university of groningen

University Medical Center Groningen

University of Groningen

Cooperation and social control

Bakker, Dieko Marnix

DOI:
10.33612/diss.98552819

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:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Bakker, D. M. (2019). *Cooperation and social control: effects of preferences, institutions, and social structure*. Rijksuniversiteit Groningen. https://doi.org/10.33612/diss.98552819

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.

Download date: 24-10-2021

3

Peer punishment and retaliation in ongoing interactions

Dieko M. Bakker, Jacob Dijkstra, Andreas Flache

This chapter is currently under review at an international peer-reviewed journal

Social life is full of cooperation problems, and lack of cooperation can result in outcomes that are neither collectively nor individually desirable (Kollock, 1998; Olson, 1965). Examples abound, ranging from pollution, across conflicts at work, to lack of safety in public places or failure of interest groups to mobilize collective action. Numerous solutions to cooperation problems have been proposed. In this study, we focus on peer punishment, an informal "bottom-up" solution that has received great attention from biologists, psychologists and behavioral economists. Peer punishment rests on human predispositions towards cooperative behavior and enforcement of others' cooperation, which are argued to be stable across different contexts (Au & Kwong, 2004; Peysakhovich et al., 2014; Van Lange, Balliet, et al., 2014; Van Lange et al., 2007) and developed in the evolutionary past (Bowles & Gintis, 2004, 2011; Kurzban et al., 2015). Behavioral economists specifically provided evidence for the presence of a considerable portion of strong reciprocators in the human population: individuals inclined to cooperate with others and punish those who do not cooperate (Fehr & Gächter, 2002; Fehr & Gintis, 2007; Simpson & Willer, 2015).

Much evidence for the effectiveness of peer punishment in promoting cooperation has been collected in behavioral experiments (Chaudhuri, 2011) implementing peer punishment institutions in collective good games These institutions allow members of a group to punish other group members, at a cost to themselves (Guala, 2012; Elinor Ostrom, 2000). In behavioral experiments, costs and effects of punishments are typically expressed in material terms; both the punisher and the punished are materially less well off after punishment has been imposed. In peer punishment institutions, group members can neither coordinate their punishments nor share the costs associated with punishing. Despite having to bear significant individual costs, many people do punish (Falk, Fehr, & Fischbacher, 2005; Fehr & Gächter, 2000, 2002). Most punishments are given to non-cooperators, resulting in increased cooperation and thus more contributions to collective goods. In groups with peer punishment, contribution rates were found to increase over the course of repeated interaction, and the threat of punishment often increases cooperation even before any punishments have been inflicted (Fehr & Gächter, 2002; Fehr & Gintis, 2007; Elinor Ostrom et al., 1992).

However, peer punishment is also risky, due to potential *antisocial* punishment and the threat of *retaliation*. The punishment of cooperative behavior, called antisocial punishment, is in fact fairly common. A cross-national comparison of antisocial punishment behavior showed that punishment aimed at cooperators occurs throughout the world and can be as frequent as punishment aimed at non-

cooperators (called *prosocial* punishment (Herrmann et al., 2008)). Additionally, peer punishment outside of laboratory environments tends to come with a risk of retaliation (Balafoutas & Nikiforakis, 2012; Guala, 2012). Yet, peer punishment institutions implemented in laboratory experiments are traditionally designed such that they exclude the possibility of retaliation (Guala, 2012; Nikiforakis, 2008). Punishments can only be handed out once per period of the experiment. At the end of each period, groups may be disbanded and recomposed or the groups may remain constant while experimenters ensure that group members are unaware of who punished them (Fehr & Gächter, 2000, 2002; Nikiforakis, 2008). Under these circumstances, directed retaliation is practically impossible even when group members interact repeatedly.

Outside of laboratory experiments, it is difficult to punish anonymously. When peer punishment is not anonymous, recipients of punishment can retaliate. In fact, the existence of antisocial punishment and retaliation is one of the main reasons why it has been proposed that alternative institutions such as coordinated sanctioning (e.g. Van Miltenburg, Buskens, Barrera, & Raub, 2014) or the selection of designated punishers (e.g. Baldassarri & Grossman, 2011; Kuwabara & Yu, 2017) may be more effective (Simpson & Willer, 2015).

Evidence that retaliation is a serious problem impairing the effectiveness of peer punishment is found in experiments that lift the veil of anonymity behind which enforcers could hide in earlier studies. A series of experiments has shown that, by excluding the possibility of retaliation, the effectiveness of peer punishment may have been significantly overstated (Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008; Nikiforakis & Engelmann, 2011). In fact, Nikiforakis (2008) shows that peer punishment institutions may not be able to enforce cooperative behavior at all in the presence of retaliation. A full quarter of all punishments imposed in his study triggered retaliation. Consequently, group members were deterred from punishing non-cooperators, in turn resulting in low levels of cooperation.

