort choices and payoffs, subjects could (but did not have to) assign punishment points to fellow members. Each point inflicted a cost of 10 cents on the punisher and 20 cents on the punished subject. After all players had assigned points, the feedback screen showed to each player the sum and costs of points assigned by her and to her in the current round, the resulting payoff (or profit) for the current round, and the cumulative payoff for Stage 2. Then the next round started. Note that to the extent that effort choices and payoffs of other group members were ordered from the lowest to the highest effort in each round and hence players' identity was concealed, "retaliation" or punishment of past effort choices was not possible (although we cannot rule out that some players mistakenly believed so). This form of post-punishment feedback was chosen to parallel the one used in public goods games with punishment (e.g., Fehr and Gaechter, 2000, or Anderson and Putterman, 2006).

In order to give players the opportunity to punish in the very first round of Stage 2 independently of their earnings in that round, subjects received an initial endowment of 4 euros. The endowment meant that a player with an effort choice of 7 facing seven other group members choosing effort level 1 was able to almost equalize the profit of all members in the current round (i.e., by assigning 6 points to each of the other members). This ensures comparability between rounds and limits the effect of past earnings on punishment decisions. That being said, the size of the endowment seems innocuous and not suggestive of any punishment strategy. For reasons of symmetry, the 4 euro endowment was granted in both stages of all treatments. Punishment was limited by the punishing player's own cumulative payoff up to the previous round (including the 4 euro endowment). This resulted in an overall payoff for player i of

$$\pi(s)_{Punish} = \max\{0; 0.4 \times \min_{j \in N}\{e_j\} - 0.2e_i + 1.20 - 0.1 \sum_{j \in N} P_{ij} - 0.2 \sum_{j \in N} P_{ji}\} \quad (2)$$

where P_{ij} denotes the punishment points that player i assigns to player j. Anderson and Putterman (2006) use a similar punishment design in a public goods game, with the slight difference that their subjects could only use their current round's game payoff to punish.

The simultaneous choice of punishment points in any given round generates a second order public goods game where players may free-ride on others carrying the cost of punishing group members with low effort choices. This problem is magnified by the fact that punishment points could only reduce other members' game payoff from the current round at the most to zero, so some of the assigned points may be "wasted" in case the points assigned to a given member were to reduce her game payoff to below zero. Subjects of course did not know *ex ante* how many points other members would assign, but the full cost of assigning points had to be born *ex post*.

To compare the effect of monetary and non-monetary sanctions, we ran a third treatment called *Disapproval*. The procedure in this treatment was as similar as possible to the one in *Punishment*, with the important difference that disapproval points did not inflict monetary costs on either the disapproving or the disapproved group member. The points were merely a means of communicating one's opinion about the other members' behavior. After receiving the feedback on effort choices and payoffs, a player could assign between zero and six points (only integer) to each other group member, with six points expressing the maximum disapproval. To parallel the post-punishment feedback provided in the *Punishment* treatment, the last screen in each round showed to each player the sum of points assigned by her and to her (in the current round), the payoff for the current round, and the cumulative payoff for Stage 2. The feedback slightly differs from the disapproval treatment in Dugar (2010) where subjects could in addition observe the sum of points assigned to their fellow group members. Other differences between ours and Dugar's design are the number of group members and the number of rounds in a given stage - in both cases eight in ours and 10 in Dugar's. Judged from the literature surveys of Devetag and Ortmann (2007) and Engelmann and Normann (2010), such minor variation in these design features appears to have little or no (consistent) effect on coordination outcomes.

A likely more important design difference is the absence of Stage 1 in Dugar's experiment. There are at least two reasons for including the initial baseline Stage 1 in all our treatments. First, we wished to examine the effect of our treatment manipulation after a history of inefficient effort choices (anticipated on the basis of the findings of previous studies with similar design features), which arguably allows us to draw stronger conclusions regarding the effect of monetary and non-monetary sanctions, if any. An initial baseline stage is not present, e.g., in the cheap-talk communication treatment of Blume and Ortmann (2007) and in the disapproval treatment of Dugar (2010), but a baseline stage similar to ours is present in the communication treatment of Brandts and Cooper (2007) and in public goods games with punishment (e.g., Fehr and Gächter, 2000). The second reason for having Stage 1 is that it permits a difference-in-differences comparison of behavior across treatments. In other words, in addition to the standard contemporaneous across-treatment comparison of behavior in Stage 2, we are able to compare treatments in terms of between-stage *changes* in behavior, hence accounting for across-treatment differences in groups' and individuals' initial propensity to coordinate efficiently.

Treatment	Stage 1 (rounds 1-8)	Stage 2 (rounds 9-16)
Baseline	Baseline MEG	Baseline MEG
Disapproval	Baseline MEG	MEG with disapproval point assignment
Punishment	Baseline MEG	MEG with punishment point assignment

Table 2: *Treatments and Stages*

To put it in a nutshell, the three treatments are presented in Table 2 (MEG stands for minimum effort game). As mentioned already, the first stage adds a certain within-subject feature: Given that groups are mostly expected to fail to reach efficient equilibria, and given the strong path-dependence of coordination behavior documented in earlier studies, our design constitutes a unfavorable within-subject test of *Disapproval* and *Punishment*. Put differently, finding a between-stage efficiency improvement can be interpreted as a strong effect of the punishment or disapproval devices. In addition, controlling for across-treatment differences observed in the common first stage yields a more balanced between-subject comparison of the three treatments in the second stage.

3.2 Participants and procedures

Eight sessions of 32 subjects were run, at the lab of the MPI of economics in Jena in november and december 2011, for a total of 256 subjects composing eight groups for *Baseline*, 12 groups for *Disapproval*, and 12 groups for *Punishment*.[2] The experiment was programmed and conducted in Z-Tree (Fischbacher, 2007) and took on average 80 minutes. Including a 2.50 euro show-up fee, the average earnings in the experiment were 18.18 euros (around 24 USD), ranging between 7.10 and 27.30 euros.

The participants were recruited among students of various disciplines at the local university using the ORSEE software (Greiner, 2004). In each session, gender composition was approximately balanced and each subject took part only in one session. Upon arrival at the laboratory, subjects were randomly assigned to one of the computer terminals. Each terminal is in a cubicle that does not allow communication or visual interaction among the participants. Participants privately read the instructions at their own pace and could ask for clarification. In order to check the understanding of the instructions, subjects were asked to answer several control questions. After all subjects had answered the questions correctly, the experiment started. At the end of the session, participants were asked to complete a questionnaire on standard sociodemographics and several debriefing open questions about the experiment. Eventually, participants were paid in cash according to their performance, with privacy being also guaranteed during the payment phase.

[2]Another session was run with the baseline condition (32 subjects, 4 groups), but because of a technical problem, the second part of the experiment could not be run. The results of this session are not reported here, but are similar to what is observed in the first stage of the experiment in all treatments.

4 Results

We first provide a descriptive account of observed behavior (attained efficiency and coordination outcomes) and a statistical analysis of the across-treatment differences. Next, we analyze punishment and disapproval behavior. Finally, we study how treatments affected welfare and overall efficiency.

4.1 Coordination and efficiency

For each treatment, Figure 1 shows the evolution of average effort, and Figure 2 displays the evolution of average minimum effort (i.e., the average of groups' minimum effort). In Stage 1, both figures suggest little or no across-treatment differences. In all treatments, the average effort is initially around 5 and gradually falls to 2. Average minimum effort starts off at about 2 and does not diverge much from that level throughout the stage (ending slightly further below the initial level in *Punishment*). Note that at the end of the stage, the average effort is only marginally above the average minimum effort, especially in *Disapproval* and *Baseline*. This implies low within-group variance of effort choices – i.e., groups mostly coordinate on particular equilibria and hence solve the *individual coordination problem*. At the same time, the low average efficiency means that groups mostly converge to inefficient equilibria and hence do not overcome the *collective coordination problem*, using Van Huyck, Battalio, and Beil's (1990) terminology. Hence all treatments feature a large scope for efficiency gains in Stage 2, the across-treatment comparison of which is the primary aim of our study.

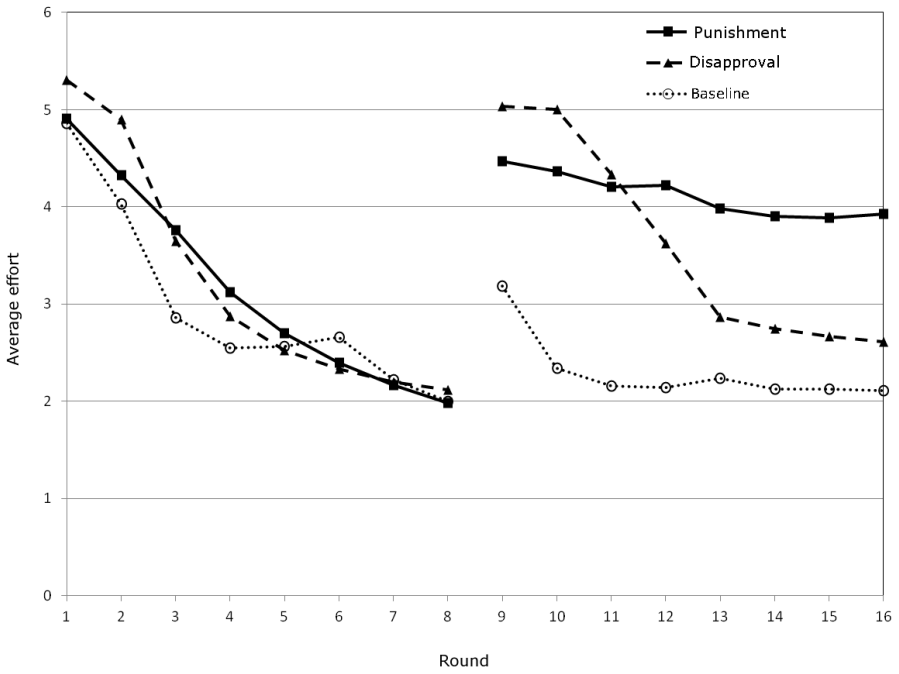

Figure 1: *Average effort per round and treatment*

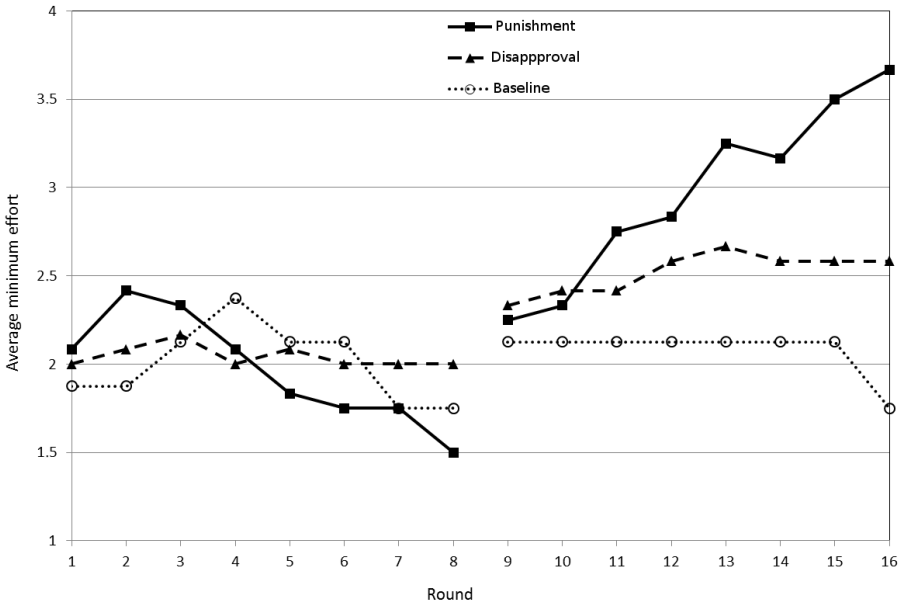

Figure 2: *Average minimum effort in groups per round and treatment*

Figure 3 presents a more disaggregate look at effort choices. In all treatments, the highest effort level is initially the most frequent choice and the lowest effort level is chosen by less than a tenth of subjects. Throughout Stage 1, effort-choice distributions in *Baseline* and *Disapproval* gradually polarize towards the highest and especially the lowest effort level, the latter eventually comprising over three-quarters of choices in both treatments. This is so because six of the eight *Baseline* groups and 10 of the 12 *Disapproval* groups converge or almost converge to the least efficient equilibrium by the end of Stage 1. The remaining two *Disapproval* groups and one *Baseline* group converge to the most efficient equilibrium, while one *Baseline* group coordinates less successfully and features effort choices 4 and 5 as well as three lowest-effort choices at the end of Stage 1. The effort-choice distribution in *Punishment* remains less polarized throughout Stage 1, with almost no-one choosing the highest or even an above-average effort level in the last two rounds. However, matching the other two treatments, the lowest two effort levels taken together eventually comprise over three-quarters of choices. This is due to six of the 12 groups converging or almost converging to the least efficient equilibrium and four groups converging to the level-2 equilibrium (i.e., the equilibrium in effort level 2). The remaining two groups attempt to coordinate on effort level 4, one of them with less success.[3] Although the effort-choice distributions in Stage 1 visually somewhat differ across treatments, Figures 1 and 2 suggest that the differences are not very pronounced in aggregate, nor are they significant (see subsection 4.2). On a general note, the observed patterns of individual and group behavior qualitatively match typical findings in the literature: First, effort levels decrease with rounds, and second, following an initial period of miscoordination, groups tend to coordinate on low-efficient equilibria, mostly the least efficient one.

