University of Groningen

The double edge of counter-sanctions

Flache, Andreas; Bakker, Dieko; Maes, Michael; Dijkstra, Jacob

Published in:
Social dilemmas, institutions, and the evolution of cooperation

DOI:
[10.1515/9783110472974-014](https://doi.org/10.1515/9783110472974-014)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Flache, A., Bakker, D., Maes, M., & Dijkstra, J. (2017). The double edge of counter-sanctions: Is peer sanctioning robust to counter-punishment but vulnerable to counter-reward? In B. Jann, & W. Przepiorka (Eds.), *Social dilemmas, institutions, and the evolution of cooperation* (pp. 279-302). De Gruyter Oldenbourg. <https://doi.org/10.1515/9783110472974-014>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Andreas Flache, Dieko Bakker, Michael Mäs, and Jacob Dijkstra

The Double Edge of Counter-Sanctions. Is Peer Sanctioning Robust to Counter-Punishment but Vulnerable to Counter-Reward?

Abstract: Peer sanctioning institutions are powerful solutions to the freerider problem in collective action. However, counter-punishment may deter sanctioning, undermining the institution. Peer-reward can be similarly vulnerable, because peers may exchange rewards for rewards (“counter-reward”) rather than enforce contributions to the collective good. Based on social exchange arguments, we hypothesize that peer-reward is vulnerable in a repeated game where players are fully informed about who rewarded them in the past. Social preference arguments suggest that peer-punishment is robust under the same conditions. This contrast was tested in an experiment in which counter-sanctioning was precluded due to anonymity of enforcers in one treatment and allowed in another treatment by non-anonymity of enforcers. This was done both for a reward and for a punishment institution. In line with the exchange argument, non-anonymity boosted reward-reward exchanges. Punishment was only somewhat reduced when enforcers were not anonymous. In contrast with previous experiments, we found no effects of counter-sanctioning on contributions. Thus, non-anonymity did not undermine the effectiveness of the peer sanctioning institutions in our experiments, neither for reward nor for punishment. Our results suggest that previous claims about the vulnerability of peer-punishment to counter-punishment may not generalize to non-anonymous repeated interactions.

1 Introduction

Human societies depend on the successful provision of collective goods, which typically require the contribution of many to be produced.¹ However, as Olson (1965) prominently argued, groups may fail to supply collective goods due to members’ rationally “freeriding”, unless there are selective incentives eliciting contributions (Olson 1965).

¹ An important exception is the so-called volunteer’s dilemma introduced by Diekmann (1985). In this game, one contributor is enough to generate the collective good.

Note: This work has benefitted from insightful comments by the editors and reviewers, as well as from stimulating discussions with the members of the Norms and Networks research group at the Department of Sociology/ICS of the University of Groningen. Any remaining deficiencies are, of course, sole responsibility of the authors.

Environmental pollution, lack of effort in joint team production, or the failure to maintain a valuable community resource are examples of Olson's famous "logic of collective action" (Bouma, Bulte, and van Soest 2008; Hardin 1968; Petersen 1992). But the empirical picture is not all that bleak. Collective goods are provided even without formal institutional solutions, as in mass protests to overthrow an oppressive regime (Opp, Voss, and Gern 1995), effective lobbying associations (Marwell and Oliver 1993), Wikipedia (Anthony, Smith, and Williamson 2009), or successful "self-managing teams" at the workplace (Barker 1993).

Our chapter focuses on peer sanctioning as an informal social institution supporting collective action. Sociologists (Coleman 1990; Homans 1951; Homans 1974), economists (Fehr and Gächter 2000; Fehr and Gächter 2002; Kandel and Lazear 1992), and political scientists (Ostrom, Walker, and Gardner 1992) have long recognized the importance of peer sanctioning for the enforcement of cooperation. Under a peer sanctioning institution, group members failing to pull their weight face expressions of disapproval, physical punishment, or ostracism by their peers (Homans 1951). Contributors are rewarded with peer approval, praise, or affirmation of their social standing (Willer 2009). Research in work groups has provided ample evidence of the power of peer sanctioning, starting with Roethlisberger and Dickson's classic Hawthorne studies (1939).

How robust, however, are peer sanctioning institutions as solutions to collective action problems? Our contribution focuses in particular on how the robustness of peer sanctioning depends on whether the sanction is reward (for cooperative behavior) or punishment (of uncooperative behavior), and we report results from a laboratory experiment (see Van Miltenburg et al. 2014 for a similar approach).

The robustness of peer sanctioning is debated in the literature because peer sanctioning institutions are threatened by the "second order free rider problem" (Oliver 1980). Peer sanctioning being both individually costly and in the collective interest, it is itself a collective good (Coleman 1990). Rational egoists should, therefore, refrain from sanctioning. Incentivized experiments with collective good games showed, however, that players are willing to invest in costly peer-punishment (Diekmann and Przepiorka 2015; Fehr and Gächter 2000; Fehr and Gächter 2002) and costly peer-reward (Flache and Bakker 2012; Flache 1996; Van Miltenburg et al. 2014). Fehr and Gächter (2000; 2002) and Fehr and Gintis (2007) explain "altruistic punishment" by positing a behavioral disposition or social preference in at least part of the human population "to punish violations of cooperative norms even at a net cost to the punisher" (Fehr and Gintis 2007:45).

Even when taking social preferences into account, theories of peer sanctioning are challenged to explain its robustness in the face of counter-sanctioning. In work teams, collaborative projects in science, and close-knit communities, for instance, punishers are not anonymous and face future interactions with those they punish. Consequently, fear of retaliation may deter peer-punishment. The experimental design employed by Fehr and Gächter, and virtually all follow-up studies in this paradigm, excluded

counter-punishment. To study the effects of counter-punishment, Nikiforakis (2008) added a stage to each period of the standard collective good experiment in which subjects first learned how strongly other group members had punished them and were given the opportunity to counter-punish. Participants were deterred from using punishment against freeriders in the first place, reducing contributions to the collective good (see also Denant-Boemont, Masclet, and Noussair 2007).

Research on counter-rewards similarly suggests that peer-reward may be vulnerable to the effects of counter-reward. Flache (1996; 2002; see also Flache and Macy 1996; Flache, Macy, and Raub 2000) analyzed a repeated game in which players could reward each other based on information about others' contribution to the collective good in previous periods. They compared two conditions, one in which enforcers remained anonymous and one in which their identity was revealed prior to the next round of the game. Both social learning theory (Flache and Macy 1996) and game-theory (Flache 2002) predicted that without anonymity players would refrain from using rewards as an instrument to enforce contributions under a large range of conditions. The reason is that in the non-anonymous condition players establish *mutually beneficial reward exchanges*, even with freeriders. Experimental tests confirmed the related prediction that subjects contribute less to the collective good in the non-anonymous condition (Flache and Bakker 2012; Flache 1996).