However, the evidence from this type of experiments is mixed on how strongly and consistently the possibility of retaliation undermines the effectiveness of peer punishment as an enforcer of cooperation (Chaudhuri, 2011). For example, one study where retaliatory punishment was possible shows no significant impact on cooperative behavior (Cinyabuguma, Page, & Putterman, 2006). There are also indications that antisocial punishment is more likely to trigger retaliation than punishment directed at non-cooperators (Cinyabuguma et al., 2006; Nikiforakis, 2008). If retaliation can strongly discourage sanctions against cooperators, in the right

circumstances the net effect of retaliation on cooperative behavior may even be positive. More importantly, we believe that besides mixed evidence, there is a much more serious potential problem that the literature on retaliation faces. Most previous experimental studies also preclude what can be considered a key feature of many of the cooperation problems of interest to social scientists. Decisions whether to contribute, to punish, or to retaliate in real life are often embedded in ongoing social relations, having not just a present, but also a past and a future. To fathom the impact of punishment and retaliation on contributions in cooperation problems, we need to consider how this embeddedness affects decisions to cooperate, punish or retaliate (Granovetter, 1985; Raub & Weesie, 1990).

Some experimental studies that allow for a limited extent of ongoing social interaction in a cooperation problem, paint an unclear picture. One study has shown that when retaliation can itself be retaliated against, retaliation is less frequent and the level of cooperative behavior is *below* that observed without retaliation (but with punishment), but *above* that observed when only the initial punishment is subject to retaliation (Denant-Boemont et al., 2007). But research on feuds, indefinitely extended cycles of retaliation, shows that the threat of feuds can lead to both more and less cooperation than usual (Nikiforakis & Engelmann, 2011; Nikiforakis, Noussair, & Wilkening, 2012). When interactions are non-anonymous and punishments can be used to respond to punishment, some retaliation occurs but there is no evidence that this decreases contribution levels (Flache, Bakker, Mäs, & Dijkstra, 2017).

Present study

Previous research suggests that embedding punishment decisions in ongoing relations decreases the prevalence and impact of retaliation. But these studies could not consistently exclude alternative explanations. Research on feuds, for instance, has not compared to the original single-step retaliation which had such a dramatic impact on punishment and cooperation (Nikiforakis, 2008; Nikiforakis & Engelmann, 2011; Nikiforakis et al., 2012). A study which did include multiple steps of retaliation and a direct comparison to the original single-step retaliation (Denant-Boemont et al., 2007) simultaneously altered several other aspects of the punishment institution. For example, contrary to Nikiforakis (2008) the treatment in Denant-Boemont et al. (2007) allowed participants to 'retaliate' against non-punishers and to punish those who punished or retaliated against third parties. This makes it difficult to assess the impact of ongoing social relations alone. For instance, any effect of repeated retaliation on contributions in Denant-Boemont et al. (2007) may also be due to second-order norm

enforcement (punishment of non-punishers; Bendor & Swistak, 2001). In this study, we therefore contribute to the literature by performing a systematic investigation of the effect of ongoing social interaction on cooperation, punishment, and retaliation. First, we replicate the experiment performed by Nikiforakis (2008). Then, we extend this experiment by gradually increasing the extent to which cooperation, punishment, and retaliation are embedded in ongoing social interactions, and punishments and retaliations can be responded to in future interactions. We find that retaliation is common whenever it is possible, and that retaliation deters punishment. However, we find no difference in cooperation between treatments with and without retaliation.

Additionally, our design allows us to differentiate between prosocial punishment (of defectors) and antisocial punishment (of cooperators). We find, consistent with previous research (Cinyabuguma et al., 2006; Nikiforakis, 2008), that antisocial punishment is subject to retaliation at a higher rate than prosocial punishment. We explore this finding further and show that, because of this difference in retaliation, prosocial punishments make up a greater proportion of all total punishments when retaliation is possible.

THEORY AND HYPOTHESES

Our key assumption is that whether effectiveness of a peer punishment institution suffers from retaliation depends on the extent to which participants are anonymized between interactions. In our experiment we vary this degree of anonymization systematically in a peer punishment institution with ongoing interaction, and assess how this variation affects the degree of cooperation, punishment and retaliation compared to a baseline condition in which retaliation is not possible at all (called *Punishment Only* treatment).