[3]To get a more detailed picture of group behavior, an interested reader can inspect Tables A1, A2 and A3 in the Appendix dedicated to group-level data, see on-line Additional Material.

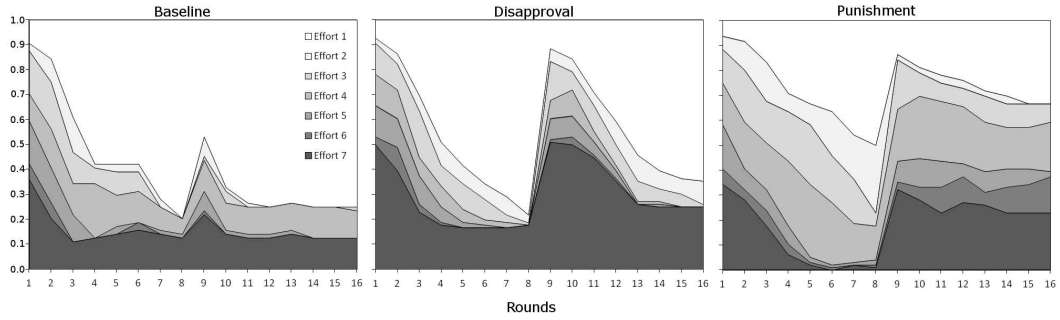

Figure 3: *Distribution of effort choices per round and treatment*

Turning to Stage 2, Figure 1 shows that the average effort jumps up to 4.5 in *Punishment* and to 5.0 in *Disapproval* in the restart round 9, in both cases almost reaching the initial round 1 level. Subsequently, the *Punishment* effort falls slowly over time to reach 3.9 in the final round 16, whereas the *Disapproval* effort falls much faster from round 11 onwards to eventually reach 2.6. These developments can be compared with the pure restart effect in *Baseline* where the average effort jumps up much less and then falls almost immediately back to the lowest level of 2.1 reached at the end of Stage 1. Figure 2 indicates a small positive restart effect also for average minimum efforts. In *Punishment*, minimum effort subsequently rises markedly to eventually reach 3.7, whereas in *Disapproval* minimum effort increases only mildly for several rounds, thereafter remaining at 2.6. The average minimum effort in *Baseline* remains at 2.1 throughout Stage 2, except for a slight drop in the final round. As for Stage 1, average efforts are just above the respective average minimum efforts in all treatments at the end of Stage 2, implying that groups mostly manage to coordinate on particular equilibria.[4] The attained equilibria involve on average more efficient effort levels in *Punishment* compared to *Disapproval* and especially to *Baseline*.

The effort-choice distributions in Figure 3 confirm the aggregate picture observed in Figure 1. The *Baseline* distribution reflects the small, temporary restart effect, while the rest of Stage 2 resembles the last two rounds of Stage 1. At the group level, *Baseline* features strong between-stage inertia: The six groups converging to the least efficient equilibrium in Stage 1 also remain or quickly converge to that equilibrium in Stage 2; one group sustains coordination on the most efficient equilibrium reached in Stage 1; and the remaining group manages to better coordinate on the level-4 equilibrium. The *Disapproval* distribution is similar across stages, except for eventually shifting slightly upwards. For instance, comparing the last two rounds of each stage, the fraction of subjects choosing the two highest effort levels increases from 17 % to 25 %, while the fraction choosing

[4]Thus Figure 3 conveniently portrays not only aggregate effort-choice distributions but approximately also the fraction of groups coordinating on particular equilibria in the final rounds of each stage.

the two lowest effort levels decreases from 80 % to 72 %. This efficiency gain is almost solely due to one group fully recovering from the least efficient equilibrium reached in Stage 1 to eventually coordinate on the most efficient equilibrium in Stage 2. The other nine groups reaching the least efficient equilibrium in Stage 1 also converge to that equilibrium in Stage 2 despite larger positive restart effects compared to *Baseline* (except for one group that manages to eventually coordinate on the level-2 equilibrium). The two remaining groups sustain coordination on the most efficient equilibrium reached in Stage 1.

Figure 3 documents a much stronger efficiency gain in *Punishment* where the fraction of subjects choosing the two highest effort levels rises from 2% to 36 % between the last two rounds of each stage, while the fraction choosing the two lowest effort levels falls from 70% to 33%. The *Punishment* distribution also remains less polarized throughout Stage 2 compared to the other two treatments. *Punishment* features a much lower extent of between-stage inertia at the group level compared to the other treatments, as about half the groups make substantial collective coordination improvements between stages. Of the six groups converging to the least efficient equilibrium in Stage 1, two strongly recover and reach high efficiency on average, though they do not manage to fully coordinate on particular equilibria; the remaining four groups converge to the least efficient equilibrium, despite large and durable restart effects for a couple of them. The four groups converging or almost converging to the level-2 equilibrium in Stage 1 reach the level-3, level-4, level-6 and level-7 equilibria by the end of Stage 2, respectively. The group coordinating on effort level 4 in Stage 1 reaches the most efficient equilibrium in Stage 2. The remaining group manages to better coordinate on the level-4 equilibrium but otherwise makes no efficiency gain.

In sum, Stage 2 generates across-treatment efficiency differences in the posited direction. From about the same aggregate starting point at the end of Stage 1, the efficiency gains in Stage 2 are initially slightly larger in *Disapproval* than in *Punishment* - perhaps reflecting subjects' initial hopes of the effectiveness of the cheap-talk communication device - but these hopes fade off rather quickly and the efficiency gains are eventually considerably larger in *Punishment* than in *Disapproval*. Except for a small positive restart effect, Stage 2 brings about no efficiency gains in *Baseline*.

4.2 Statistical tests on treatment effects

The above described behavioral patterns are mostly confirmed by statistical tests. To study across-treatment differences in efficiency, we first establish whether the contemporaneous across-treatment differences (i.e., differences in a given stage or round) presented in Figures 1 to 3 are statistically significant. We then inspect effort *changes* (gains or losses) between stages as well as between *matched round-pairs* (i.e., rounds 1 and 9, 2 and 10, 3 and 11, and so forth), and assess whether these changes are economically meaningful and statistically different across treatments. We test for treatment effects in this difference-in-differences

manner in order to ensure that across-treatment differences observed in Stage 2 do not stem from differences originating in Stage 1. This concern is even more imminent given the substantial between-stage inertia of group behavior mentioned above. The analysis is performed both for within-subject changes and within-group changes, whenever appropriate. All tests are two-sided. Throughout the section, any across-treatment difference that is not reported as being significant is actually not significant at the 10 % level.

We first compare effort choices by the Mann-Whitney U test applied to average efforts at the group level.[5] In Stage 1, the across-treatment differences do not turn out significant both overall and in each round, reflecting the identical design setup across treatments. In Stage 2, groups' average efforts are significantly higher in *Punishment* compared to *Baseline* both overall ($p < 0.05$) and in the first six rounds ($p < 0.05$ in rounds 10-12 and 14; $p < 0.10$ otherwise), and also significantly higher in *Disapproval* compared to *Baseline* both overall ($p < 0.05$) and in the first five rounds ($p < 0.05$ in round 10; $p < 0.10$ otherwise).

Parametric tests provide a similar degree of statistical support for the across-treatment differences. In particular, Wald tests from ordered probit estimation indicate that effort does not significantly differ across treatments in Stage 1 overall as well as in each round.[6] As an exception, effort in round 1 is significantly higher in *Disapproval* compared to both *Punishment* and *Baseline* ($p < 0.10$ in both cases). In Stage 2, effort is higher in *Punishment* compared to *Baseline* both overall ($p < 0.10$) and in the first five rounds ($p < 0.05$ in round 10; $p < 0.10$ otherwise). Effort is also higher in *Disapproval* compared to *Baseline* in the first three rounds ($p < 0.01$ in round 10; $p < 0.05$ otherwise).

Turning to minimum effort instead of average effort, groups' minimum efforts do not significantly differ across treatments in Stage 1 overall and in individual rounds, both by the Mann-Whitney U test and the Wald test.[7] Confirming the

[5] Depending on the type of comparison, groups' average efforts in a given treatment are calculated for each stage or each round. In round 1, we apply the test directly to effort choices since these are not correlated within groups.

[6] We regress effort choices on treatment dummies interacted with a stage dummy or round dummies (for across-treatment comparison at the stage level or the round level, respectively). The estimations are based on a panel of 256 subjects with 16 rounds of effort choices each. We use the cluster-robust estimator of variance allowing for intra-group correlation of effort choices. The number of clusters (i.e., groups) seems sufficient given the perfectly balanced cluster sizes (e.g., Kezdi (2004); Rogers (1993)). The results are unaffected if including a second level of clustering at the subject level, or instead including group- and individual-level random effects. Since regressors comprise only categorical variables and their interactions, it turns out practically irrelevant for the standard errors and hence the Wald tests whether the treatment-round interactions are estimated simultaneously or round-by-round (thus saving degrees of freedom). Wald tests from ordered logit models and t-tests from linear probability models (i.e., OLS regressions) yield very similar results in terms of significance levels, as do separate estimations for round 1 performed without group clustering (since effort choices are independent).

[7] We regress groups' minimum efforts on treatment dummies interacted with a stage dummy or round dummies. The estimations are based on a panel of 32 groups with 16 rounds of minimum efforts each. As above, we use the cluster-robust estimator of variance allowing for intra-group

observation in Figure 2, the effect of punishment opportunities is most pronounced towards the end of Stage 2. For both tests, minimum effort is significantly higher in *Punishment* compared to *Baseline* in the final round 16 ($p < 0.10$).