There is a crucial difference in the experimental designs used to study counter-punishment and counter-reward. On the one hand, counter-punishment was extremely salient and unambiguous in previous studies (Denant-Boemont, Masclet, and Noussair 2007; Nikiforakis 2008). After having been exposed to a sanction, subjects were immediately given the opportunity to strike back, and only players who had imposed a punishment in the preceding punishment stage could be targeted for counter-punishment. In the counter-reward experiments (Flache and Bakker 2012; Flache 1996), on the other hand, counter-sanctioning was more ambiguous. Participants were always fully informed about all group members' contributions and reward decisions, but the rules of the game did not favor any particular reaction to others' behaviors in the reward-stage of the game. Unlike in the counter-punishment experiments, there was no explicit stage of the game where counter-rewards were possible. Instead, participants could respond with changing contributions and reward decisions in subsequent periods.

The latter implementation of counter-sanctioning makes it more difficult for the researcher to disentangle rewards from counter-rewards, but comes with the important advantage of increased external validity. In real-life collective good situations, sanctions are inherently ambiguous signals, often embedded in long-term exchange processes in which players can respond with either contributions or sanctions to previous contributions or sanctions from others.

It is an open question whether punishment institutions are vulnerable to counter-sanctioning under these more realistic conditions of the counter-reward experiments. The ambiguity of sanctions in repeated non-anonymous interactions leaves room for

at least two possibilities. One is that potential enforcers are deterred by the prospect of counter-sanctions (Nikiforakis and Engelmann 2011). Actors contemplating punishment thus fear counter-punishment, and actors contemplating withholding a reward fear forfeiting future rewards themselves. The other possibility is that sanctioned freeriders refrain from retaliation because the enforcers against whom they retaliate can strike back in future encounters (Nikiforakis and Engelmann 2011). If the latter mechanism prevails, lack of anonymity combined with a “shadow of the future” should not undermine the effectiveness of a peer sanctioning institution as a solution to the problem of collective action. If the former mechanism predominates, however, repeated non-anonymous interactions will severely curtail the effectiveness of peer sanctioning.

In this study, we propose and experimentally test the theoretical prediction that the vulnerability of a peer sanctioning institution to counter-sanctions depends on whether the sanctioning institution is peer-reward or peer-punishment. In the subsequent section, we elaborate our theoretical predictions. We then describe the experiment and our results.

2 Theory and hypotheses

Under both peer-reward and peer-punishment institutions, group members provide selective incentives to fellow group members using social or material resources. Actors considering freeriding face the possibility of peers using the incentive conditionally on a target’s sufficient contribution to the common good.

Theoretical approaches in the literature differ when it comes to explaining why group members sanction in the first place. The social preference explanation (e.g., Bolton and Ockenfels 2000; Charness and Rabin 2002; Dijkstra 2012; Fehr and Schmidt 1999) posits that norm enforcers derive satisfaction from punishing a perpetrator, such that the material or social costs of enforcement are subjectively more than compensated. Peer-reward institutions, by contrast, are typically addressed from the perspective of social exchange theory (Coleman 1990; Dijkstra 2012; Dijkstra 2015; Holländer 1990; Homans 1974). In this view, group members reward contributors because the costs of rewarding are more than compensated by future contributions to the collective good that the reward recipient will make in response. This exchange theoretic model of peer sanctioning replaces the intrinsic benefits of sanctioning central to theories of peer-punishment, with the extrinsic motivation of obtaining future compensation for a present investment in an ongoing exchange relation.

In the present research, we follow previous work in assuming that peer-punishment is predominately motivated by its intrinsic benefits (e.g., vengeance, relief) for enforcers, whereas peer-reward is mainly driven by the enforcers’ desire to maintain a beneficial exchange with the target of the sanction. At the same time, we acknowl-

edge that future work should move beyond this distinction and explore the possibility that peer-rewards may be emotionally motivated, just as peer-punishment decisions may also be affected by exchange considerations. Evidence of “altruistic rewarding” has, for example, been provided by experiments with reward-based sanctioning institutions in which players invested in costly peer-reward, although reciprocation was precluded by a one-shot design (Van Miltenburg et al. 2014).

For an anonymous sanctioning institution, both emotionally driven peer-punishment and exchange driven peer-reward lead to the same qualitative prediction about effects of the institution. Both institutions will increase contribution rates in a repeated collective good game, in comparison with a game without a peer sanctioning institution (Hypothesis 1). More precisely, and in line with the common standard in the literature, we formulate this claim for a collective good game that imposes the incentive structure of an indefinitely repeated N -person Prisoner’s Dilemma game in which players are aware of the aggregate contribution of their peers in previous rounds but have – in the absence of a peer sanctioning institution – no means of responding other than by adapting their own contribution decision. For games like this, a common pattern frequently observed in experimental research is that participants initially make substantial contributions in the absence of a peer sanctioning institution, but cooperation rates gradually collapse as the game progresses (Andreoni 1988; Camerer 2003; Ledyard 1995).

An *anonymous* peer-punishment institution adds a second stage in every round, in which players learn individual peers’ contribution levels in the previous stage and can then respond with a negative sanction targeted at a particular peer. The institution is anonymous in the sense that targets do not know who imposed a sanction upon them. At the same time, in our design, the group remains stable throughout the game and all players keep the same labels, such that a sanction can always be given conditionally on their past *contribution* behavior. Imposing the sanction comes at a cost to both the recipient and the enforcer. An anonymous peer-reward institution differs from this only in that the sanction benefits the recipient.

The expected positive effects of anonymous peer sanctioning derive from the assumption that peer sanctions will be imposed, and that players expect this. This implies that the reward (punishment) a player receives from her peers in the sanctioning stage of a round of the game is higher if this player contributed (did not contribute) to the collective good in the contribution stage (Hypothesis 2 and Hypothesis 3).

A game-theoretic analysis can highlight more precisely the scope conditions under which effectiveness of a peer-reward institution is in line with players’ strategic rationality, even without intrinsic motivation to sanction. Drawing on the theory of repeated games (Friedman 1971; Friedman 1986; Taylor 1987), Flache (1996; see also Flache, Macy, and Raub 2000; Flache 2002) showed that the exchange-theoretic explanation of the effectiveness of peer-reward can be reconstructed in terms of individually rational conditional cooperation in an indefinitely repeated game (see also Spagnolo 1999 for a similar analysis). More precisely, given a sufficiently large contin-

uation probability, effective peer-reward under an anonymous peer-reward institution can be sustained by a subgame-perfect equilibrium in which all players condition their contributions to the collective good as well as the rewards they give for their peers' contribution behavior in previous rounds of the game.

This model of peer-reward imposes constraints on the incentive structure under which the institution can be effective. In a nutshell, the institution can be effective only if there is sufficient interest in future payoffs: the value of the reward is large enough to discourage players from risking losing future rewards despite the benefits of freeriding, and the costs of providing a reward are sufficiently low to guarantee a net benefit as long as the reward recipients respond with cooperation. But will the institution remain effective once enforcers are identifiable (thus making counter-reward possible), all other things being equal? To answer this, we will focus theoretically and experimentally on payoff parameters for the game that meet the constraints under which effective peer-reward is sustained by a subgame perfect equilibrium.