The theoretical difference of interest between experimental treatments is how easy it is for participants to identify which decisions their fellow group members made previously and to respond to those decisions. The easier the identification, the better a participant can respond to retaliation with renewed punishment in a subsequent interaction. At the operational level, the veil of anonymity behind which retaliators can hide is lifted stepwise in three different treatments, with corresponding theoretical expectations about how this affects the consequences of retaliation. In the *Anonymous Retaliation treatment*, retaliation is free of future consequences. This treatment replicates the counter-punishment treatment from Nikiforakis (2008), in which participants only receive general feedback on the amount of retaliation they receive and in which group members are anonymous from one period to the next,

since they are randomly given a new label at the start of each period. In the next treatment, the *Non-Anonymous Retaliation treatment*, participants are recognizable from one period to the next. However, participants are only informed of the total amount of retaliation they received and not of who retaliated against them. Finally, in the *Retaliation with Reminder treatment*, participants are also recognizable from one period to the next. In addition, we inform participants of how strongly their peers retaliated against them in the period before.

From the *Anonymous Retaliation treatment*, through the *Non-Anonymous Retaliation* treatment to the *Retaliation with Reminder* treatment, we gradually reduce participants' anonymity between periods. In principle, this reduction in anonymity applies to all decisions (contributions, punishments and retaliations) participants make. This allows, for instance, the emergence of alternative enforcement mechanisms for cooperation, such as reputations (e.g. Diekmann, Jann, Przepiorka, & Wehrli, 2014). It also allows participants to respond to punishments received in previous periods. However, in the context of the present study, we expect the impact of reduced anonymity on cooperation and punishment levels to be minimal. Both cooperation and punishment decisions can already be sanctioned even in the *Anonymous Retaliation treatment*, through punishment and retaliation respectively. Retaliations, however, are risk-free in the *Anonymous Retaliation treatment*. Reducing anonymity between periods opens up the possibility that participants sanction received retaliations, by renewed or increased punishment in subsequent periods. These sanctions, or the knowledge that these sanctions are possible, introduce potential negative future consequences for participants who retaliate. It has already been established that participants avoid (norm-violating) contributions (Chaudhuri, 2011; Simpson & Willer, 2015) as well as punishments (Nikiforakis, 2008) for which they expect to be sanctioned, and we expect the same to apply to retaliation. This leads us to our first hypothesis.

Hypothesis 1: There is less retaliation when participants are less anonymous between periods

Our second and third hypotheses formulate the expectations about the effects of retaliation on peer punishment and on cooperation which are suggested by the earlier literature. In this view, the main consequence of retaliation is that it undermines the effectiveness of peer punishment. When retaliation is not possible, peer punishment is risk-free. Yet, following earlier work, we would expect that when the possibility of

retaliation is introduced, this creates potential negative future consequences of punishing (Nikiforakis, 2008). In effect, the costs of punishment become higher. As a result, less punishment would be given. Second, assuming that less punishment is indeed given, peer punishment would be expected to become less effective at enforcing cooperative behavior. This adds up to hypotheses 2 and 3 which express the view currently prevailing in the literature.

Hypothesis 2: There is less punishment in treatments where there is more retaliation

Hypothesis 3: There is less cooperation in treatments where there is less punishment

Logically connecting hypotheses 2 and 3 with hypothesis 1 points to a new implication that our study allows to test. If, following hypothesis 1, treatments with less anonymity induce less retaliation, then according to hypothesis 2 this should also be the treatments with more punishment and thus, following hypothesis 3, with more cooperation. This suggests a more complex relationship between the anonymity of interactions and the viability of peer punishment institutions than previous research (e.g. Guala, 2012; Nikiforakis, 2008) has supposed.

Table 3.1. Overview of treatments

Treatment	*Retaliation possible*	*Anonymity between periods*
Punishment Only	No	High (Random relabeling)
Anonymous Retaliation	Yes	High (Random relabeling)
Non-Anonymous Retaliation	Yes	Moderate (No relabeling, no reminder)
Retaliation with Reminder	Yes	Low (No relabeling, with reminder)

METHOD

Treatments

Table 3.1 gives an overview of the main features of each treatment. Each treatment consists of 10 periods in which participants play a variant of the public good game (Ledyard, 1995) in a fixed group of 4 participants. Playing the public good game (PGG) repeatedly is necessary in order to allow subjects to respond to punishment and retaliation from their peers, by adjusting their behavior in subsequent periods. Keeping the group composition constant throughout the 10 periods of a treatment is

necessitated by our implementation of future consequences as something that happens across different interactions with the same group.