	Treatment	Stage 1-2	Round 1-9	Round 2-10	Round 3-11	Round 4-12	Round 5-13	Round 6-14	Round 7-15	Round 8-16
Average effort change	Punishment	0.95 ww,ss bbb	-0.44 bb	0.04 bbb	0.45 bb	1.09 ww,s bb	1.28 www,s bbb	1.51 www,s bbb	1.72 www,ss bbb,d	1.95 www,ss bbb,d
	Disapproval	0.38 bb	-0.27 bbb	0.10 bbb	0.69 bb	0.75	0.34	0.42	0.47	0.50
	Baseline	-0.66 ww,ss	-1.67 www,ss	-1.69 www,ss	-0.70 w,ss	-0.41	-0.33 ss	-0.53 w,ss	-0.09	0.11
Fraction of groups with an average-effort increase	Punishment	0.54 bbb	0.33 bb	0.50 bb	0.58	0.58	0.58 bb	0.58 bb	0.58 b	0.58
	Disapproval	0.36 bb	0.42 bb	0.58 bbb	0.67 b	0.42	0.17	0.17	0.25	0.25
	Baseline	0.11	0.13	0.13	0.13	0.13	0.00	0.00	0.13	0.25
Average minimum-effort change	Punishment	1.00 www,sss bb	0.17	-0.08	0.42	0.75 b	1.42 www,ss bbb	1.42 www,ss bbb	1.75 www,sss bb,d	2.17 www,sss bbb,dd
	Disapproval	0.48	0.33	0.33	0.25	0.58 b	0.58	0.58 b	0.58	0.58 b
	Baseline	0.08	0.25	0.25	0.00	-0.25	0.00	0.00	0.38	0.00
Fraction of groups with a minimum-effort increase	Punishment	0.45 bb,d	0.33	0.17	0.42	0.42 b	0.50 bb,d	0.50 bb	0.58 b,dd	0.67 bbb,dd
	Disapproval	0.24	0.33	0.33	0.33	0.25	0.17	0.17	0.17	0.17
	Baseline	0.06	0.13	0.13	0.13	0.00	0.00	0.00	0.13	0.00

The "w", "t" and "s" superscripts denote a significant difference across stages or across a round-pair (see the top row), using an appropriate ordered probit Wald test, t-test, and Wilcoxon signed-rank test, respectively, as described in Section 4.2. The "b" resp. "d" symbols denote a significant difference across stages or across a round-pair between the treatment directly above the symbol and *Baseline* resp. *Disapproval*, using an appropriate ordered probit Wald test (in the first and third blocks) or Mann-Whitney U test (in the second and fourth blocks). Significance levels are 1%, 5% resp. 10% for three, two resp. one superscripts or symbols of a kind in a given cell.

Table 3: *Between-stage and between-round effort changes in each treatment*

As mentioned at the beginning of this section, we now turn to analyzing behavioral changes between stages, the across-treatment comparison of which provides for a cleaner test of treatment effects. The results are provided in Table 3. The first row (i.e., block of results) displays effort changes and their statistical significance between Stages 1 and 2, both overall and for each round-pair. From the same level of about 3 in Stage 1, the average effort in Stage 2 increases by 0.95 (30 percent) in *Punishment* and 0.38 (12 percent) in *Disapproval*, whereas it decreases by 0.66 (22 percent) in *Baseline*. The overall efficiency gain in *Punishment* as well as the overall efficiency loss in *Baseline* are significant by both the ordered probit Wald test described above and the Wilcoxon signed-rank test applied to groups' average efforts. *Punishment* features an initial average-effort decrease in the first round-pair followed by increases that become larger over time. The effort increases in the last five round-pairs are significant. A pattern of initial average-effort decrease followed by an increase also occurs in *Disapproval*, but the increase fades off after the fifth round-pair and subsequently remain much smaller compared to the ones in *Punishment*; effort changes are not significant in any

correlation of observations. The results are unaffected if instead including group-level random effects. Other estimation details are identical to the estimation for effort choices.

round-pair. *Baseline* generally features average-effort decreases of declining magnitude (except for a small increase in the last round-pair) which are significant in the first three round-pairs and in the fifth and sixth round-pairs.

The overall picture is therefore one of rising efficiency gains in *Punishment* which increasingly outweigh those in *Disapproval*, and one of efficiency losses in *Baseline*. The treatment effect tests presented in the first block of Table 3 show that the efficiency differences between *Punishment* and *Baseline* are strongly significant. In particular, Wald tests from ordered probit estimation indicate that the positive treatment effect between *Punishment* and *Baseline* is significant both overall (see column titled "Stage 1-2") and in each round-pair, while the positive treatment effect between *Disapproval* and *Baseline* is significant overall and in the first three round-pairs. Last, the positive treatment effect between *Punishment* and *Disapproval* is weakly significant in the last two round-pairs.[8]

The second block of results in Table 3 displays the fraction of groups with an average-effort increase (i.e., efficiency gain) between the stages. In *Punishment*, the fraction rises to seven out of 12 groups in the third round-pair and remains at that level till the end. In *Disapproval*, the fraction is initially higher compared to *Punishment* in the first three round-pairs, but at most three out of 12 groups register an average-effort increase in the last four round-pairs. In *Baseline*, at most one out of eight groups register an average-effort increase (except for two groups in the last round-pair). We also present another set of treatment effect tests, comparing groups' average-effort changes across stages and across round-pairs by the Mann-Whitney U test. The positive treatment effect between *Punishment* and *Baseline* is significant overall and for all but the third, fourth and the last round-pair. The positive treatment effect between *Disapproval* and *Baseline* is significant overall and for the first three round-pairs, while the treatment effect between *Punishment* and *Disapproval* is never significant. Although the non-parametric test results are weaker compared to the Wald tests in the first block, both blocks together suggest a persistent shift towards more efficient effort choices in *Punishment*, as opposed to (eventually) a much weaker shift in *Disapproval* and no such shift in *Baseline*.

The third and fourth blocks in Table 3 display minimum-effort changes between Stages 1 and 2. From the same level of about 2 in Stage 1, minimum effort in Stage 2 increases on average by 1.00 (51 percent) in *Punishment*, 0.48 (24 percent) in *Disapproval*, and 0.08 (4 percent) in *Baseline*. The third block further shows that the overall minimum-effort efficiency gain is significant only in *Punishment*, by both the ordered probit Wald test (see footnote 7) and the Wilcoxon signed-rank test applied to groups' average minimum efforts. The efficiency gains are also

[8] We regress within-subject effort-choice changes on treatment dummies, and their interaction with round-pair dummies whenever performing separate tests for each round-pair. The estimations are based on a panel of 256 subjects with eight effort-choice changes each (i.e., changes between rounds 1 and 9, 2 and 10, etc.). As above, we use the cluster-robust estimator of variance allowing for intra-group correlation of observations. Other estimation details are identical to the estimation for effort choices.

significant in the last four round-pairs by both the Wald test and the Wilcoxon signed-rank test applied to groups' minimum efforts. In fact, starting from the third round-pair, the minimum-effort efficiency gains in *Punishment* are similar in magnitude, and hence also growing at a similar pace, as are the corresponding average-effort efficiency gains presented in the first block. On the other hand, the minimum-effort efficiency gains remain relatively small and not significant in *Disapproval*, and even smaller (or none) and also not significant in *Baseline*.

The third block also presents Wald tests of treatment effects for the minimum-effort changes.[9] The positive treatment effect between *Punishment* and *Baseline* is significant both overall and in the last five round-pairs. The positive treatment effect between *Disapproval* and *Baseline* is significant in the fourth, sixth and last round-pair. The positive treatment effect between *Punishment* and *Disapproval* is significant in the last two round-pairs.

The fourth block of results in Table 3 complements the third block by displaying the fraction of groups with a minimum-effort increase between stages. In *Punishment*, the fraction is initially four out of 12 groups and it gradually doubles by the last round-pair. In *Disapproval*, the fraction is initially the same as in *Punishment* but drops to two out of 12 groups in the last four round-pairs. In *Baseline*, at most one out of eight groups, and eventually no group, registers a minimum-effort increase. The fourth block also presents another set of treatment effect tests, comparing groups' minimum-effort changes across stages and across round-pairs by the Mann-Whitney U test. The positive treatment effect between *Punishment* and *Baseline* is significant overall and in the last five round-pairs, whereas there are no significant treatment differences between *Disapproval* and *Baseline*. The positive treatment effect between *Punishment* and *Disapproval* is significant overall and in the fifth and the last two round-pairs. In sum, both the Wald test and the Mann-Whitney U test support the overall picture of persistent differences in minimum-effort efficiency gains between *Punishment* and *Baseline* in the last five round-pairs and between *Punishment* and *Disapproval* in the last two round-pairs.

Overall, the results yield a consistent picture. *Baseline* replicates tightly the typical findings in the literature on experimental Pareto-ranked games, namely, gradual convergence to low-efficiency coordination and a very small and temporary efficiency improvement in a restart stage (such as our Stage 2). Both *Disapproval* and *Punishment* bring about substantial efficiency gains following the restart, but only in *Punishment* does this positive effect persist throughout the restart stage and gets stronger over time in terms of the outcome of the game, i.e., minimum effort. The strong positive effect of *Punishment* vis-a-vis the other treatments is

[9] We run ordered probit estimations of within-group minimum-effort changes on treatment dummies, and their interaction with round-pair dummies whenever performing separate tests for each round-pair. The estimations use a panel of 32 groups with 8 minimum-effort changes each (i.e., changes between rounds 1 and 9, 2 and 10, etc.). As above, we cluster observations at the group level. Other estimation details are identical to the estimation for minimum efforts.

evident not only in terms of the plain between-subject comparison in Stage 2, but also in terms of the unfavorable within-subject and within-group comparison of efficiency gains between the stages. Voluntary monetary sanctions in *Punishment* hence seem capable of persistently increasing coordination efficiency levels, even in groups that previously converged to very inefficient coordination outcomes. By contrast, the effect of *ex post* cheap talk in *Disapproval* does not seem strong enough to stabilize coordination at a substantially higher efficiency level than in the *Baseline* treatment.

4.3 Punishment and disapproval behavior

It is of course of interest to know what kind of punishment behavior, and perhaps to a less extent what kind of disapproval behavior, may drive the observed coordination outcomes. In *Punishment*, 657 points are assigned overall - 80% in the first four rounds - inflicting a total cost of 65.7 euros on the punishers and 131.4 euros on the punished (i.e., about 9% of punishment points are not actually implemented because they would decrease a punished subject's round payoff to below zero). Figure 4 shows that the percentages of punishers and punished start off at 44% and 53%, respectively, and both the percentages decline gradually to 9% in the final round. Each punisher initially assigns four points on average. This figure declines gradually to below two points in the penultimate round and then jumps back to four points in the final round.

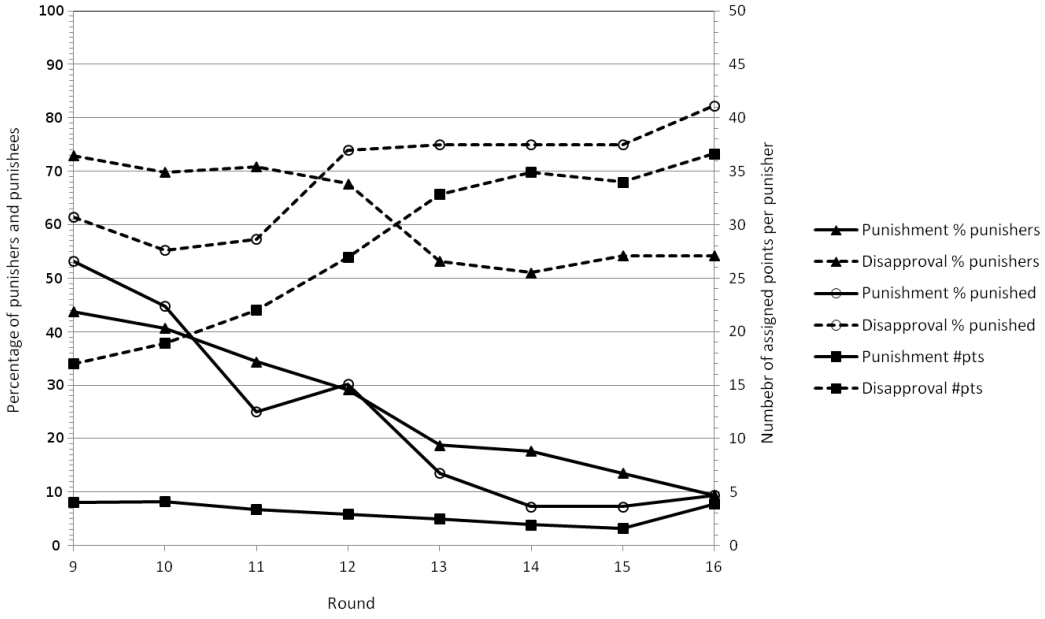

Figure 4: *Punishment and disapproval points assignment*

In *Disapproval*, 12766 points are assigned overall - 45% in the first four rounds - which is almost 20 times higher than in *Punishment*. The percentage of disapprovers starts off at 73 % and is still at 54% at the end, while the percentage of

disapproved begins at 61% and eventually rises to 82%. Each disapprover initially assigns 17 points on average, and this figure steadily rises to eventually reach 37 points, rather close to the maximum of 42 points. Thus disapproval is much more widespread than punishment and the gap widens over time. These differences are unlikely to be driven solely by the across-treatment differences in effort-choice distributions, especially not in the initial rounds of Stage 2 where the distributions are still relatively close to one another. It seems more plausible that the across-treatment differences stem – in addition with the obvious gap regarding monetary consequences – from differences in the nature of and the underlying motives behind punishment and disapproval, as is also apparent from how the points are targeted.