The conditions under which effective peer-reward is possible will also be taken as starting points for studying peer-punishment in our contribution. We assume that the provision of peer-punishment – unlike the provision of peer-reward – is more robust against higher costs or lower benefits of contribution and sanctions due to the additional intrinsic benefits of punishing. We therefore assume that if anonymous peer-reward is effective for given payoff parameters, so is anonymous peer-punishment. To render peer-reward and peer-punishment comparable, we equalize the material unit costs and benefits of the collective good and the unit costs of sanctioning under both institutions. Moreover, the material value of avoiding a punishment is equal to the material value of receiving a reward.² Alignment of the two institutions allows isolating the effects of adding the possibility of counter-sanctions to either institution.

² To be precise, the games are not exactly aligned in this way. A difference that remains is that perpetual conditional cooperation under the reward institution requires that players bear the costs of providing rewards throughout the game, whereas under the punishment institution perpetual conditional cooperation requires that no punishments need ever be carried out (see van Miltenburg et al. 2014 for a similar argument). This implies that when all costs and benefits are equal in the way described above, the conditions for conditional cooperation to constitute a Nash equilibrium in the repeated game are actually less restrictive for a punishment institution than they are for a reward institution. However, for the punishment game, a Nash equilibrium does not assure individual rationality in the stricter sense that the threat of actually carrying out the punishment is credible once it has failed to deter freeriding. Unless we assume intrinsic motivation, the fact that punishment is costly may withhold players from imposing the sanction in that situation. Technically, the strategy of conditional punishment may constitute a Nash equilibrium, but it is not necessarily a subgame perfect one. This is different for the reward game, where carrying out the threat of not rewarding is cost-free. Flache (1996; 2002) proved that in a reward game, the corresponding Nash equilibrium is also subgame perfect. For these reasons, the two games cannot readily be aligned in terms of the conditions under which conditional cooperation is individually rational. The best we can do, therefore, is to align all cost and benefit parameters such that they are equal in absolute value.

The non-anonymous sanctioning institutions we study differ from the anonymous ones only in that enforcers are identifiable for their peers in the interactions following enforcement. It has been shown that this small difference can profoundly affect a peer-reward institution (Flache and Bakker 2012; Flache and Macy 1996; Flache 1996). In a repeated game, the loss of anonymity provides players with additional information to condition their behavior. In particular, there is a new conditional strategy in which present rewards are conditioned on rewards received in the past. The problem for the effectiveness of the peer-reward institution is that in the corresponding equilibrium of the repeated game, the prospect of attaining future rewards from peers is no longer an incentive for a player to contribute to the collective good. Instead, it is an incentive to keep rewarding the peer even if she is a freerider.

To be sure, everyone would still be better off if the collective good were to be provided and players rewarded each other at the same time. Technically, the equilibrium corresponding to only mutual reward is payoff inferior to a competing one in which everyone both contributes and rewards all peers. However, the mutual-reward equilibrium also turns out to be more robust under random deviations (Flache 2002), easier to coordinate upon for boundedly rational backward-looking players (Flache and Macy 1996), and consistent with individual rationality under a less restrictive set of conditions (Flache, Macy, and Raub 2000; Flache 1996). Intuitively, the reason for this discrepancy is that a mutually beneficial exchange of reward for reward between two “friends” in a group is easier to establish and maintain than the more complex multilateral exchange between a sufficient number of contributors to the collective good and a sufficient number of peers needing to reward them for their efforts to motivate them into contributing (see Manhart and Diekmann 1989 for a similar argument about group size effects on the robustness of conditional cooperation).

Building on this theoretical work, we hypothesize that contribution rates will be lower under a non-anonymous peer-reward institution than under an otherwise equivalent anonymous one (Hypothesis 4). We expect that the rewards a player receives from her peers depend less on being a contributor when the peer-reward institution allows counter-reward than in an equivalent anonymous peer-reward institution (Hypothesis 5).

Effects of lifting the anonymity behind which enforcers can hide should be different when the motivation to sanction freeriders in the first place is intrinsic. To be sure, the experiments of Nikiforakis (2008) and others indicate that the costs of future retaliation are not entirely disregarded by enforcers in a peer-punishment institution. However, these experiments are crucially different from the peer-punishment institution we consider here. In our experiment, a player who counter-punishes an enforcer must face the possibility of retaliation by that same enforcer in the future. Other than for counter-reward, we cannot draw on elaborated formal modelling work to form theoretical expectations about effects of non-anonymity of punishers in a repeated game framework. The reason is that previous research on counter-punishment has focused both theoretically and empirically on one-shot games. Nevertheless, the underlying

logic of intrinsic motivation can also be applied to informally hypothesize effects of anonymous vs. non-anonymous punishment institutions in repeated games. Our reasoning starts from the notion that the possibility of counter-punishment evoking revenge is very real if we follow the arguments of Fehr and Gächter (2000; 2002; see also Bowles and Gintis 2011). Players can be expected to anticipate or learn from interactions with their peers that a considerable fraction of a group is prepared to punish norm violators at a cost to themselves. While non-anonymity of enforcement may thus reduce the extent to which freeriders are punished compared to an anonymous punishment institution, we expect that the prospect of future retaliation strongly deters punished freeriders from counter-punishing. In other words, under an anonymous punishment institution, the sanctions imposed on a player should be less contingent on the sanctions imposed by that player on her peers in the past than under a non-anonymous reward institution (Hypothesis 6). Accordingly, we also expect that the degree to which non-anonymity of the sanction will reduce rates of contribution to the collective good is higher if the sanction is reward than if the sanction is punishment (Hypothesis 7).

3 Method

3.1 Game and conditions

We compared five different institutions in our experiment, using five treatments in a between-subjects design. In all treatments, subjects played a repeated five-person Prisoner's Dilemma game. Following Flache (1996), we adopted a stage game with a dichotomous choice. In every round of the game, all players had to decide simultaneously whether or not to contribute to a collective good. After each round, they learned how many members contributed, and how many points they had earned from the collective good. The group playing the game consisted of the same five subjects with the same (anonymized) identities in all rounds. The first treatment of our experiment represented the baseline situation without a peer-sanctioning institution. The remaining four treatments added four different sanctioning institutions to the collective good game, implemented by a sanctioning stage following the contribution stage in every round of the game.