All treatments: The contribution stage and anonymity

The first part of each period, in all treatments, is the contribution stage. In this stage, all participants are endowed with 20 points. Participants then decide, simultaneously and without communicating, how many of their points to contribute to a group project. They can contribute anywhere from 0 (keeping all points to themselves) to 20 whole points (contributing all they have). Each point contributed is multiplied by 1.6 (as per Fehr & Gächter, 2000 & Nikiforakis, 2008), so that the total benefit to the group of each point contributed is 1.6 points. The collective productivity is thus maximized when all group members contribute all of their 20 points. The points produced in the group are then evenly divided across all 4 group members. This implies that any point a participant contributes to the group results in a collective return of 1.6 and a personal return of 0.4 (1.6 / 4) for that participant. Because contributions to the public good are collectively beneficial at a cost to the individual, high contributions are the measure of cooperation in a public good game. The contribution stage ends by reminding participants of their own contribution and showing them the total contribution of the group, their income from the group project, and their total income after the contribution stage.

In all treatments, the members of a particular group are randomly labeled 1 through 4. In the Punishment Only treatment and the Anonymous Retaliation treatment, participants are randomly relabeled with these numbers at the start of every period. This relabeling was originally done to avoid reputation effects (Nikiforakis, 2008). A side-effect is that retaliation is anonymous. That is: in the Anonymous Retaliation treatment victims of retaliation will not be able to tell which of their group members retaliated against them after the period ends. In both the Punishment Only treatment and the Anonymous Retaliation treatment, players cannot know in future periods who punished them in the punishment stages of previous periods. In the Non-Anonymous Retaliation treatment and the Retaliation with Reminder treatment, group members are randomly labeled only in the first period of the treatment and then keep the same label throughout the 10 periods of this treatment, so that identification across periods is possible.

Punishment Only treatment: contribution and punishment

The contribution stage is followed, in every period of every treatment, by the punishment stage. The punishment stage is replicated from Fehr & Gächter (2000) and Nikiforakis (2008). It starts by informing participants of the individual contributions of each of their fellow group members and giving them the opportunity to reduce the income of their peers. These decisions are made simultaneously, with no communication between participants. Punishing happens by the distribution of punishment points to other group members. Each point assigned to a fellow group member reduces that person's income, gained from the contribution stage, by 10% of the total. Income can never be reduced by more than 100%. Any participant who receives 10 or more punishment points will thus see their income reduced to 0. Assigning these points also carries costs for the punisher. The cost of punishing a particular peer follows a nonlinear progression, given by Table 3.2. Participants are not allowed to assign more punishment points than they can afford from their income gained from the contribution stage. At the end of the punishment stage, participants are reminded of their income from the contribution stage and are shown the punishment points they assigned to others with the associated cost, the punishment points they received with the associated reduction in income, and their income from this period after the punishment stage. Since participants pay for punishments before they know how many punishments they themselves are subjected to, participants can incur a net loss from a period of the experiment.

Table 3.2. Costs of punishment and counter-punishment

Punishment points you assign	0	1	2	3	4	5	6	7	8	9	10
Costs in points	0	1	2	4	6	9	12	16	20	25	30

In the Punishment Only treatment, each period ends after the punishment stage and participants are randomly relabeled for the next period. In the three following treatments, the period continues to the retaliation stage when the punishment stage completes.

Anonymous Retaliation treatment: contribution, punishment, and anonymous retaliation

In the three treatments where retaliation is possible, the punishment stage is followed by the retaliation stage. The retaliation stage is replicated from Nikiforakis (2008). This stage starts by informing participants of the number of punishment points they

received from each of their fellow group members. They are then given the opportunity to reduce the income of those who punished them (retaliation). To avoid strategic delay of punishment, participants cannot reduce the income of group members who did not punish them (Nikiforakis, 2008). Retaliation is done by assigning counter-punishment points, whereby the effects and costs associated with these points are exactly the same as in the punishment stage (Table 3.2), including the restriction that participants cannot distribute more counter-punishment points than they can afford from their remaining income of the current period. The retaliation stage ends by reminding participants of their income after the punishment stage and showing them the number of counter-punishment points they assigned to others with the associated cost, the number of counter-punishment points they received from others with the associated reduction in income, and their final income after the retaliation stage. In the Anonymous Retaliation treatment, punishers who receive retaliation do not learn which of the other group members retaliated against them. In addition, participants are randomly relabeled for the next period.