		punished subject's effort level							Row
		1	2	3	4	5	6	7	total
	1	3.2	0	0	0	0	0	1.1	4.7
punisher's	2	0	0	0	0	0	0	0	0
	3	3.7	4.6	0.8	0.3	0	0	0	9.3
effort	4	5.5	2.9	5.0	0.8	0.2	0.2	0.9	15.4
	5	3.5	1.5	2.0	3.0	0.2	0.5	0	10.7
level	6	1.5	0.6	1.4	5.3	4.0	0.2	0.2	13.1
	7	19.6	1.2	6.5	11.9	2.9	3.5	1.2	46.9
Col. Total		37.0	10.8	15.7	21.8	7.2	4.3	3.3	657 pts

Table 4: *Percentage of points assigned in* Punishment

		disapproved subject's effort level							Row
		1	2	3	4	5	6	7	total
	1	49.9	2.8	1.6	0.3	0.4	0	0.4	55.5
disapprover's	2	6.8	3.2	1.5	0.3	0.1	0	0.2	12.0
	3	3.4	1.7	2.0	0.7	0.5	0.0	0.3	8.6
effort	4	1.5	0.6	0.9	0.2	0.1	0	0.1	3.4
	5	1.7	0.6	1.4	0.7	0.2	0	0	4.6
level	6	0	0	0.1	0.2	0.1	0	0	0.5
	7	6.2	1.0	3.7	1.9	1.4	0.8	0.3	15.3
Col. Total		69.6	9.9	11.7	4.3	2.8	0.9	1.3	12,766 pts

Table 5: *Percentage of points assigned in* Disapproval

In particular, Tables 4 and 5 display the distribution of punishment and disapproval points, respectively, aggregated across Stage 2, conditional on effort choices of the subjects by whom and to whom the points were assigned. In *Punishment*, punishers assign 90% of points to group members with a lower effort than theirs, i.e., the assigned points appear below the main diagonal of Table 4. The most populated bottom-left cell contains points of punishers with effort level 7 assigned to subjects with effort level 1. As could be expected, punishment points are mainly targeted at 'shirkers' (i.e., subjects with the lowest effort in the group). The second-order public good problem is typically present since most points are assigned by few group members (not always those with the highest effort); other

members seemingly prefer to instead 'signal' their desire to raise efficiency by choosing a high effort level. As to the remaining 10% of points that punishers assign to group members with the same or even higher effort level than theirs - i.e., the points located on or above the main diagonal of Table 4 - most of the cases appear in just four groups featuring various degree of coordination success. The reasons for this kind of punishment are hard to judge (recall that subjects observe neither the identity nor the effort level of the punisher) since the cases are rather scarce and erratic, with only two subjects punishing in this manner repeatedly, namely three and four times.[10] Both belong to the shirkers in their groups which eventually reach almost the highest efficiency, so both the subjects gradually increase their effort alongside punishing members with the same or higher effort.

In this sense, punishment behavior in our minimum effort game replicates quite well the typical findings for cooperation games, e.g., Fehr and Gaechter (2000); Fehr and Gächter (2005). In both studies, subjects who by their behavior reduce the efficiency of the outcome tend to be punished. However, while a direct comparison between cooperation (public good) and coordination games is difficult, it seems that cooperation games produce more punishment overall and that more group members take part in it. See for instance Fehr and Gächter (2005) where 1,270 punishment points roughly equivalent to ours were assigned in 10 rounds, to be compared with less than 700 points in our 8 rounds.

In *Disapproval*, only 35% of points are assigned by disapprovers to group members with a lower effort than theirs. The main reason for this much lower percentage compared to *Punishment* is that half of all disapproval points are assigned from shirkers to other shirkers choosing the same effort level 1 (see the top-left cell of Table 5). These points are assigned in the eight groups that converge to the least efficient equilibrium, mostly in the last several rounds where the groups already reached or almost reached the equilibrium. Even if one leaves out this rather special category of disapproval behavior, disapproval points are generally less consistently targeted at shirkers compared to *Punishment*, especially towards the end where group coordination outcomes are more or less settled.

Responses from a debriefing questionnaire shed some light on motives underlying disapproval and punishment behavior. We asked subjects in the *Punishment* and *Disapproval* treatments whether they assigned points to others and believed that this would influence others' behavior; and whether they themselves got assigned points and were influenced by them. In the *Punishment* treatment, the majority of players who got assigned points stated that they responded by increasing their effort choice.[11] By contrast, although many players got assigned points in *Disapproval*, most reported that it had no influence on their choice as "the points have no impact." These responses differ from those in earlier studies

[10] A possible explanation is spitefulness as put forth by Falk, Fehr, and Fischbacher (2005).

[11] A typical statement would be "I had to choose higher numbers as I would lose money otherwise."

suggesting that expressing disapproval with others' behavior is sufficient to make them reconsider their actions. For example, in Lopez-Perez and Vorsatz (2010), subjects who know they may *ex post* receive a disapproving message from their partner tend to cooperate more in a prisoner's dilemma game. Such considerations are expressed only by a handful of our subjects. Therefore, the questionnaire responses strengthen our conclusions from the data analysis, namely that disapproval is in our setting insufficient to consistently improve coordination patterns, whereas monetary punishment and especially its consequences for shirkers seem to be able to achieve that.

Given that the percentage of punishers as well as the average number of assigned punishment points decrease steadily with rounds, this may be considered as evidence that shirkers understand punishers' motivations quite well. The decrease also fits the idea that punishers are mostly driven by the perspective of influencing shirkers' future behaviors. Yet, in our design it is impossible to distinguish purely instrumental punishment (i.e., sanctions aimed at changing the other player's behavior in subsequent rounds) from punishment induced by social preferences. Indeed, the observed pattern of decreasing punishment and increasing coordination could also be compatible with other punishment motives such as altruism (Fehr and Gächter, 2005), reciprocity or other social preferences.

4.4 Welfare

The fact that *Punishment* leads to higher efficiency does not guarantee that welfare – defined here in a restricted way as subject's total payoff – is improved: The losses due to punishment (to both parties involved) may exceed efficiency gains in the stage game. We therefore study welfare differences across treatments. Figure 5 shows for each treatment the evolution of average payoff as a fraction of the maximum achievable payoff (i.e., 2.60 euros per subject if everyone chooses the highest effort level 7 in a given round). In *Punishment*, we distinguish between payoff in the stage game and profit, i.e., the payoff minus punishment costs (more precisely, the cost of punishing and being punished, if any). For the other treatments, payoff and profit are obviously equal. Starting off at about 40 % in all treatments, the average payoff increases throughout Stage 1 to eventually reach 60 % in *Disapproval*, 58 % in *Baseline* and 54 % in *Punishment* (*Baseline* features a small peak in the middle of Stage 1 that corresponds to the minimum effort peak in Figure 2). The upward trend and the magnitude of the average payoff reflect the improving individual coordination on mostly inefficient equilibria in all treatments.

Turning to Stage 2, the average payoff in *Baseline* initially slightly rises above the level reached at the end of Stage 1 and then stays at that level (the small drop in the last round corresponds to the minimum effort drop in Figure 2). This relatively flat pattern arises because *Baseline* individual and collective coordination outcomes in Stage 2 remain at or quickly return to the outcomes attained at the end of Stage 1. The average payoff in *Punishment* and *Disapproval* initially drops

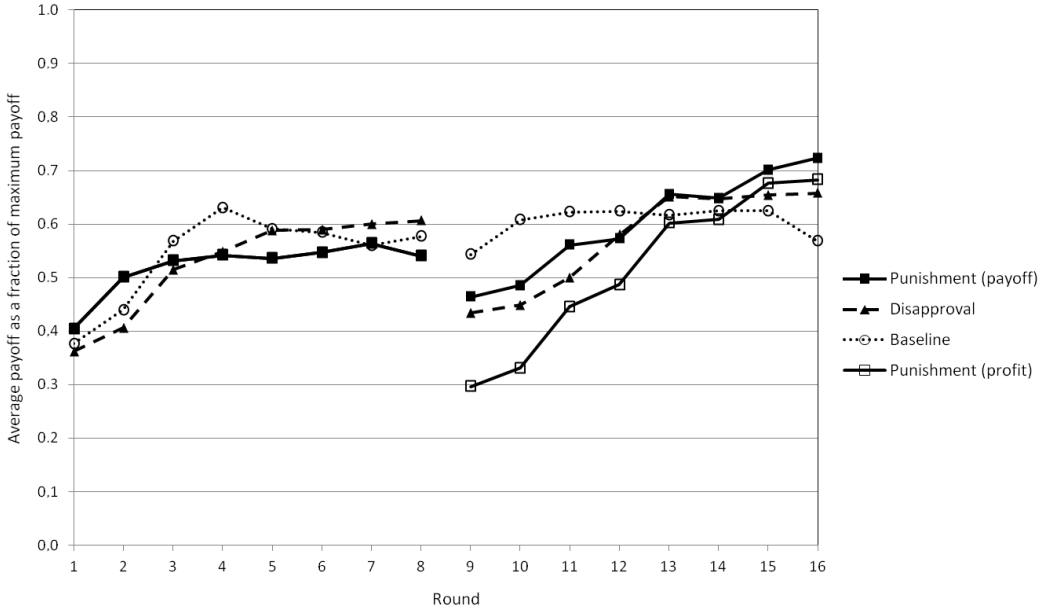

Figure 5: *Payoffs and profits per round and treatment*

substantially to just above the initial round 1 level, subsequently rising steadily and surpassing the *Baseline* average payoff in the second half of Stage 2. The initial drop is due to the extensive initial attempts in both treatments to raise efficiency, with negative consequences for individual coordination outcomes. The subsequent upward trend in average payoff stems from the gradual individual coordination improvements as well as from about half the groups in *Punishment* and two groups in *Disapproval* improving their collective coordination outcomes, as already described above.

Figure 5 further shows that at the beginning of Stage 2, the average profit in *Punishment* is initially only at 30% of the maximum achievable payoff, i.e., 17 percentage points (or 36%) below the average payoff. The welfare consequences of punishment are thus initially considerable. The extent and welfare consequences of punishment decrease over time (i.e., the profit curve converges to the payoff curve) and both the average payoff and average profit eventually reach about 70%, which is higher compared to *Disapproval* and especially to *Baseline*. Nonetheless, the across-treatment welfare differences at the end of Stage 2 are minor compared to the efficiency differences observed in Figures 1 and 2.

In Stage 1, groups' average payoffs do not significantly differ across treatments by the Mann-Whitney U test, both overall and in each round. In Stage 2, average payoffs are significantly higher in *Punishment* compared to *Baseline* in the final round 16 ($p < 0.05$). When including the cost of punishment, profit is significantly lower in *Punishment* compared to *Baseline* in the first two rounds of Stage 2 ($p < 0.10$ in round 9; $p < 0.01$ in round 10) but still significantly higher in the final round 16 ($p < 0.10$). Payoff is also significantly lower in *Disapproval*

compared to *Baseline* in the first three rounds of Stage 2 ($p < 0.05$ in rounds 10 and 11; $p < 0.10$ otherwise).

Similar to the effort-choice comparisons in section 4.2, we also report t-tests from OLS estimation.[12] In Stage 1, payoff differences are not significant across treatments both overall and in each round. In Stage 2, payoff is significantly higher in *Punishment* than in *Baseline* in the final round 16 ($p < 0.10$). When including the cost of punishment, profit is significantly lower in *Punishment* than in *Baseline* in the first three rounds of Stage 2 ($p < 0.05$ in round 9; $p < 0.01$ in round 10; $p < 0.10$ in round 11). Other across-treatment differences are not significant. Hence the parametric t-tests yield slightly weaker results but otherwise confirm the picture in Figure 5, most importantly that payoffs in Stage 2 are initially lower in *Punishment* compared to *Baseline* but the pattern eventually reverses.