We manipulated the peer sanctioning institution in a full factorial design along two dimensions. The first dimension was whether the sanction was reward or punishment. In both cases, players learned in the second stage of every round who had contributed and who had failed to contribute to the collective good in the first stage of that round. They also learned the amount of contributions every group member had made in all previous rounds of the game. Next, subjects had to decide simultaneously whether or not to impose a sanction on each of the other group members, at a

cost to themselves. Under the reward (punishment) institution, the sanction increased (decreased) the wealth of the target. Unlike in the design of Fehr and Gächter (2000; 2002; see also Nikiforakis 2008; Denant-Boemont, Masclet, and Noussair 2007), players could not vary the magnitude of a sanction and there was no budget limitation constraining the number of sanctions they could impose. Instead, each sanction had a fixed cost for the enforcer and a fixed effect on the wealth of the recipient. These design choices were made to represent the situation of an informal peer sanctioning institution in which actors' instruments for sanctioning are typically social resources that do not depend on the material output of the collective action (e.g., the force of reprobation does not depend on material wealth).

The second dimension along which the peer sanctioning institutions varied was the anonymity of the enforcer. With anonymous peer sanctioning, players learned after the second stage of every period only the *number* of other players who had punished (or rewarded) them in the previous stage. They were not told by *whom*. When peer sanctioning was not anonymous, players also learned *who* had imposed sanctions upon them. Moreover, subject labels were never changed so that group members remained identifiable and could be linked to their previous actions by all participants of the game, in all rounds. In the non-anonymous treatments, subjects also received information of how often they had sanctioned each of their fellow group members on average, in all previous rounds of the game. Only in the non-anonymous sanctioning treatment did they additionally learn how often on average they had been sanctioned by each of their fellow group members in previous periods.

3.2 Payoff structure and wealth

Subjects received an initial endowment and were informed that their total wealth at the end of the game consisted of the initial endowment plus the total number of points earned in all periods of the game. To align wealth effects of punishment and reward we calculated endowments such that the maximal and minimal number of money units (MUs) subjects could end up with at the end of the game was equal across all peer sanctioning institutions.

We chose payoff parameters such that players faced a considerable incentive to freeride in the collective action game, such that sanctions could provide a strong incentive to either earn rewards from the other players or avoid being punished. The payoff a player received in one round of the game consisted of the sum of the payoffs obtained in the collective good stage and (except for the baseline) the sanctioning stage of the game. The payoff rule for the collective good part implied a conventional linear *N*-person Prisoner's Dilemma structure. A contribution added 30 MUs to the total value of the collective good in the current round at a cost of 20 MUs to the contributor. The total value of the collective good generated by the group in one round was divided by the number of players and the resulting amount added to every player's

Tab. 1: Payoff of a player in the collective good game.

Number of peers who contributed	0	1	2	3	4
Player did not contribute	0	6	12	18	24
Player contributed	-14	-8	-2	4	10

payoff in that round. This imposed a social dilemma in that the marginal costs of contribution (20) were more than three times as large as the marginal benefits ($30/5 = 6$), yielding a marginal net benefit of -14 MU. Table 1 shows the resulting payoff structure of the collective good stage game.

In the sanctioning stage of one round, a player could impose a sanction on each other group member at a cost of 3 MUs per target to the enforcer. The benefits of being rewarded or the costs of being punished increased linearly in the number of group members who rewarded (or punished) the target. Being rewarded (or punished) by one other player increased (or decreased) the recipient's wealth by 10 MUs. The relatively low cost of imposing a sanction compared to the value of being sanctioned implements the notion that social rewards or punishments are relatively easy to "produce" in an informal peer sanctioning institution, but can have a substantial impact on a recipient's subjective wellbeing (Coleman 1990). This choice made peer sanctions potentially an effective instrument to elicit contributions to the collective good. Tables 2 and 3 describe the payoff structure of the stage game in the sanctioning phase for reward and punishment respectively.

Being rewarded by two peers, or avoiding the punishment of two peers, could offset the costs of a contribution to the collective good. This also created a considerable incentive to use sanctions as instruments to elicit reward or avoid punishment in the institutions with counter-sanctions. At 3/10, the cost-to-benefit ratio of mutual reward

Tab. 2: Payoff of player at the sanction stage under a peer-reward institution.

	Number of peers rewarding the player	0	1	2	3	4
Number of peers rewarded by the player						
0		0	10	20	30	40
1		-3	7	17	27	37
2		-6	4	14	24	34
3		-9	1	11	21	31
4		-12	-2	8	18	28

Tab. 3: Payoff of player at the sanction stage under a peer-punishment institution.

	Number of peers punishing the player	0	1	2	3	4
Number of peers punished by the player						
0		0	-10	-20	-30	-40
1		-3	-13	-23	-33	-43
2		-6	-16	-26	-36	-46
3		-9	-19	-29	-39	-49
4		-12	-22	-32	-42	-52

was considerably better than the cost-to-benefit ratio of contributing to the collective good in a situation of universal cooperation (20/30).

3.3 Duration of the game and payment

Subjects received no information about the duration of the game other than that the maximum duration of an experimental session was one hour. Thus, since subjects did not know the number of rounds to be played, we regard this as an indefinitely repeated game. In fact, the game ended after 20 rounds and endowments were calculated such that, across all sanctioning treatments, subjects could in the worst case end up with zero MU, whereas in the best case they would end up with 1,800 MU. In the punishment treatments, subjects started with a wealth of 1,320 MU in their accounts, whereas they started with 520 MU in the reward treatments. In the baseline treatment, subjects could never end up with fewer than zero MUs and they could earn maximally 760 MUs after 20 rounds. All subjects were paid a show-up fee of 5 € and also received a payment proportional to the final wealth in their account at a conversion rate of 1 € per 180 MU. Unlike many experiments with designs that prevent reciprocation (e.g., Fehr and Gächter 2002), the subjects of our experiments received a sizeable endowment at the outset rather than a series of small endowments per round. This approach allows comparison with previous similar experiments using peer-reward in a repeated setting (Flache 1996). It also allows subjects better to realize the long-term consequences of their actions by showing how their wealth develops over time, in accordance with the theoretical notion of a longer-term exchange perspective in the reward treatment.

3.4 Participants

The experiments were conducted at the “Sociological Laboratory” of the University of Groningen, The Netherlands (<http://www.gmw.rug.nl/~orsee/public/>). The Sociological Laboratory’s subject pool mostly consists of undergraduate students from a variety of disciplines including sociology, economics, law, biology, physics, etc. Students of psychology and sociology are overrepresented in the subject pool. The rules of the Sociological Laboratory guarantee that subjects will not be deceived, and that they will be paid for their efforts.

We conducted nine sessions with 120 subjects in total, where 20 subjects (4 groups) participated in the baseline treatment. Of the subjects, 25 were assigned to the reward treatment and 35 to the reward and counter-reward treatment. The treatments with only punishment and with punishment and counter-punishment each comprised of 20 participants.

Experiments took place in computer rooms, with subjects in separate cubicles. The experiments were programmed as a web application by the second author. After an oral introduction by the experimenter, each participant randomly picked an envelope containing detailed instructions and a login code to start the experiment. This ensured random assignment of participants to groups.