Non-Anonymous Retaliation treatment: contribution, punishment, and non-anonymous retaliation

The Non-Anonymous Retaliation (short: Non-Anonymous) treatment is identical to the Anonymous Retaliation treatment described above when it comes to the actions taken by participants. The only difference between the Anonymous and Non-Anonymous Retaliation treatments is that participants are not randomly relabeled at the start of each period, but keep the same labels throughout the 10 periods of the experiment. This makes it possible for participants to recognize group members in period *t* who retaliated against them in period *t-1, as long as they only punished one of their peers in period t-1*. For example, imagine that you punish only your peer labeled '3' and are retaliated against. You will know that group member '3' is the retaliator and that this person will also be labeled '3' in all future periods of the treatment. The implication is that with this information participants can use the punishment stage of period *t* to retaliate for punishment or retaliation received in period *t-1*. However, the non-anonymity is imperfect when participants punish more than one of their peers. Imagine that you punish two of your peers, players '3' and '4', and receive retaliation. Since at the end of the retaliation stage you are only shown the *total* number of counter-punishment points received, you are in the dark about whether peer '3', peer '4' or both retaliated. However, even in this case of incomplete information, an infuriated victim of retaliation has a better chance of striking the perpetrator than in

the *Anonymous Retaliation* treatment. Thus, potential retaliators are effectively less anonymous in the *Non-Anonymous treatment*.

Retaliation with Reminder treatment: contribution, punishment, and non-anonymous retaliation with reminder

In the Retaliation with Reminder (short: Reminder) treatment, we introduce a reminder which ensures that participants know who retaliated against them even if they punished more than one of their peers. In the punishment stage of every period, when participants are shown the individual contributions of their fellow group members, we also show them the number of punishment and counter-punishment points received from that particular peer in the previous period. This removes all anonymity from retaliation and quite explicitly introduces the opportunity to use the punishment stage of period *t* to respond to punishment or retaliation received in period *t-1*.

Procedures

The experiments were conducted at the Sociological Laboratory of the University of Groningen (http://www.soclab.nl). The Sociological Laboratory has a subject pool consisting of students at the University of Groningen. These students come from a variety of disciplines including sociology, economics, law, biology, physics, etc. Within the subject pool, psychology students and sociology students are overrepresented, compared to the population of students at the University. The rules of the Sociological Laboratory guarantee subjects that they will not be deceived in the experiment, and that they will be paid for their efforts.

Experiments took place in computer rooms prepared in such a way that, once they were seated, participants could not see the screens on which the other participants were playing. The experiments were programmed using the zTree software for economic experiments (Fischbacher, 2007). The experiment started with an introduction by the experimenter, explaining the rules of conduct within the lab and asking the participants to start reading the instructions. Instructions were provided on paper and were kept on hand throughout the experiment. Before the start of each treatment, participants were asked to read the section of the instructions pertinent to that treatment. During the experiment, subjects were always allowed to take notes. This ensured that in the treatments where participants might want to remember information across periods (e.g. the past behavior of their peers) they would not be required to memorize this information.

During the experiment, participants earned Monetary Units (MUs) dependent on their decisions and those of their group members. At the end of the experiment, these MUs were converted to Euros at a fixed rate, such that on average participants earned between €2 and €3 from the experiment, on top of their €5 show-up fee.

Data

Data were collected from a total of 152 participants across 11 sessions. One additional session had been planned, but had to be canceled as too few participants showed up. Participants played two, three or four of our treatments depending on the time available in the session. The treatments played in each session were chosen such that after the final session each treatment had been played at least once before and at least once after every other treatment.

All in all, we collected data on 3840 contribution decisions, 10440 punishment decisions, and 663 retaliation decisions. These decisions will form the units of analysis. Each decision is cross-nested in an individual (the participant who took the decision) and a group (the group this participant was part of at the time of the decision). This multilevel structure will be taken into account in all statistical analyses performed in the following section. We estimate multilevel models using the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015). In each analysis we include uncorrelated random intercepts for participants and groups, in addition to the level-1 error.