	Treatment	Stage 1-2	Round 1-9	Round 2-10	Round 3-11	Round 4-12	Round 5-13	Round 6-14	Round 7-15	Round 8-16
Average payoff change (in p.p.)	Punishment (payoff)	8.07 ^{tt,ss}	5.93 ^{t}	-1.60 $_{bb}$	2.96	3.13	11.94 $^{tt,ss}_{bb}$	10.18 ^{tt,ss}	13.70 ^{ttt,sss}	18.35 $^{ttt,sss}_{bbb,dd}$
	Punishment (profit)	-0.46	-10.82 $^{tt,ss}_{bbb,dd}$	-17.03 $^{ttt,ss}_{bbb,ddd}$	-8.61 $_{b}$	-5.49	6.61	6.21	11.18 ^{ttt,sss}	14.30 $^{tt,ss}_{bbb}$
	Disapproval	4.49	7.21	4.33	-1.44	3.21	6.33 ^{s}	5.77 ^{s}	5.37 ^{ss}	5.13
	Baseline	6.31 ^{ttt,ss}	16.71 ^{tt}	16.83 ^{ttt,ss}	5.41	-0.72	2.52 ^{tt,ss}	4.09 ^{tt,ss}	6.49 ^{s}	-0.84
Fraction of groups with an average-payoff increase	Punishment (payoff)	0.73	0.67 $_{b}$	0.58 $_{bb}$	0.67	0.75	0.83	0.75	0.75 $_{b,dd}$	0.83 $_{bbb,dd}$
	Punishment (profit)	0.53	0.33 $_{bbb,d}$	0.17 $_{bbb,dd}$	0.42	0.58	0.58	0.67	0.75 $_{b,dd}$	0.75 $_{bbb,d}$
	Disapproval	0.50	0.58	0.58	0.50	0.58	0.50	0.42	0.50	0.33 $_{b}$
	Baseline	0.55	0.88	0.88	0.88	0.25	0.50	0.63	0.38	0.00

The "t" and "s" superscripts denote a significant difference across stages or across a round-pair (see the top row), using an appropriate t-test and Wilcoxon signed-rank test, respectively, as described in section 4.4. The "b" resp. "d" symbols denote a significant difference across stages or across a round-pair between the treatment directly above the symbol and *Baseline* resp. *Disapproval*, using a t-test (in the first block), or Mann-Whitney U test (in the second block). Significance levels are 1%, 5% and 10% for three, two and one superscripts or symbols of a kind in a given cell.

Table 6: *Between-stage and between-round welfare changes in each treatment*

The first block of results in Table 6 displays payoff changes (i.e., welfare gains and losses) between Stages 1 and 2. From about the same level of 52-54% of the maximum achievable payoff in Stage 1, average payoff in Stage 2 increases by 8.1 percentage points (0.21 euros) in *Punishment*, 4.5 percentage points (0.12 euro) in *Disapproval*, and 6.3 percentage points (0.16 euro) in *Baseline*. The overall welfare gains in *Punishment* and *Baseline* are significant by both the t-test described above and the Wilcoxon signed-rank test applied to groups' average payoffs. In *Punishment*, the overall profit (including punishment costs) decreases not significantly by 0.5 percentage points (0.01 euro). In the last four round-pairs, the welfare gains in *Punishment* reach over 10 percentage points and are

[12] We regress individual payoffs on treatment dummies interacted with a stage dummy or round dummies. The estimations are based on a panel of 256 subjects with 16 rounds of payoffs each. As above, we use the cluster-robust estimator of variance allowing for intra-group correlation of observations. Other estimation details are identical to the estimation for effort choice changes.

significant by the t-test as well as the Wilcoxon signed-rank test applied to groups' average payoffs. Because of high punishment costs, *Punishment* initially features relatively large profit decreases that are significant in the first two round-pairs; the pattern eventually reverses to reach significant profit increases in the last two round-pairs. *Disapproval* generally features small welfare gains throughout, while *Baseline* features large and significant welfare gains in the first two round-pairs. In the last four round-pairs, the welfare gains in *Disapproval* and *Baseline* are often significant but their magnitude is much smaller compared to *Punishment*, in the last two round-pairs even if punishment costs are included.

The first block also presents t-tests of treatment effects for payoff changes.[13] Due to the large initial welfare gains in *Baseline*, the treatment effect between *Punishment* and *Baseline* is negative in the first three round-pairs, significantly so in the second round-pair. When including punishment costs, *Punishment* fares even worse in the first three round-pairs, not only compared to *Baseline* but also compared to *Disapproval*. This pattern then gradually reverses and in the fifth round-pair, the treatment effect between *Punishment* and *Baseline* becomes significantly positive. The strongest positive treatment effect is observed in the last round-pair where the welfare gains are significantly higher in *Punishment* compared to both *Baseline* and *Baseline*, in the former case even if punishment costs are included.

The second block in Table 6 complements the first one by displaying the fraction of groups with an average payoff increase (i.e., welfare gain) between stages. In *Punishment*, the fraction rises gradually from seven out of 12 groups to eventually reach 10 out of 12 groups in the last round-pair. When including punishment costs, the fraction of groups with an average profit increase is initially only four out of 12 groups but eventually rises to nine out of 12 groups. In *Disapproval*, seven out of 12 groups initially register an average payoff increase but the fraction falls gradually over time to reach four out of 12 groups in the last round-pair. In *Baseline*, seven of the eight groups initially register an average payoff increase but the fraction then falls unevenly, with 3 groups respectively no group featuring a welfare gain in the penultimate respectively the last round-pair. The second block also presents another set of treatment effect tests, comparing groups' average-payoff changes across stages and across round-pairs by the Mann-Whitney U test. Confirming the general picture from the first block, the treatment effect between *Punishment* and *Baseline* is significantly negative in the first two round-pairs. When including punishment costs, the negative treatment effect is even more pronounced and significant, also between *Punishment* and *Disapproval*. In the last two round-pairs, by contrast, the welfare gains are significantly higher in *Punishment* compared to both *Baseline* and *Disapproval*, even when punishment costs are included.

[13] We run OLS regressions of within-subject payoff changes on treatment dummies, and their interaction with round-pair dummies whenever performing separate tests for each round-pair. The estimations are based on a panel of 256 subjects with 8 payoff changes each.

More broadly, our data reveal that, first, *Punishment* fosters efficiency in a robust and stable way, both in comparison with Stage 1 and the other treatments in Stage 2, whereas *Disapproval* seem to have only a transient and limited effect. Second, this seems to be achieved by a rather substantial initial incidence of voluntary sanctions imposed mainly by high-effort players on low-effort ones. And third, the efficiency gains associated with introducing the sanctioning mechanism are initially negative (partly due to the high punishment costs) but ultimately turn out significantly positive. Hence after the initial episode of miscoordination and adjustment to the new conditions, and "coordination costs" incurred by using the sanctions, the sanctions can substantially improve coordination outcomes.

5 Discussion and concluding remarks

Our findings raise several issues which we structure around four main points. First, our results suggest that communication - more precisely, *ex post* disapproval communication - may not be a strong enough efficiency-enhancing coordination device in particularly adverse conditions, such as relatively large group size, anonymity of actions, and prior history of inefficient coordination. By contrast, punishment opportunities seem more powerful under the same conditions, despite the fact that they imply a monetary cost to their user (i.e., the punisher) unlike cost-free disapproval. In this sense, our findings resemble the effect of punishment in cooperation games (e.g., Fehr and Gaechter, 2000), thus possibly contributing to an explanation of why cooperation arises in real economic settings (Ostrom, Walker, and Gardner, 1992). In particular, a common explanatory factor in the two types of strategic interactions may be that organizations attain relatively high efficiency due to individuals fearing (or experiencing) others' retaliation in case of shirking or free-riding. Individuals in organizations generally have various opportunities to (*ex post*) retaliate, more or less formally, e.g., by hiding work-related information, refusing help, malevolent gossiping, exposure of shirking or free-riding actions, and so forth. The nature of these retaliation opportunities clearly relies on being able to observe others' effort, which is likely the case at least in part in stable organizations; the retaliation is also likely to be less anonymous compared to our as well as other experimental settings. As the examples suggest, although the nature of the retaliation is non-pecuniary, its short- and long-term consequences in real organizations may be quite severe and hence more likely resembling our punishment rather than our disapproval mechanism. This is not to say, however, that "pure" *ex post* communication does not work *per se*, or that our results necessarily contradict those of Dugar (2010) or other studies documenting a stronger effect of (predominantly *ex ante*) communication. Yet, our results suggest that monetary sanctioning opportunities provide a more powerful device for recovering from a history of low-efficiency coordination.

A related methodological point would be that it may generally be more appropriate to test the (relative) power of efficiency-enhancing coordination devices after allowing for a history of low-efficiency coordination, as has been demonstrated in cooperation settings (e.g., Fehr and Gaechter, 2000). In coordination

settings such as ours, the reasons are even more imminent due an even stronger evidence of high initial (first-round or one-shot) efficiency as well as strong path-dependence of coordination outcomes. Consequently, a rather mild initial nudge provided by an otherwise weak coordination device may be sufficient to improve efficiency (which in any case would have been close to an upper bound without the device) and to sustain it. Hence the relative strength of different mechanisms is hard to assess. Furthermore, since only a minor change in initial behavior or expectations of players may be required, what is tested may in fact not be the power of the devices in enhancing coordination *per se*, but rather their power in a "second order coordination game": As in a beauty contest, rather than finding out whether the tested mechanism is soundly efficiency-improving, one might merely observe whether participants believed so. Put differently, one may observe the effect of self-fulfilling prophecies rather than a stable and reliable effect of the newly implemented mechanism. On the contrary, in our setting that (in most groups) necessitates a strong recovery from previously inefficient coordination, the mechanism presumably needs to exert influence on the coordination propensities themselves, and the self-fulfilling prophecy concern is mitigated by experiencing a history of inefficient coordination. In this respect, we observe that a pure restart effect is not strong enough, either in the baseline or the disapproval treatment. Moreover, the positive kick given by disapproval opportunities is initially very similar to (or even slightly stronger than) the effect of punishment opportunities, suggesting that subjects did not hold different expectations regarding the power of the two devices. A potentially important implication for future studies is that the discriminatory power of various efficiency-enhancing mechanisms may crucially depend on the conditions in which the mechanisms are implemented (naturally depending on the research question of interest).

Our third point gathers issues about the motivation behind voluntary punishment, which is a critical aspect of the broader applicability of the proposed coordination device. In coordination as well as cooperation settings, while punishing shirkers, respectively free-riders, is costly and hence constitutes a second-order public goods game, this issue seems at least partly overcome in experimental settings given the ample evidence of low-effort defectors being punished and of punishment being mostly targeted at them. Yet, unlike in cooperation games, it is not straightforward to refer to social preferences such as inequity aversion and (perhaps more widely acknowledged) reciprocity-based preferences as motives driving punishment in coordination settings.[14] In particular, free-riding intentions in cooperation games are not difficult to interpret as being driven by self-interest, independent of one's beliefs about others' behavior, generating negative reciprocity. By contrast, exerting a low effort in a coordination situation such as our minimum effort game is somehow linked to heterogeneity of beliefs: Since a shirker partly hurts herself, her choice likely reflects wrong beliefs regarding

[14]Based on a comprehensive investigation of punishment motives in cooperation settings, Falk, Fehr, and Fischbacher (2005) conclude that inequity aversion models can only explain a small share of punishing choices (around 10 %), whereas reciprocity (Falk and Fischbacher, 2006; Charness and Rabin, 2002; Rabin, 1993) seems to be the main driver of punishment behavior.

others' choices rather than mean intentions.[15] Thus it is not clear why an individual motivated by reciprocity would punish low-effort players. To provide a reciprocity-based explanation, one may need a broader definition, for instance arguing that punishment of shirkers is justified by them hurting the group as a whole.[16] Another potential explanation is that shirking is viewed not as being due to wrong beliefs (i.e., mis-coordination) but rather as an intentional choice, i.e., a deliberate decision to hurt others (even at own cost), or in a more charitable interpretation, to earn more than others. The underlying motive for punishment could then be spitefulness or competitiveness.[17] Punishment behavior could also be seen as purely instrumental, that is, motivated by the building of reputation in initial rounds to achieve higher efficiency in later ones. This would be consistent with our finding that punishment incidence decreases steadily throughout Stage 2, yet this might simply be driven by the fact that coordination improves over time as well. Our setting therefore provides only very limited discriminatory power as regards motives for punishment. Alongside examining beliefs of low-effort players, investigating these motives in coordination contexts certainly constitutes an interesting area of future research.