4 Results

Hypothesis 1 states that contribution rates are higher in the anonymous sanctioning institutions than in the baseline treatment. Figure 1 pictures the development of contribution rates in the five treatments of the experiment. Dots show observed average rates in sets of five subsequent periods. The error bars show 95 % confidence intervals with robust standard errors clustered on subjects.³ In the baseline condition, the contribution rate was 0.37 averaged over all periods. Supporting Hypothesis 1, Figure 1 shows fewer contributions in the baseline condition than in all other treatments. Compared to the baseline treatment, random-effects logistic regressions with random intercepts at the level of subjects and groups showed significantly more contributions in the anonymous reward treatment ($z = 6.45$ [$p = 0.000$], log odds model), the non-anonymous reward treatment ($z = 4.82$ [$p = 0.000$], log odds model), the punishment treatment ($z = 2.08$ [$p = 0.037$], log odds model) and the non-anonymous punishment treatment ($z = 2.61$ [$p = 0.009$], log odds model). The same regression models with round as an additional independent variable showed significantly decreasing contribution rates in the baseline treatment ($z = -3.98$ [$p = 0.000$]) and the treat-

³ Due to the small number of observations on the levels of subjects and groups, it was not possible to estimate multi-level logistic regressions for this figure.

ments with anonymous punishment and non-anonymous punishment ($z = 2.46$ [$p = 0.014$]). There was no significant trend in the remaining treatments (all $\text{abs}(z) < 0.87$, [$p > 0.38$]). These results not only support Hypothesis 1 in showing that contribution rates in the conditions with anonymous peer sanctioning exceed those in the baseline, but also demonstrate that, under the institutions allowing counter-sanctioning, contribution rates are higher than in the baseline.

Fig. 1: Treatment effects on contribution rates over time.

Supporting Hypotheses 2 and 3, Figure 2 shows that in peer sanctioning institutions, subjects who contributed received more rewards (fewer punishments) in the subsequent sanctioning stage. Multi-level Poisson regressions with random intercepts at the level of groups and subjects showed that the effect of the subject's own contribution behavior on group members' sanctioning decisions directed at the subject was significant in all peer sanctioning treatments. After a subject contributed, the expected log count of the number of received rewards increased by 2.17 ($z = 8.33$ [$p = 0.000$]) in the anonymous reward treatment and by 1.33 ($z = 10.96$ [$p = 0.000$]) in the non-anonymous reward treatment. The effect of a subject's contribution on the amount of sanctions received was significantly stronger in the anonymous reward treatment than in the non-anonymous reward treatment ($z = -2.89$ [$p = 0.004$]), supporting Hypothesis 5.

After a subject contributed, the expected log count of the number of received punishments decreased by 1.56 ($z = -9.97$ [$p = 0.000$]) in the anonymous punishment treatment and by 3.32 ($z = -11.05$ [$p = 0.000$]) in the non-anonymous punishment treatment. The difference in the two effects is significant ($z = -4.53$ [$p = 0.000$]).

Figure 1 shows that contribution rates hardly differed between the anonymous and non-anonymous reward treatments, contradicting Hypothesis 4 that contribution rates should be lower in the latter than in the former. Average contribution rates in the non-anonymous reward treatment were about 5% lower than in the anonymous reward treatment, but random-effects logistic regressions with random intercepts at the level of subjects and groups revealed that this difference is not significant ($z = -0.96$ [$p = 0.337$]). A trend analysis showed that neither the slopes ($z = -0.85$ [$p =$

Fig. 2: Treatment effects on average number of received sanctions in Period t , over subject's contribution choice in the same period.

0.400)] nor the intercepts ($z = -0.31$ [$p = 0.757$]) of the contribution rates differed significantly between the two reward treatments.

In the non-anonymous punishment treatment contribution rates were 5 % higher than in the anonymous punishment treatment. This effect too, was insignificant ($z = 0.03$ [$p = 0.976$]). The trend analysis also showed no treatment effect (intercept: $z = 0.33$ [$p = 0.741$]; slope: $z = -1.20$ [$p = 0.230$]).

Figure 3 depicts the development of sanctioning rates over time. Each dot shows the rate of sanctioning of all subjects of the respective treatment at the respective period of the experiment. The solid lines show linear trends. We estimated these trends with multi-level logistic regressions with random intercepts at the level of groups and subjects, using a dataset with 4 observations per subjects and period (one sanctioning decision for each of the 4 fellow group members). We found no significant difference in initial sanctioning rates between the anonymous reward treatment and the non-anonymous reward treatment ($z = -0.05$ [$p = 0.960$], log odds model). However, in the anonymous reward treatment, sanctioning rates significantly decreased ($z = -6.28$ [$p = 0.000$]), whereas there was a significant linear increase in sanctioning in the non-anonymous-reward treatment ($z = 2.69$ [$p = 0.007$]). Based on the linear trend, the estimated reward rate dropped from 0.682 initially to 0.480 in

Fig. 3: Development of sanctioning rates.

round 20 in the anonymous reward treatment, and rose from 0.650 to 0.716 in the non-anonymous reward treatment. Figure 3 shows that subjects sanctioned much less in the two punishment treatments. In the anonymous punishment treatment, punishment rates fell significantly over time ($z = -2.25$ [$p = 0.024$]). The same trend was found in the non-anonymous punishment treatment ($z = -2.66$ [$p = 0.008$]). Neither the intercept ($z = -1.74$ [$p = 0.082$]) nor the slope (-0.95) of the trend line differed between the two treatments (in the log odds model). Based on the linear trend, the estimated punishment rate dropped from 0.167 initially to 0.104 in round 20 in the anonymous punishment treatment and from 0.082 to 0.030 in the non-anonymous punishment treatment.

Fig. 4: Effects of other players' contribution and sanction of focal player on sanctioning behavior focal player.

Figure 4 shows sanctioning rates in the four treatments with sanctioning institutions. To generate Figure 4, we analyzed dyadic data, where each case represents the decision of subject i to sanction or not sanction one of her four team members j in a given period t . Figure 4 shows the sanctioning rates of i in Period t depending on j 's contribution decision in the same period and whether or not j had sanctioned i in the previous period. The numbers above the bars of Figure 4 indicate how often one of our subjects faced the respective decision. The error bars shown in Figure 4 visualize the 95 % confidence intervals estimated with robust standard errors (clustered on subjects). In the following, however, we report only results from multi-level logistic regressions with random intercepts at the levels of groups and subjects.

The two left-hand panels of Figure 4 show that contributors (grey bars) were rewarded more often than defectors (white bars) in both treatments with a peer-reward institution. This difference is significant ($z = 23.16$ [$p = 0.000$]). The two right-hand panels show that contributors were punished less often than defectors ($z = -13.97$ [$p = 0.000$]) in treatments with a peer-punishment institution.