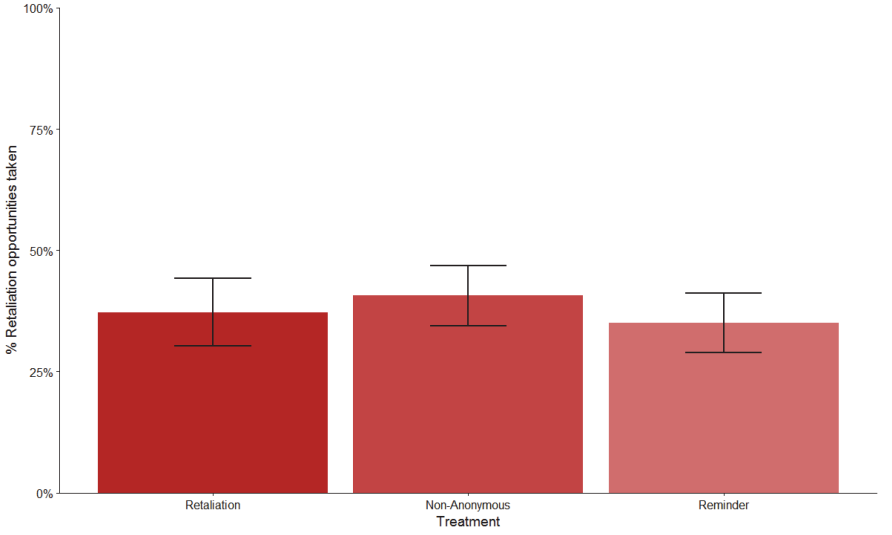
RESULTS

Retaliation

Overall, 686 punishments were dealt in the three treatments in which retaliation was possible (Anonymous, Non-Anonymous, and Reminder). In some cases, subjects on the receiving end of these punishments may have wanted to retaliate, but were unable to afford even a single counter-punishment point after the punishment stage. We identified all cases where subjects could not afford any retaliation (23 in total) and removed them, as we are interested only in decisions where our subjects *could* retaliate. This leaves us with 663 punishments which could have triggered retaliation. Of these 663 opportunities for retaliation, 250 (37.7%) were taken. This overall frequency of retaliation is higher than in the experiments by Nikiforakis (2008), where 25.7% of punishments triggered retaliation.

The frequency of retaliation differed little between the three treatments, ranging from 35.0% in the Reminder treatment to 40.8% in the Anonymous treatment (Figure 3.1). In the majority of cases in which retaliation occurred, retaliators dealt the minimum level of retaliation (1 counter-punishment point, 135 out of 250 cases). There are very few cases in which more than two counter-punishment points were given (36 out of 250 cases, never more than 6 points).

Figure 3.1. Retaliation opportunities taken by treatment



Note. Error bars represent 95% confidence intervals

Hypothesis 1 predicts that retaliation will be less prevalent in treatments with less anonymity. To test this hypothesis, we estimate a multilevel Poisson model suitable for the structure of this data. The 663 retaliation decisions our participants made are the units of analysis in this model. Previous research using the same or similar experimental designs (Nikiforakis, 2008) elected to estimate hurdle models (Cameron & Trivedi, 2013), which consist of a logistic regression model for the decision to retaliate or not and a linear regression model for the number of counter-punishment points given if retaliation occurred. A linear regression model would not be appropriate for our data, as the assumptions of homoscedasticity, linearity, and normality would all be violated even if we excluded all cases in which no counter-punishment points were given. We opt instead for Poisson models because the

counter-punishment points assigned closely resemble count data. All values are small positive integers (no less than 0, no more than 10) and, while they do not originate from a Poisson process as they are not based on the repeated occurrences of single events, their distribution does closely resemble one generated by such a process and they do represent counts (numbers of points given).

In the Poisson model, the number of counter-punishment points assigned in a retaliation decision is the dependent variable. The main independent variable in this model is the treatment in which the decision was taken. We include dummies for the Non-Anonymous treatment and the Reminder treatment, using the Anonymous treatment as our reference category. We control for the order in which treatments were played by including dummy variables for the position within the session in which the treatment was played (Position 2 and Position 3). Additionally, we control for the severity of the received punishment, for the period in which the decision was taken, and for the extent to which the received punishment was antisocial or unjustified. By the extent to which punishment was antisocial we mean by how much the target of punishment contributed more than the average of their peers, and by the extent to which it was unjustified we mean by how much the target of punishment contributed more than the punisher. Exact formulations of these concepts can be found in Nikiforakis (2008). Table 3.3 shows the results of this model.

We find that participants who received more severe punishment retaliated more (b = 0.281, p < 0.001 one-sided). We expected that as more potential future consequences to retaliation are introduced there would be less retaliation. We find that only in the Reminder treatment there is significantly less retaliation than in the Anonymous Counter-Punishment treatment (b = -0.343, p < 0.05 one-sided).

Result 1: Retaliation is common whenever it is possible. Introducing more future consequences to counter-punishment makes retaliation less frequent in the reminder treatment but not in the non-anonymous treatment. This result partially supports Hypothesis 1.

Table 3.3. Maximum likelihood estimates of a multilevel Poisson model for retaliation

Treatment variables	
Non-Anonymous	-0.133 (0.225)
Reminder	-0.343 (0.207)*
Control variables	
TreatmentPos:2	-0.114 (0.169)
TreatmentPos:3	0.292 (0.272)
Period	-0.023 (0.019)
Received punishment	
Extent unjustified	0.281 (0.048)***
Extent antisocial	-0.046 (0.088)
Constant	-1.186 (0.258)
Random effects (SD)	
Subject	1.029
Group	0.174
Case	0.0
N	663
Deviance	1283.1