Finally, we suggest future line of research about the effect of punishment opportunities in more general coordination games, i.e., not Pareto-ranked coordination games. If the motive for punishment is purely instrumental (inducing others to raise their effort) or based on a group-level reciprocity, it is possible that sanctioning opportunities may improve stability of an arbitrary equilibrium: If players who occasionally or randomly deviate from an equilibrium face retaliation and its consequences, such deviations may become less frequent. Deviations, even if only erroneous, are made much more costly than in the absence of such punishment. In particular, this may have interesting consequences in the forming of conventions or social norms, even though the roots of their transgression may not actually be malevolence or bad intentions. The specific conditions under which informal sanctions may help coordination at a broader level are of course an empirical question, but may yield interesting insights into the decentralized enforcement of social norms and conventions, a question critical to economics in many respects (Knack and Keefer, 1997).

References

Anderson, C., and L. Putterman (2006): "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution

[15] In case of several shirkers (i.e., lowest-effort players), none of them individually hurt higher-effort players, and arguably, the latter individually had wrong expectations rather than the former.

[16] This is sometimes referred to as indirect reciprocity, in the sense that an individual retaliates against (or is benevolent towards) someone who was malevolent (resp. benevolent) towards a third party (Nowak and Sigmund, 2005)

[17] Interestingly, Falk, Fehr, and Fischbacher (2005) suggest that such motives, in particular spitefulness, have so far been overlooked and may explain why defectors tend to punish cooperators in cooperation games. Such preferences (i.e., their anticipation) could of course reinforce defection in cooperation games as well as shirking in Pareto-ranked coordination games.

mechanism," *Games and Economic Behavior*, 54(1), 1–24.

Becker, G., and K. Murphy (1992): "The division of labor, coordination costs, and knowledge," *The Quarterly Journal of Economics*, 107(8), 1137–1160.

Blume, A., and A. Ortmann (2007): "The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria," *Journal of Economic Theory*, 132, 274–290.

Brandts, J., and D. Cooper (2006): "A change would do you good: An experimental study on how to overcome coordination failure in organizations," *American Economic Review*, 96(3), 669–693.

——— (2007): "It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure," *Journal of the European Economic Association*, 5(6), 1223–1268.

Cachon, G., and C. Camerer (1996): "Loss-avoidance and forward induction in experimental coordination games," *Quarterly Journal of Economics*, 111(1), 165–194.

Camerer, C. F., and M. Knez (1994): "Creating 'expectational assets' in the laboratory: 'Weakest-link' coordination games," *Strategic Management Journal*, 15, 109–109.

Carpenter, J. (2007): "The demand for punishment," *Journal of Economic Behavior and Organization*, 62(4), 522–542.

Casari, M. (2005): "On the design of peer punishment experiments," *Experimental Economics*, 8, 107–115.

Charness, G. (2000): "Self-serving cheap talk: A test Of Aumann's conjecture," *Games and Economic Behavior*, 33, 177–194.

Charness, G., and M. Rabin (2002): "Understanding social preferences with simple tests," *Quartely Journal of Economics*, 117, 817869.

Chaudhuri, A., A. Schotter, and B. Sopher (2009): "Talking ourselves to efficiency: Coordination in inter-generational minimum effort games with private, almost common and common knowledge of advice," *Economic Journal*, 119, 91–121.

Cooper, R., D. DeJong, R. Forsythe, and T. W. Ross (1992): "Communication in coordination games," *Quarterly Journal of Economics*, 107(2), 739–771.

Devetag, G., and A. Ortmann (2007): "When and why? A critical survey on coordination failure in the laboratory," *Experimental Economics*, 10(3), 331–344.

Dugar, S. (2010): "Nonmonetary sanctions and rewards in an experimental coordination game," *Journal of Economic Behavior and Organization*, 73(3), 377–386.

Engelmann, D., and H.-T. Normann (2010): "Maximum effort in the minimum-effort game," *Experimental Economics*, 13(3), 249–259.

Falk, A., E. Fehr, and U. Fischbacher (2005): "Driving forces behind informal sanctions," *Econometrica*, 73(6), 2017–2030.

Falk, A., and U. Fischbacher (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293–315.

Fatas, E., T. Neugebauer, and J. Perote (2006): "Within-team competition in the minimum effort coordination game," *Pacific Economic Review*, 11(2), 247–266.

Fehr, E., and S. Gächter (2005): "Altruistic punishment in humans," *Nature*, 415, 137–140.

Fehr, E., and S. Gaechter (2000): "Cooperation and punishment in public goods experiments," *American Economic Review*, 90, 980–994.

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

Galbiati, R., K. Schlag, and J. van der Weele (2009): "Can sanctions induce pessimism? An experiment," *Working Paper 1150, Universitat Pompeu Fabra*.

Goeree, J. K., and C. A. Holt (2005): "An experimental study of costly coordination," *Games and Economic Behavior*, 51(2), 349–364.

Greiner, B. (2004): "An online recruitment system for economic experiments," in *Forschung und wissenschaftliches Rechnen 2003*, ed. by K. Kremer, and V. Macho. Gttingen : Ges. fr Wiss. Datenverarbeitung,.

Horwitz, A. (1990): *The logic of social control.* New York : Plenum Press.

Kezdi, G. (2004): "Robust standard error estimation in fixed-effects panel models.," *Hungarian Statistical Review Special*, 9, 96–116.

Knack, S., and P. Keefer (1997): "Does social capital have an economic payoff? A cross-country investigation," *The Quarterly Journal of Economics*, 112(4), 1251–1288.

Lopez-Perez, R., and M. Vorsatz (2010): "On approval and disapproval: Theory and experiments," *Journal of Economic Psychology*, 31(4), 527–541.

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval (2003): "Monetary and nonmonetary punishment in the voluntary contributions mechanism," *American Economic Review*, 93, 366–380.

Nowak, M., and K. Sigmund (2005): "The evolution of indirect reciprocity," *Nature*, 437, 1291–1298.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants with and without a sword: Self governance is possible," *American Political Science Review*, 86, 404–417.

Rabin, M. (1993): "Incorporating fairness into game theory and economics," *American Economic Review*, 83, 1281–1302.

Rogers, W. H. (1993): "sg17: Regression standard errors in clustered samples.," *Stata Technical Bulletin*, 13, 19–23.

Romero, J. (2011): "The effect of hysteresis on equilibrium selection in coordination games," *Purdue University Economics Working Papers 1265, Purdue University, Department of Economics*.

Van Huyck, J., R. Battalio, and R. Beil (1990): "Tacit coordination games, strategic uncertainty, and coordination failure," *American Economic Review*, 80(1), 234–248.

——— (1991): "Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games," *Quarterly Journal of Economics*, 106(3), 885–911.

Van Huyck, J., R. Battalio, and F. Rankin (2007): "Evidence on learning in coordination games," *Experimental Economics*, 10(3), 205–220.

Weber, R., C. Camerer, Y. Rottenstreich, and M. Knez (2001): "The illusion of leadership: Misattribution of cause in coordination games," *Organizational Science*, 12(5), 582–598.

Jena Economic Research Papers 2012 - 030

Appendix to *Punishment Fosters Efficiency in the Minimum Effort Coordination Game*

Fabrice Le Lec

Astrid Matthey

and Ondřej Rydval*

June 19, 2012

Group-level data

Our supplementary analysis is of qualitative nature and focuses on group behavior. Broadly in line with our examination of treatment effects, we discuss across-treatment differences in groups' efficiency and welfare gains (or losses) in Stage 2 relative to Stage 1. The discussion is based on Tables A1, A2 and A3, which show for each group the evolution of minimum, average and maximum effort, average payoff (or profit) and punishment and disapproval behavior. Within each treatment, groups are ordered by their efficiency gains between stages, namely descending in average-effort gain between the last four rounds of each stage (the ordering is inconsequential and we make a couple of exceptions for ease of exposition).

Baseline

In *Baseline*, six of the eight groups, B3-B8, make various coordination attempts in the first several rounds of Stage 1, at least to the extent that several group members choose a high effort level. However, these attempts are always undercut by shirkers (i.e., subjects with the lowest effort in the group) choosing in most cases the lowest effort level, which seems to drive the groups to eventually coordinate

*Le Lec: Lille Economics and Management UMR CNRS 8179, Catholic University of Lille, Lille, France. fabrice.lelec@icl-lille.fr, phone: +33 359 56 69 75. (corresponding author)
Matthey: Max Planck Institute of Economics, Jena, Germany. matthey@econ.mpg.de, phone: +49 3641 686644
Rydval: Max Planck Institute of Economics, Jena, Germany, and CERGE-EI, Charles University Prague and Academy of Sciences of the Czech Republic, Prague, Czech Republic, E-mail: rydval@econ.mpg.de, phone: +49 3641 686641

on the least efficient equilibrium.[1] The groups could be described as having a very low "collective coordination potential" for Stage 2 – in the sense that they fail to solve the collective coordination problem, but also a large scope for efficiency gains (and little or no scope for efficiency losses). In fact, except for a variety of small and temporary restart effects and a couple of other sporadic attempts to raise efficiency, groups B3-B8 remain at the least efficient equilibrium throughout Stage 2, registering either no efficiency changes or small efficiency losses (comparing the average effort in the last four rounds of each stage). With a couple of minor exceptions (see the first half of Stage 1 for group B8), the groups' average payoff increases up to 54% of the maximum achievable payoff, i.e., the payoff achieved in the least efficient equilibrium.

		round # (Stage 1)								choices in	round # (Stage 2)								choices in
Group		1	2	3	4	5	6	7	8	round 8	9	10	11	12	13	14	15	16	round 16
B1	Min e	4	4	4	4	4	4	1	1	1114	4	4	4	4	4	4	4	1	1444
	Avg e	5.4	5.5	4.5	4.0	4.1	4.3	3.8	3.0	4445	4.4	4.1	4.1	4.1	4.1	4.0	4.0	3.6	4444
	Max e	7	7	5	4	5	6	5	5		5	5	5	5	5	4	4	4	
	Payoff	66	65	73	77	76	75	33	38		74	76	76	76	76	77	77	34	
B2	Min e	3	3	5	7	7	7	7	7	7777	7	7	7	7	7	7	7	7	7777
	Avg e	6.0	6.1	6.8	7.0	7.0	7.0	7.0	7.0	7777	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7777
	Max e	7	7	7	7	7	7	7	7		7	7	7	7	7	7	7	7	
	Payoff	46	45	71	100	100	100	100	100		100	100	100	100	100	100	100	100	
B3	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	3.4	3.1	1.8	1.4	1.1	1.0	1.0	1.0	1111	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.3	1113
	Max e	7	7	3	3	2	1	1	1		1	1	1	1	1	1	1	3	
	Payoff	36	38	48	51	53	54	54	54		54	54	54	54	54	54	54	52	
B4	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.5	3.3	1.1	1.0	1.0	1.0	1.0	1.0	1111	1.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1111
	Max e	7	7	2	1	1	1	1	1		5	1	1	1	1	1	1	1	
	Payoff	27	37	53	54	54	54	54	54		50	54	54	54	54	54	54	54	
B5	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	5.4	3.6	1.6	1.0	1.0	1.8	1.0	1.0	1111	3.4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1111
	Max e	7	7	3	1	1	7	1	1		7	1	1	1	1	1	1	1	
	Payoff	20	34	49	54	54	48	54	54		36	54	54	54	54	54	54	54	
B6	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.9	2.3	1.4	1.0	1.8	1.8	1.0	1.0	1111	1.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1111
	Max e	7	5	3	1	7	7	1	1		7	1	1	1	1	1	1	1	
	Payoff	24	44	51	54	48	48	54	54		47	54	54	54	54	54	54	54	
B7	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.5	3.6	1.8	1.4	1.6	1.8	1.8	1.0	1111	2.4	1.8	1.0	1.0	1.0	1.0	1.0	1.0	1111
	Max e	7	7	3	4	5	6	7	1		7	7	1	1	1	1	1	1	
	Payoff	27	34	48	51	49	48	48	54		43	48	54	54	54	54	54	54	
B8	Min e	3	3	3	3	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.9	4.8	4.0	3.6	2.9	2.8	1.3	1.0	1111	4.0	1.9	1.1	1.0	1.8	1.0	1.0	1.0	1111
	Max e	7	7	5	4	4	4	2	1		7	3	2	1	7	1	1	1	
	Payoff	55	56	62	64	39	40	52	54		31	47	53	54	48	54	54	54	

Min e, *Avg e* and *Max e* stands, respectively, for minimum effort, average effort and maximum effort in the group under consideration.