Our results are compatible with counter-rewarding having taken place. Even for the anonymous reward treatment, Figure 4 shows that subjects rewarded by a team mate j in the previous period rewarded that team mate at a higher rate in the current period ($z = 14.62$ [$p = 0.000$]). Subjects in this treatment were not informed about which group member rewarded them. However, this effect suggests that subjects nevertheless inferred from past periods who had rewarded them and tried to reciprocate, despite the risk of rewarding the 'wrong' group member. In line with the notion that subjects use reward as an instrument of exchange when this is possible, Figure 4 also shows that there was clearly more counter-rewarding in the non-anonymous reward treatment. When also controlling for the group member's contribution, we found a significant increase in reward rates when the group member receiving the reward had previously rewarded the subject ($z = 13.89$ [$p = 0.000$]). This was observed both when the other group member contributed and when she defected.

There is weaker evidence for counter-punishment. In the treatment with anonymous punishment, there is no significant effect of j 's previous sanctioning behavior on i 's punishment decision ($z = -0.27$ [$p = 0.787$]). This is different from what we observed in the anonymous reward treatment. In the non-anonymous punishment treatment, subjects punished a group member more often if they had been punished by the same subject in the previous period ($z = 2.91$ [$p = 0.004$]). Subjects even counter-punished contributors (with a rate of 9.4 %) if the contributor had punished them in the previous period. Note, however, that our subjects encountered this situation only 64 times. Thus, this very strong form of "antisocial counter-punishment" was observed only six times.

Hypothesis 6 predicts that, in the non-anonymous punishment institution, the sanctions imposed on a player by her peers should be less contingent on the sanctions imposed by that player on her peers in the past than in a non-anonymous reward institution. Figure 5 provides no support for this hypothesis, showing sanctioning rates

Fig. 5: Test of interaction treatment with effects of sanctioning behavior other player on focal player's sanction (hypothesis 6).

estimated with multi-level logistic regressions with random intercepts at the level of groups and subjects. Figure 5 shows that we did not find for the comparison of the two reward treatments a stronger effect for having been sanctioned by a team member in the past than we find for the comparison of the two punishment treatments. To be precise, we did find a significant difference between these two treatments in the log odds model ($z = 2.26$ [$p = 0.024$]), but as Figure 5 shows, this effect is not meaningful in terms of probability differences.

Hypothesis 7 predicts that the possibility of counter-sanctioning will reduce contribution rates more in the reward treatments than in the punishment treatments. Figure 6 compares contribution rates in the four treatments estimated with multi-level logistic regressions with random intercepts at the level of groups and subjects. It turns out that there is no significant difference in the contribution rates between the two reward treatments ($z = -0.18$ [$p = 0.857$]), nor is there a difference between the two punishment treatments ($z = -0.03$ [$p = 0.976$]). Thus, our data lend no support to Hypothesis 7.

The different contribution and sanctioning patterns in the five treatments also translated into payoff differences. In the baseline treatment, participants earned on average 3.7 MU per round ($t = 3.74$ [$p = 0.001$]). In stark contrast, participants earned on average 25.1 MU ($t = 30.00$ [$p = 0.000$]) in the anonymous reward treatment and 26.6 MU ($t = 21.59$ [$p = 0.000$]) in the non-anonymous reward treatment. The difference between the two reward treatments is not significant ($t = -1.68$ [$p = 0.098$]).

Fig. 6: Test of interaction effect of reward vs. punishment by anonymity of sanction on contribution (hypothesis 7).

Obviously, these high payoffs per round resulted from the very high contributions and the fact that receiving a reward increases one's payoff. Payoffs were much lower in the punishment treatments. In the anonymous punishment treatment, participants lost 0.1 MU on average ($t = -0.07$ [$p = 0.945$]). In 40 % of all played rounds in this treatment participants had a negative payoff. In the non-anonymous punishment treatment, participants earned 4.50 MU on average ($t = 7.09$ [$p = 0.000$]). Subjects could earn more MUs in the reward conditions due to the rules of the game. We corrected for this with lower endowments in the reward conditions as compared to the punishment conditions, aligning the maximum possible wealth in all sanctioning conditions to 1,800 MU (see above). This resulted in higher average net wealth at the end of the game for subjects in punishment conditions (anonymous: 1318, non-anonymous: 1,410) compared with subjects in reward conditions (anonymous: 1,022, non-anonymous: 1,052).

5 Discussion and conclusion

Previous research has found that peer sanctioning institutions as a solution to the freerider problem can be vulnerable to counter-sanctioning. However, these studies did not allow for an assessment of whether peer-punishment is vulnerable under the same conditions as peer-reward. In this chapter we focused on conditions reflecting peer sanctioning in many empirical settings, that is, small group situations in which enforcers cannot remain anonymous and retaliation against sanctions and counter-sanctions is possible in future encounters. We expected that the non-anonymity of enforcers in repeated interaction may eliminate the vulnerability of peer-punishment to counter-punishment postulated by previous research, while we expected non-anonymity to reduce the effectiveness of peer-reward. Comparing the effects of counter-punishment and counter-reward for this setting in a collective good experiment, we found no evidence that the possibility for counter-sanctioning undermined

peer sanctions as a solution to the freerider problem. These results put into perspective the results of previous studies of counter-punishment (Nikiforakis 2008) and support our expectation that counter-punishment may not be a problem in repeated non-anonymous interactions, as they often occur in real life collective action settings. What we did not expect was that a similar robustness to non-anonymity would be found for a reward institution, in contrast with previous research (Flache 1996). Our results thus raise probing questions.

Strikingly, our experiments did support the theoretical mechanisms on the basis of which we expected that peer sanctioning would be vulnerable to counter-reward and robust to counter-punishment. We found that subjects traded rewards for rewards rather than using them to enforce contributions to the collective effort in the non-anonymous reward treatment. This was in line with the notion that reward is driven by a logic of exchange and conditional cooperation that can undermine enforcement. We also found, as hypothesized, that the possibility of counter-punishment only slightly reduced enforcement rates in the non-anonymous peer-punishment institutions, consistent with the explanation that peer-punishment is intrinsically motivated.

A reason why the theoretical mechanisms we found did not induce a corresponding difference in terms of the vulnerability of the institutions could be that the vulnerability of peer-reward only shows up in the longer term. Our data (see Figure 3) suggest that the differences between sanctioning behavior with and without the possibility of counter-reward gradually increase over time, which is in line with theoretical predictions from learning theory and results of earlier studies using longer games than ours (Flache and Bakker 2012; Flache 1996). These differences did not yet apparently translate into differences in contribution decisions, but it is possible that over more than the 20 rounds that we employed contributions will be affected. An extrapolation of the increasing trend in reward rates that we found in the non-anonymous reward treatment allows for speculation that, after sufficient time, participants in this treatment may discover that they can obtain high earnings even without having to make costly contributions, eventually yielding a decline in contribution rates similar to the one found by Flache (1996). Likewise, an extrapolation of the simultaneous increase in contribution rates and decline in punishment rates, which we observed in the non-anonymous punishment treatment after the initial phase of the game, could indicate that eventually both freeriding and counter-punishment may be sufficiently deterred by experiences of retaliation to guarantee sustained contribution without the need for actual punishment. Future research should, therefore, study longer games and manipulate the length of the game to assess how this affects the link between contribution behavior, sanctioning and the possibility of counter-sanctioning.