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

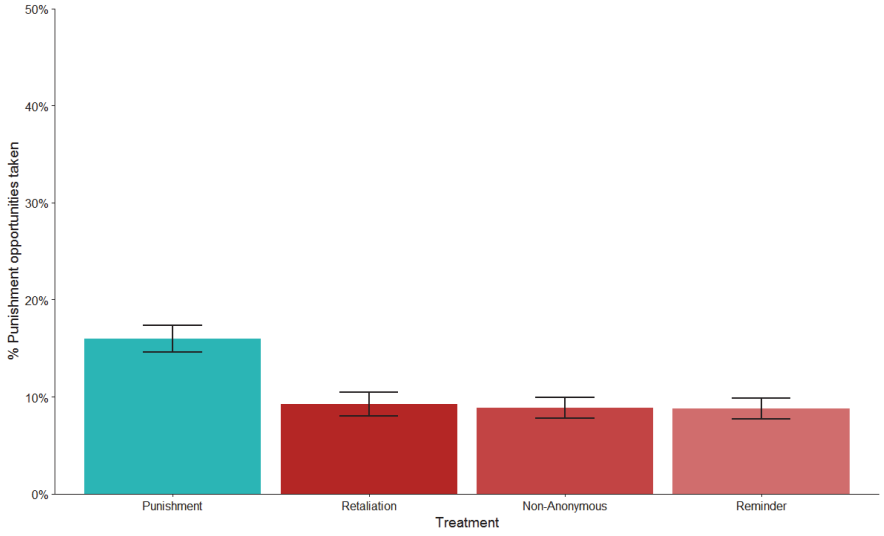
Punishment

The total of 10440 punishment decisions made by subjects across all four treatments resulted in 1127 actual punishments. This means that, overall, 10.8% of all punishment opportunities are taken. This percentage differs between treatments (Figure 3.2), with the highest observed percentage of punishment taking place in the Punishment Only treatment (16.0%)[3].

As expected based on Hypothesis 2, punishment is less frequent in the treatments with retaliation than in the Punishment Only treatment. In the majority of the cases in which punishment occurred, punishers dealt the minimum level of punishment (1 point, 651 out of 1127 cases). There are relatively few cases in which punishments of more than two points were given (171 out of 1127 cases, of which 3 cases with more than 6 points).

[3] This is lower than in comparable treatments of previous research (Nikiforakis, 2008), which can be explained by the fact that our results also show comparatively high average contributions. This implies that less punishment is required to enforce contribution norms.

Figure 3.2. Punishment opportunities taken by treatment



Note. Error bars represent 95% confidence intervals

The pattern of punishment levels, with a large number of zero-punishments and very few punishments of more than two points, is very similar to what we observed for retaliation decisions. To test Hypothesis 2, that there is less punishment where there is more retaliation, the same analytical strategy therefore applies: we will estimate a multilevel Poisson model where the dependent variable is the number of punishment points assigned. We use the Punishment Only treatment as our reference category and include as main independent variables three dummies representing the treatments with retaliation (Anonymous, Non-Anonymous and Reminder). We control for the order in which treatments were played by including dummy variables for the position within the session in which the treatment was played (Position 2, Position 3 and Position 4). We also control for the extent to which the contribution of a potential target of punishment was below that of the punisher (i.e. the extent to which punishment was justified), and for the extent to which the target's contribution was below the average contribution of other group members (i.e. the extent to which punishment was prosocial, see Nikiforakis, 2008). Finally, to control for trends over time, we include as further variable the period in which the decision was taken.

Table 3.4 shows the result of this model. We find that more punishment is given to a specific target the further that target's contribution is below the average of the group (b = 0.126, p < 0.001 one-sided). Similarly, more punishment is given when the target's contribution is further below that of the punisher (b = 0.116, p < 0.001 one-sided).

We expected to observe less punishment in treatments with more retaliation. As we know from the analysis of retaliation decisions (Table 3.3), retaliation occurs similarly often in the Anonymous and Non-Anonymous Retaliation treatments and significantly less frequently in the Reminder treatment. Nonetheless there is still a substantial amount of retaliation in the Reminder treatment. This implies that less punishment is expected in all of the treatments with retaliation than in the Punishment Only treatment, with the effect possibly being less pronounced for the Reminder treatment.

Table 3.4. Maximum likelihood estimates of a multilevel Poisson model for punishment

Treatment variables	
Anonymous	-0.546 (0.196)**
Non-Anonymous	-0.568 (0.169)***
Reminder	-0.702 (0.182)***
Control variables	
TreatmentPos:2	-0.458 (0.124)***
TreatmentPos:3	-0.338 (0.240)***
TreatmentPos:4	-0.399 (0.448)
Period	-0.068 (0.011)***
Punishment	
Extent justified	0.116 (0.018)***
Extent prosocial	0.126 (0.018)***
Constant	-2.625 (0.175)
Random effects (SD)	
Subject	1.114
Group	0.395
Case	0.965
N	10440
Deviance	7467.5

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

Table 3.4 shows that the probability a given punishment opportunity results in punishment is indeed significantly lower in all of the treatments with retaliation than in the Punishment Only treatment (all $p < 0.01$ one-sided). Somewhat unexpectedly, in the light of Hypothesis 2, the retaliation treatment in which retaliation is the least frequent (Reminder) is not the treatment with the most punishment.