Table A1: *Group data for* Baseline

[1] For these groups as well as for similarly behaving groups in the other treatments, the movement towards the least efficient equilibrium does not seem to be driven by an "endgame" effect, at least not primarily so.

The remaining two *Baseline* groups, B1 and B2, build up a different kind of collective coordination potential for Stage 2 compared to groups B3-B8, but otherwise experience a similarly strong degree of between-stage inertia. Group B2 manages to coordinate on the most efficient equilibrium in the last five rounds of Stage 1 (and hence achieves 100% payoff), which clearly represents the best collective coordination potential for Stage 2 but also no scope for efficiency gains (i.e., only a large scope for efficiency losses). It therefore seems a valuable achievement - also given the extensive prior evidence of inefficient coordination in the baseline minimum-effort game - that the group sustains coordination on the most efficient equilibrium throughout Stage 2.

Finally, group B1 initially sets out for achieving high efficiency and welfare and reaches the level-4 equilibrium by round 4, but subsequent attempts to further raise efficiency are hindered by shirkers. The group comprises effort choices 4 and 5 as well as three lowest-effort choices at the end of Stage 1 (bringing the average payoff down to 38%), which could be described as an intermediate degree of collective as well as individual coordination potential and implies a scope for both efficiency gains and losses. In Stage 2, the group quickly solves the individual coordination problem by converging to the level-4 equilibrium (except for a small downward endgame effect), hence returning to the efficiency and payoff level reached in the middle of Stage 1. What might have prevented further efficiency gains is that no-one chooses the highest two effort levels in Stage 2 to signal a desire to raise efficiency. To sum up, *Baseline* features no group with a substantial efficiency gain (or loss) between stages, although the scope for efficiency gains is large for most groups.

Disapproval

For the most part, *Disapproval* features a similarly strong degree of between-stage inertia at the group level as does *Baseline*. Two of the 12 groups, D3 and D4, are similar to *Baseline* group B2 in that they manage to coordinate on the most efficient equilibrium in the last several rounds of Stage 1 and to sustain the coordination throughout Stage 2. These are also the only groups that do not assign any disapproval points.[2] Eight *Disapproval* groups, D5-D12, in both stages share the fate with *Baseline* groups B3-B8. Particularly, they make various coordination attempts in the first several rounds of Stage 1 but eventually converge or in a couple of cases almost converge to the lowest-effort equilibrium, again seemingly due to shirkers choosing a low and in most cases the lowest effort level throughout the stage. In Stage 2, despite larger and more durable restart effects compared to *Baseline* - especially in D9 and D11 - which are costly in terms of average payoff, all the groups eventually converge or in one case almost converge to the least efficient equilibrium.

[2] As an exception, one subject in group D3 assigns the maximum number of points to all other members in the final round. This action is of course inconsequential and the reasons behind it are unclear given the fully efficient coordination achieved by the group.

		round # (Stage 1)								choices in	round # (Stage 2)								choices in
Group		1	2	3	4	5	6	7	8	round 8	9	10	11	12	13	14	15	16	round 16
D1	Min e	1	1	1	1	1	1	1	1	1111	1	3	4	6	7	7	7	7	7777
	Avg e	5.1	5.5	4.0	1.4	1.0	1.0	1.3	1.0	1111	5.4	6.1	6.5	6.9	7.0	7.0	7.0	7.0	7777
	Max e	7	7	7	3	1	1	3	1		7	7	7	7	7	7	7	7	
	Pts/subj										87/6	64/6	64/6	28/6	.	.	.	.	
	Payoff	22	19	31	51	54	54	52	54		20	45	58	86	100	100	100	100	
D2	Min e	1	1	1	1	1	1	1	1	1111	3	3	3	3	3	2	2	2	2222
	Avg e	5.5	5.1	2.3	1.8	1.0	1.1	1.0	1.0	1111	5.6	6.1	6.0	5.0	3.6	3.3	2.5	2.1	2223
	Max e	7	7	4	6	1	2	1	1		7	7	7	7	7	5	3	3	
	Pts/subj										91/7	84/7	101/7	170/7	169/7	171/6	193/7	180/5	
	Payoff	19	22	44	48	54	53	54	54		49	45	46	54	64	52	58	61	
D3	Min e	3	5	5	5	7	7	7	7	7777	7	7	7	7	7	7	7	7	7777
	Avg e	5.5	6.3	6.5	6.8	7.0	7.0	7.0	7.0	7777	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7777
	Max e	7	7	7	7	7	7	7	7		7	7	7	7	7	7	7	7	
	Pts/subj										.	.	.	.	.	.	.	42/1	
	Payoff	50	75	73	71	100	100	100	100		100	100	100	100	100	100	100	100	
D4	Min e	4	5	6	7	7	7	7	7	7777	7	7	7	7	7	7	7	7	7777
	Avg e	6.1	6.6	6.9	7.0	7.0	7.0	7.0	7.0	7777	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7777
	Max e	7	7	7	7	7	7	7	7		7	7	7	7	7	7	7	7	
	Pts/subj										.	.	.	.	.	.	.	.	
	Payoff	61	72	86	100	100	100	100	100		100	100	100	100	100	100	100	100	
D5	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	5.6	4.4	1.8	1.3	1.4	1.1	1.5	1.1	1112	4.3	4.0	2.8	1.3	1.4	1.1	1.3	1.3	1122
	Max e	7	7	5	2	3	2	4	2		7	7	7	2	3	2	2	2	
	Pts/subj										158/7	201/8	221/8	233/6	196/6	224/8	210/7	208/7	
	Payoff	18	28	48	52	51	53	50	53		29	31	40	52	51	53	52	52	
D6	Min e	3	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	5.5	4.5	2.5	1.5	1.3	1.3	1.1	1.0	1111	3.6	3.0	2.1	1.4	1.3	1.1	1.3	1.0	1111
	Max e	7	7	5	3	2	3	2	1		7	5	5	3	2	2	3	1	
	Pts/subj										144/6	157/6	188/7	191/5	169/5	120/3	129/5	126/3	
	Payoff	50	27	42	50	52	52	53	54		34	38	45	51	52	53	52	54	
D7	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	3.9	1.9	1.3	1.0	1.1	1.0	1.0	1.0	1111	3.9	4.1	1.5	1.5	1.0	1.0	1.0	1.0	1111
	Max e	7	4	3	1	2	1	1	1		7	7	4	5	1	1	1	1	
	Pts/subj										121/7	144/6	190/6	240/7	182/5	189/5	189/5	273/7	
	Payoff	32	47	52	54	53	54	54	54		32	30	50	50	54	54	54	54	
D8	Min e	3	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	6.1	5.8	4.9	1.9	1.0	1.0	1.0	1.0	1111	5.1	5.0	3.4	1.8	1.0	1.0	1.0	1.0	1111
	Max e	7	7	7	5	1	1	1	1		7	7	7	7	1	1	1	1	
	Pts/subj										112/8	120/8	187/7	258/7	252/6	252/6	238/6	238/6	
	Payoff	45	17	24	47	54	54	54	54		22	23	36	48	54	54	54	54	
D9	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	5.0	4.4	1.9	1.1	1.1	1.0	1.0	2.0	1137	4.4	4.8	5.5	4.8	1.0	1.0	1.0	1.0	1111
	Max e	7	7	7	2	2	1	1	7		7	7	7	7	1	1	1	1	
	Pts/subj										131/7	90/5	72/6	105/7	168/4	168/4	168/4	210/5	
	Payoff	23	28	47	53	53	54	54	46		28	25	19	25	54	54	54	54	
D10	Min e	3	3	2	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.6	4.1	3.4	2.6	1.9	1.5	1.1	1.0	1111	4.1	2.3	1.5	1.1	1.0	1.0	1.0	1.0	1111
	Max e	7	7	5	5	4	3	2	1		7	4	3	2	1	1	1	1	
	Pts/subj										122/7	189/6	180/6	123/4	112/4	196/6	199/7	194/7	
	Payoff	57	61	51	41	47	50	53	54		30	44	50	53	54	54	54	54	
D11	Min e	1	2	3	1	1	1	1	1	1111	2	1	1	1	1	1	1	1	1111
	Avg e	5.5	5.1	3.9	4.1	3.0	2.1	1.5	1.0	1111	5.9	6.1	5.8	3.1	1.4	1.0	1.0	1.0	1111
	Max e	7	7	7	7	5	5	3	1		7	7	7	7	3	1	1	1	
	Pts/subj										77/7	61/7	75/7	195/8	270/8	182/5	189/5	182/5	
	Payoff	19	38	63	30	38	45	50	54		32	14	17	38	51	54	54	54	
D12	Min e	2	3	3	3	2	1	1	1	1111	2	2	1	1	1	1	1	1	1111
	Avg e	5.1	5.1	4.6	4.1	3.5	2.9	1.9	1.3	1122	4.1	4.5	3.0	2.8	1.8	1.5	1.0	1.0	1111
	Max e	7	7	7	5	5	4	4	2		7	7	5	7	2	2	1	1	
	Pts/subj										148/8	158/8	219/8	209/8	157/6	209/6	252/6	252/6	
	Payoff	38	53	57	61	50	39	47	52		45	42	38	40	48	50	54	54	

Min e, *Avg e* and *Max e* stands, respectively, for minimum effort, average effort and maximum effort in the group under consideration. *Pts/Subj* stands for the total number of points assigned in the group divided by the number of subjects targeted

Table A2: *Group data for* Disapproval

Both of the remaining *Disapproval* groups, D1 and D2, also converge to the least efficient equilibrium by the end of Stage 1 (in fact already by the middle of the stage). In Stage 2, however, D1 experiences a strong restart effect and converges to the most efficient equilibrium already in the first half of the stage. The strong attempts to raise efficiency are clearly very costly initially (because of a shirker) but pay off later. In D2, by contrast, a similarly strong restart effect seems ultimately insufficient to overcome the influence of shirkers, despite repeated and costly attempts by other members (not always the same ones) to raise efficiency by choosing a high effort level. The group in the end only manages to coordinate on effort level 2. Therefore, group D1 is almost solely responsible for any aggregate efficiency and welfare gain observed in *Disapproval* (again comparing the last four rounds of each stage).

One may wonder what makes groups D1, D2 and D5-D12 arrive at different coordination outcomes in Stage 2 despite having similarly low collective coordination potential prior to the stage. Although the extent and dynamics of assigning disapproval points vary both across and within groups, this variation does not seem to lie behind the observed differences in coordination outcomes. In both D1 and D2, disapproval points are in the initial rounds mostly targeted at shirkers or members with below-average effort in the group, yet only D1 manages to raise efficiency fully and "permanently." Similar targeting of disapproval points occurs in other groups - most notably D8, D9 and D11 - while the remaining groups generally assign points in a much less targeted way, but regardless of these differences, all the D5-D12 groups eventually converge to the least efficient equilibrium. We focus on point assignment in the initial rounds in which most collective coordination attempts occur and shirkers are clearly distinguishable. Later on when collective coordination attempts fail and shirkers become the majority, even the originally strongly targeting groups start assigning points in a non-targeted manner, which is reflected in Table 5 and discussed in Section 4.3. What seems to trigger the various group coordination outcomes is mainly the extent and persistence of shirking behavior.