Our results also differ in further interesting ways from previous experiments comparing reward and punishment. While we observe higher contribution rates in the anonymous reward condition than in the anonymous punishment condition, Van Miltenburg and coauthors (2014) find the opposite relation. More precisely, contribution rates in their punishment conditions are approximately comparable to our anonymous

punishment condition, but contribution rates in our anonymous reward conditions are clearly higher than in Miltenburg et al.'s reward condition. However, in contrast with our anonymous sanctioning conditions, their treatment uses random matching after every round. We believe that this can explain why we observe higher contributions under anonymous reward. Unlike theirs, our reward treatment allows for the emergence of a repeated exchange of reward for contribution, even when enforcers are unknown. Contributors know even in the anonymous reward condition that they can expect future reward for present contributions. This provides a stronger incentive to contribute than in a random matching design. If punishment is primarily driven by emotional gratification, this difference in design should be less relevant for the punishment treatment than for the reward treatment, because under punishment the prospect of building up long-term exchanges is not the main motivation to contribute. The fact that we do not use random matching may, however, explain why we find relatively low punishment rates compared to previous experiments with peer-punishment. Without random matching, persistent freeriders risk repeated and escalating punishment from the same emotionally-driven enforcers. Accordingly, in an anonymous punishment design without random matching, less punishment may be needed to credibly signal that freeriding will elicit future sanctions.

We believe that a promising direction for future research is the integration of theoretical models of peer-punishment based on social preferences, with models of peer-reward based on conditional cooperation and exchange. While these mechanisms have hitherto typically been separated in the literature, there is no compelling reason why this should be so. Even when “altruistic punishment” is intrinsically motivated, players can still consider the – partially subjective – costs and benefits of it in terms of long-term exchange outcomes. Conversely, even when conditional cooperation in the exchange of peer-reward against contribution is primarily driven by associated long-term costs and benefits, actors may derive some intrinsic benefit from not rewarding a freerider, or from rewarding a contributor. Future research can assess whether such combined models can better account for differences between peer-punishment and peer-reward.

Peer sanctioning is by far not the only solution to freerider problems in collective action, and under many conditions it may not be the most powerful or most efficient solution. For example, signaling and reputation systems can be highly effective in large-scale anonymous online markets (Diekmann et al. 2014) in which there is no possibility of long-term repeated interaction with the same partners. However, the conditions of relatively small scale, non-anonymous and repeated interaction that we focused upon here reflect those of many freerider problems people face in their daily lives: at the workplace, in their neighborhoods or in informal social groups. Our work suggests that under those conditions, the informal “bottom up” institution of peer sanctioning may be less vulnerable and more effective than previous research on counter-sanctioning has assumed. This is a potentially hopeful insight that warrants further scrutiny in future work.

Appendix: instructions

Below we present sample instructions for the Punishment and Counter-punishment condition. Presented first are the written instructions given to participants at the start of the experiment. The written instructions were very similar across all conditions. For the Baseline condition, sentences referring to a change in others' payoffs were removed. In conditions which employed Reward sanctioning rather than Punishment sanctioning, the word *decrease* was replaced with the word *increase*. For each condition the endowment was changed to the appropriate amount. No other changes were required. The written instructions for the Punishment and Counter-punishment condition were as follows:

Rules of the group project

The computer assigned you to a group of 5 participants. In every round of the study, all 5 group-members ("buddies") face the same decision.

At the beginning of every round, each buddy decides whether to invest 20 points into a group project or to keep the 20 points for him or herself. The group project is simple. We will add up all investments of the 5 buddies, multiply this sum by 1.5 and divide this sum amongst all 5 buddies.

Example 1: If all 5 buddies decide to invest their 20 points, the group project has a value of 5 times 20, which is 100. We multiply this by 1.5, which is 150, and divide it amongst the five of you. Thus, each buddy receives 30 points ($= 150/5$). Investing cost you 20 points, so for you the total number of points in this round is 30 minus 20, which is 10.

Example 2: Assume that you decide to keep your 20 points but your 4 buddies contribute their 20 points. Because you don't invest, you do not lose any points. The sum of the others' contributions is 4 times 20, which is 80. This is multiplied by 1.5, which is 120. This is divided amongst all 5 group members. Thus, each of you receives 24 points. For you the total number of points in this round is 24.

After every buddy has made a decision, we will inform you about the decisions of the other 4 buddies and how many points you earned. Every buddy is then given the opportunity to react to the decisions of the others. You can decrease the payoff of each of your buddies by 10 points. Thus, if you decide to decrease a buddy's payoff, then this buddy will lose 10 points. If you decide to decrease a buddy's payoff, you have to pay a fee of 3 points.

At the same time, each of your buddies can decrease your payoff in the same way. We will always inform you about whether or not your buddies decided to decrease your payoff.

180 points are worth one Euro. You start the experiment with 1320 points. You can gain and lose points during the experiment.

Fig. 7: Written instructions for the Punishment and Counter-punishment condition.

We also present a sample screen from the Punishment and Counter-punishment condition, showing the sanctioning stage. In this stage, participants are presented with their buddies' contribution decisions and contribution history. They respond by selecting *Yes* or *No* for each buddy, indicating whether they want to reduce this buddy's payoff or not. Participants are immediately shown how much the selected decisions will cost them. This screen appeared in all conditions in which sanctioning was possible.

How do you want to react to the choices of your buddies?

In the first row, you see the contributions of your buddies. In the second row, you make your decisions. Then click the continue button to move on.

	YOU	BUDDY 1	BUDDY 2	BUDDY 3	BUDDY 4
☒ Contributed	NO	YES [100%]	YES [100%]	NO [0%]	YES [100%]
☒ You decrease buddy's payoff by 10		<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
☒ Buddy decreases your payoff by 10					

Summary
3 buddies contributed.

For the selected decisions you will pay a reaction fee of: 3

Legend

For every decision made by your buddies, the percentage [in brackets] shows you how often your buddy chose 'Yes' on this decision in all rounds so far.

☒ Stage complete
☐ Stage in progress

Fig. 8: Sanctioning stage for the anonymous punishment condition and the non-anonymous punishment condition.

The decision screens throughout the experiment were designed to make sure participants had all available information from the current period on their screen at all times, while not overwhelming them. As the experiment progressed through the stages of each period, the lines of the decision table were filled from top to bottom. The condition a participant played determined which lines of the decision table were presented. Throughout the experiment, we color-coded *decisions* (blue) and *consequences* (orange) to help participants keep track of what happened during the current period.