Punishment

Punishment overall features a much lower extent of between-stage inertia compared to the other treatments, as about half of the 12 groups make substantial collective coordination improvements (i.e., efficiency and welfare gains) between stages. We first discuss the groups that do not make such improvements. Four groups, P9-P12, are similar to *Baseline* groups B3-B8 and *Disapproval* groups D5-D12 since they eventually converge or almost converge to the least efficient equilibrium by the end of both stages, despite large, durable and costly (for the members trying to raise efficiency) restart effects in groups P9 and P12. The low collective coordination success again seems to be driven by shirkers always pressing the minimum effort down to the lowest effort level.

In these as well as other *Punishment* groups (whenever relevant), punishment points are targeted at shirkers - in a more focused way compared to *Disapproval*

		round # (Stage 1)								choices in	round # (Stage 2)								choices in
Group		1	2	3	4	5	6	7	8	round 8	9	10	11	12	13	14	15	16	round 16
P1	Min e	3	3	3	3	3	2	2	2	2222	4	4	5	6	7	7	7	7	7777
	Avg e	4.6	4.4	4.3	3.9	3.5	3.0	2.8	2.5	2235	5.1	5.4	5.9	6.5	7.0	7.0	7.0	7.0	7777
	Max e	7	6	6	5	5	4	4	5		7	7	7	7	7	7	7	7	
	Pts/subj										8/4	14/4	8/4	8/2	.	.	.	.	
	Profit	57	59	60	63	65	54	56	58		57	46	66	77	100	100	100	100	
P2	Min e	3	4	4	4	1	1	1	1	1111	1	1	1	1	4	4	5	6	6666
	Avg e	5.8	5.9	6.0	5.8	3.5	1.6	1.0	1.0	1111	3.9	3.6	3.8	4.5	4.9	5.6	6.1	6.5	7777
	Max e	7	7	7	7	4	4	1	1		7	7	6	7	7	7	7	7	
	Pts/subj										17/3	35/3	21/5	16/7	10/4	10/6	7/5	13/3	
	Profit	48	63	62	63	35	49	54	54		11	1	6	11	56	50	66	70	
P3	Min e	2	2	2	1	1	1	1	1	1111	3	3	3	3	3	4	4	4	4455
	Avg e	4.8	3.5	3.4	2.3	2.3	1.9	1.5	1.3	1122	4.3	5.4	5.4	5.4	5.6	5.5	5.6	5.4	6667
	Max e	7	7	7	4	4	2	2	2		7	7	7	7	7	7	7	7	
	Pts/subj										5/1	10/4	12/3	15/4	11/4	3/1	5/3	.	
	Profit	40	50	51	44	44	47	50	52		52	37	34	29	33	61	57	66	
P4	Min e	2	3	3	2	2	2	2	2	2222	2	3	4	4	5	3	5	6	6666
	Avg e	5.0	4.8	4.6	3.9	3.3	3.3	2.6	2.5	2334	4.6	4.6	4.9	5.5	5.8	5.9	6.0	6.1	6667
	Max e	7	7	7	7	5	6	3	4		7	7	6	6	7	7	7	7	
	Pts/subj										14/4	16/5	13/5	10/5	6/2	8/3	4/2	2/1	
	Profit	38	56	57	47	52	52	57	58		21	34	51	51	70	36	71	88	
P5	Min e	4	4	4	4	4	4	4	4	4444	4	4	7	7	7	7	7	7	7777
	Avg e	6.1	6.4	6.0	4.6	4.6	4.0	4.1	4.1	4445	6.3	6.6	7.0	7.0	7.0	7.0	7.0	7.0	7777
	Max e	7	7	7	7	7	4	5	5		7	7	7	7	7	7	7	7	
	Pts/subj										22/5	11/6	.	.	.	.	.	.	
	Profit	61	59	62	72	72	77	76	76		30	42	100	100	100	100	100	100	
P6	Min e	2	3	2	1	1	1	1	1	1112	3	3	3	3	3	3	3	3	3444
	Avg e	4.4	4.0	3.8	3.4	2.8	2.9	2.4	1.8	2223	3.5	3.9	3.9	3.9	4.0	4.0	3.9	3.9	4444
	Max e	7	7	7	5	4	4	3	3		5	5	4	4	5	5	4	4	
	Pts/subj										6/2	2/1	5/3	4/3	6/2	4/3	5/3	4/3	
	Profit	43	62	48	36	40	39	43	48		57	60	55	57	53	56	55	57	
P7	Min e	1	2	2	2	2	2	2	2	2222	2	2	2	2	2	2	3	3	3333
	Avg e	4.1	3.3	3.8	3.0	3.0	3.1	2.5	2.0	2222	4.0	3.8	3.0	2.8	2.8	2.9	3.0	3.3	3344
	Max e	7	7	7	5	4	4	3	2		7	6	4	4	3	3	3	4	
	Pts/subj										11/5	16/5	12/4	6/1	10/4	8/4	.	.	
	Profit	30	52	48	54	54	53	58	62		32	25	37	47	41	43	69	67	
P8	Min e	4	4	4	4	4	4	4	1	1444	4	4	4	4	4	4	4	4	4444
	Avg e	5.3	5.1	4.9	4.1	4.0	4.0	4.4	4.3	4467	4.9	4.4	4.0	4.0	4.0	4.0	4.0	4.0	4444
	Max e	7	7	7	5	4	4	7	7		7	6	4	4	4	4	4	4	
	Pts/subj										5/1	.	.	.	.	.	.	.	
	Profit	67	68	70	76	77	77	74	29		63	74	77	77	77	77	77	77	
P9	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	5.5	4.6	3.5	1.9	1.0	1.1	1.0	1.0	1111	5.9	5.1	5.0	4.4	3.0	1.9	1.0	1.0	1111
	Max e	7	7	7	4	1	2	1	1		7	7	7	7	7	7	1	1	
	Pts/subj										29/7	34/4	26/4	10/3	.	.	.	.	
	Profit	19	26	35	47	54	53	54	54		0	0	2	15	38	47	54	54	
P10	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.1	2.5	1.5	2.1	1.1	1.1	1.0	1.3	1113	2.8	2.0	1.3	1.0	1.0	1.0	1.0	1.0	1111
	Max e	7	7	3	7	2	2	1	3		7	4	3	1	1	1	1	1	
	Pts/subj										16/3	10/2	5/1	4/1	.	.	.	16/2	
	Profit	30	42	50	45	53	53	54	52		17	32	45	48	54	54	54	33	
P11	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.9	4.5	1.5	1.6	1.4	1.3	1.0	1.1	1112	3.1	2.1	1.0	1.0	1.0	1.1	1.0	1.0	1111
	Max e	7	7	3	6	4	2	1	2		7	7	1	1	1	2	1	1	
	Pts/subj										30/2	4/1	.	.	.	.	.	.	
	Profit	24	27	50	49	51	52	54	53		6	39	54	54	54	53	54	54	
P12	Min e	1	1	1	1	1	1	1	1	1111	1	1	1	1	1	1	1	1	1111
	Avg e	4.4	3.0	2.0	1.0	2.0	1.5	1.8	1.0	1111	5.4	5.5	5.5	4.8	1.8	1.0	1.0	1.0	1111
	Max e	7	7	6	1	7	5	7	1		7	7	7	7	7	1	1	1	
	Pts/subj										7/5	8/4	9/4	9/2	2/2	.	.	.	
	Profit	28	38	46	54	46	50	48	54		10	8	7	18	46	54	54	54	

Min e, *Avg e* and *Max e* stands, respectively, for minimum effort, average effort and maximum effort in the group under consideration. *Pts/Subj* stands for the total number of points assigned in the group divided by the number of subjects targeted

Table A3: *Group data for* Punishment

- and are mostly assigned in the first several rounds or more generally only as long as punishers still see a scope for a collective coordination improvement.[3] Especially in P9 and P12, punishment considerably reduces the shirkers' (as well as the punishers') profit, but this is insufficient to set the momentum toward higher group efficiency.

Group P8 is somewhat similar to *Baseline* group B1 in that it converges from above to the level-4 equilibrium in the first half of Stage 1 but subsequently comprises a wider spectrum of effort choices, and that in Stage 2 the group manages to quickly solve the individual coordination problem by converging to the level-4 equilibrium. As expected, punishment is unnecessary to achieve this outcome which carries no efficiency gain. The only punishment points are assigned (sparingly) in the initial round by the subject choosing effort level 7 as an attempt to raise efficiency by punishing shirkers choosing effort level 4. Next, group P7 converges to the level-2 equilibrium by the end of Stage 1, presumably because shirkers keep the minimum effort at this level throughout the stage. Stage 2 mimics Stage 1, except that the group finally manages to slightly raise efficiency and welfare by (almost) coordinating on effort level 3. Punishment is consistently targeted at shirkers and in most rounds reduces their payoff to zero or by a large fraction, but this is insufficient to further raise efficiency.

We now move to the groups that make substantial efficiency and welfare gains between stages. Group P6 is mostly comprised of effort levels 1 and 2 at the end of Stage 1 since previous attempts to raise efficiency are hindered by shirkers choosing the lowest effort level. In Stage 2, the group does not experience an exceptionally strong restart effect on average, but the minimum effort jumps immediately to 3, which seems critical for being able to coordinate on effort level 4 eventually (except for one remaining shirker hindering a larger welfare improvement). Punishment is targeted at shirkers but is rather scant and insufficient to deter them completely. This also proves hard because not always the same subjects shirk, as is sometimes the case in other groups, too. As for group B1, absence of further efficiency gains in Stage 2 might also stem from lack of high-effort choices signaling a desire to further raise efficiency.

Group P5 converges to the level-4 equilibrium by the end of Stage 1, presumably because shirkers keep the minimum effort at this level throughout the stage. In Stage 2 there is a quick convergence to the most efficient equilibrium, potentially due to several contributing factors: a relatively high collective coordination potential after Stage 1; a strong restart effect with only two shirkers initially remaining at effort level 4; punishment being targeted at them and depleting their round payoff; and the other members being "disciplined" in terms of not lowering their effort in reaction to the short-term shirking. Hence it is not necessarily the punishment *per se* that brings about the coordination success. On the other hand, the strong restart effect might itself be a consequence of (anticipated) punishment. We do

[3] An endgame effect, i.e., punishment of shirkers in the final round, occurs to some extent in group P2 and P10.

not have a closely comparable group in this or the other treatments in terms of coordination potential and restart effect to tell apart the influence of the various contributing factors.

Groups P1 and P4 have in common that they more or less coordinate on effort level 2 by the end of Stage 1, despite repeated attempts to raise efficiency. In Stage 2, P1 experiences a strong restart effect and the minimum effort jumps to 4, followed by a relatively quick convergence to the most efficient equilibrium. The strong restart effect combined with well-targeted punishment and good discipline of the other members not lowering their effort could all lie behind the coordination success. In P4, the minimum effort again restarts at 2 but then it increases more or less gradually and the group eventually manages to coordinate on effort level 6. Punishment is again well targeted, and a faster and more robust coordination success seems hindered by a single shirker who gets repeatedly punished but keeps undercutting the other members. This makes their efficiency enhancement attempts costly, partly negatively influencing their discipline and lessening the coordination momentum.

Groups P2 and P3 converge or in the case of P3 almost converge to the least efficient equilibrium by the end of Stage 1, and they register a substantial efficiency gain in Stage 2 but do not manage to fully coordinate on a particular equilibrium. In P2, shirkers keep their effort at 1 for the first half of Stage 2 which harms welfare of the others, but then the minimum effort jumps to 4 and the group eventually comprises of effort choices 6 and 7 in equal proportions. Shirkers are targeted and punished quite strongly but there are initially always several of them, which may lie behind the discipline of the other members occasionally dropping and the coordination momentum weakening (compared to more successful groups). In P3, the minimum effort in Stage 2 immediately jumps up to 3 but then it only reaches 4, despite repeated (costly) attempts of the other members to further raise efficiency by targeted (though at first rather scant) punishment and high-effort choices. Consequently, the individual coordination problem prevails until the end of Stage 2, with effort choices eventually reaching between 4 and 7.