Bibliography

- [1] Andreoni, James. 1988. "Why Free Ride?: Strategies and Learning in Public Goods Experiments." *Journal of Public Economics* 37(3):291–304.
- [2] Anthony, Denise, Sean W. Smith, and Timothy Williamson. 2009. "Reputation and Reliability in Collective Goods: The Case of the Online Encyclopedia Wikipedia." *Rationality and Society* 21(3):283–306.
- [3] Barker, James R. 1993. "Tightening the Iron Cage: Concertive Control in Self-Managing Teams." *Administrative Science Quarterly* 38(3):408–437.
- [4] Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90(1):166–193.
- [5] Bouma, Jetske, Erwin Bulte, and Daan van Soest. 2008. "Trust and Cooperation: Social Capital and Community Resource Management." *Journal of Environmental Economics and Management* 56(2):155–166.
- [6] Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

- [7] Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. New York: Russel Sage Foundation.
- [8] Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117(3):817–869.
- [9] Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.
- [10] Denant-Boemont, Laurent, David Masclet, and Charles N. Noussair. 2007. "Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment." *Economic Theory* 33(1):145–167.
- [11] Diekmann, Andreas. 1985. "Volunteer's Dilemma." *Journal of Conflict Resolution* 29(4):605–610.
- [12] Diekmann, Andreas, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. 2014. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79(1):65–85.
- [13] Diekmann, Andreas, and Wojtek Przepiorka. 2015. "Punitive Preferences, Monetary Incentives and Tacit Coordination in the Punishment of Defectors Promote Cooperation in Humans." *Scientific Reports* 5:10321.
- [14] Dijkstra, Jacob. 2012. "Explaining Contributions to Public Goods: Formalizing the Social Exchange Heuristic." *Rationality and Society* 24(3):324–342.
- [15] Dijkstra, Jacob. 2015. "Social Exchange: Relations and Networks." *Social Network Analysis and Mining* 5(1):60. doi:10.1007/s13278-015-0301-1.
- [16] Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90(4):980–994.
- [17] Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415(6868):137–140.
- [18] Fehr, Ernst, and Herbert Gintis. 2007. "Human Motivation and Social Cooperation: Experimental and Analytical Foundations." *Annual Review of Sociology* 33:43–64.
- [19] Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114(3):817–868.
- [20] Flache, Andreas. 1996. *The Double Edge of Networks: An Analysis of the Effect of Informal Networks on Cooperation in Social Dilemmas*. Amsterdam: Thesis Publishers.
- [21] Flache, Andreas. 2002. "The Rational Weakness of Strong Ties: Failure of Group Solidarity in a Highly Cohesive Group of Rational Agents." *Journal of Mathematical Sociology* 26(3):189–216.
- [22] Flache, Andreas, and Dieko M. Bakker. 2012. "De Zwakke Kant van Sociale Controle. Theoretisch en Experimenteel Onderzoek naar de Spanning Tussen Groepsreciprociteit en Relationale Reciprociteit in een Sociaal Dilemma." Pp. 103–130 in *Samenwerking in sociale dilemma's. Voorbeelden van Nederlands onderzoek*, edited by V. Buskens and I. Maas. Amsterdam: Amsterdam University Press.
- [23] Flache, Andreas, and Michael W. Macy. 1996. "The Weakness of Strong Ties: Collective Action Failure in a Highly Cohesive Group." *Journal of Mathematical Sociology* 21(1–2):3–28.
- [24] Flache, Andreas, Michael W. Macy, and Werner Raub. 2000. "Do Company Towns Solve Free Rider Problems? A Sensitivity Analysis of a Rational-Choice Explanation." *The Management of Durable Relations: Theoretical and Empirical Models for Households and Organizations*, edited by W. Raub and J. Weesie. Amsterdam: Thela Thesis.
- [25] Friedman, James W. 1971. "A Non-Cooperative Equilibrium for Supergames." *Review of Economic Studies* 38(1):1–12.
- [26] Friedman, James W. 1986. *Game Theory With Applications to Economics*. New York: Oxford University Press.
- [27] Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162(3859):1243–1248.

- [28] Holländer, Heinz. 1990. "A Social Exchange Approach to Voluntary Cooperation." *The American Economic Review* 80(5):1157–1167.
- [29] Homans, George C. 1951. *The Human Group*. New York: Harcourt Brace Jovanovich.
- [30] Homans, George C. 1974. *Social Behavior: Its Elementary Forms*. New York: Harcourt Brace Jovanovich.
- [31] Kandel, Eugene, and Edward P. Lazear. 1992. "Peer Pressure in Partnerships." *Journal of Political Economy* 100(4):801–817.
- [32] Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." Pp. 111–194 in *Handbook of Experimental Economics*, edited by J. H. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press.
- [33] Manhart, Klaus, and Andreas Diekmann. 1989. "Cooperation in 2- and N-Person Prisoner's Dilemma Games: A Simulation Study." *Analyse und Kritik* 11(2):134–153.
- [34] Marwell, Gerald, and Pamela Oliver. 1993. *The Critical Mass in Collective Action*. Cambridge: Cambridge University Press.
- [35] Van Miltenburg, Nynke, Vincent Buskens, Davide Barrera, and Werner Raub. 2014. "Implementing Punishment and Reward in the Public Goods Game: The Effect of Individual and Collective Decision Rules." *International Journal of the Commons* 8(1):47–78.
- [36] Nikiforakis, Nikos. 2008. "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics* 92(1):91–112.
- [37] Nikiforakis, Nikos, and Dirk Engelmann. 2011. "Altruistic Punishment and the Threat of Feuds." *Journal of Economic Behavior and Organization* 78(3):319–332.
- [38] Oliver, Pamela. 1980. "Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations." *American Journal of Sociology* 85(6):1356–1375.
- [39] Olson, Mancur. 1965. *The Logic of Collective Action*. Cambridge: Harvard University Press.
- [40] Opp, Karl-Dieter, Peter Voss, and Christiane Gern. 1995. *Origins of a Spontaneous Revolution: East Germany, 1989*. Ann Arbor, MI: University of Michigan Press.
- [41] Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants With and Without a Sword: Self-Governance Is Possible." *The American Political Science Review* 86(2):404–417.
- [42] Petersen, Trond. 1992. "Individual, Collective and Systems Rationality in Work Groups: Dilemmas and Market-Type Solution." *American Journal of Sociology* 98(3):469–510.
- [43] Roethlisberger, Fritz J., and William J. Dickson. 1939. *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works*. Cambridge, MA: Harvard University Press.
- [44] Spagnolo, Giancarlo. 1999. "Social Relations and Cooperation in Organizations." *Journal of Economic Behavior and Organization* 38(1):1–25.
- [45] Taylor, Michael. 1987. *The Possibility of Cooperation*. Cambridge: Cambridge University Press.
- [46] Willer, Robb. 2009. "Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem." *American Sociological Review* 74(1):23–43.