Descriptively, these differences are quite large. In the Punishment Only treatment, 16.0% of punishment opportunities result in punishment. Averaged across all three treatments in which retaliation is possible only 8.9% of all punishment opportunities result in punishment. The frequency with which punishment occurs is thus roughly halved by the presence of potential retaliation.

Result 2: There is less punishment when retaliation is possible. This holds for all treatments with retaliation. We find partial support for Hypothesis 2: all treatments in which retaliation occurs show significantly less punishment than the baseline treatment. However, the retaliation treatment with the lowest amount of retaliation (Reminder treatment) is not the treatment with most punishment.

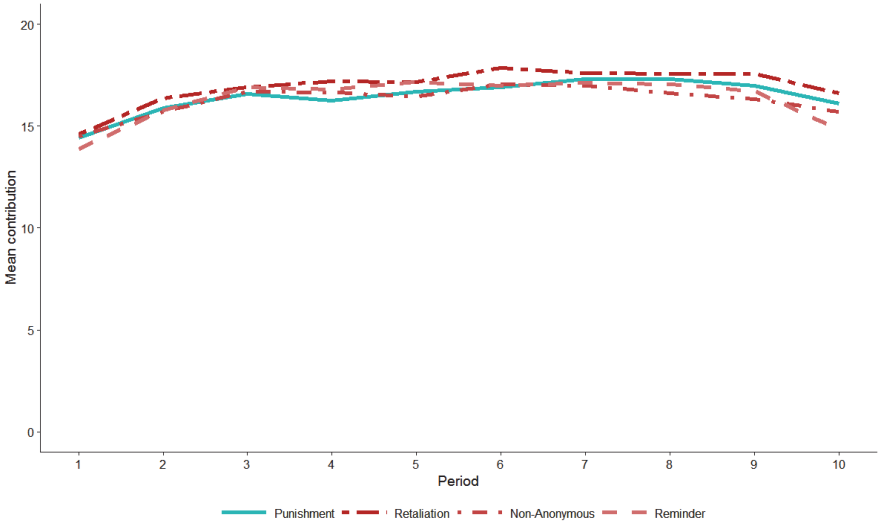
Cooperation

Contributions to the public good are the measure of cooperation in a Public Good Game. The Punishment Only treatment shows high contribution levels with a rising trend over time, consistent with earlier experiments using the same paradigm (Fehr & Gächter, 2000, 2002; Nikiforakis, 2008). Average contributions in the first round of this treatment are 14.41 points out of 20. The average contributions across all periods in the Punishment Only treatment are 16.45 points. These contribution levels are high compared to previous research (e.g. Fehr & Gächter, 2000, 2002; Nikiforakis, 2008), particularly in the first round. For example, Nikiforakis (2008) reports that in all treatments average contributions start at approximately 10 points in the first round. Contrary to our expectations (Hypothesis 3) and to previous research (Nikiforakis, 2008) we find little difference between contributions in the Punishment Only treatment and contributions in the three treatments which allow retaliation (Figure 3.3).

There are very few zero-contributions (only 103 out of 3480 contribution decisions resulted in a contribution of 0) and very many maximum contributions (1659 contribution decisions resulted in a contribution of 20). We see a very similar rising trend over time in all four treatments (Figure 3.3) up until the last few periods, in which endgame effects start to play a role.

To statistically test the impressions gained from the descriptive statistics we estimate two multilevel linear regression models. In these models, the dependent variable is the contribution by a particular subject in a particular period (i.e. one contribution decision).

Figure 3.3. Contributions by treatment over time



The main independent variables are once again three dummy variables for the retaliation treatments (Anonymous, Non-Anonymous and Reminder). The reference category is formed by our baseline: the Punishment Only treatment.

We also include the period in which a contribution decision was taken. In the second model, we add the interaction between the period and the treatment in which a decision was taken, to test if the trend over time differs between treatments. We control for the order in which treatments were played by including dummy variables for the position within the session in which the treatment was played (Position 2, Position 3 and Position 4).

Table 3.5 shows the results of these models. Model 1 shows that contribution levels tend to increase over time by an average of 0.141 points in every period ($p < 0.001$ one-sided). According to the theoretical mechanism we investigate, average contribution levels should be lower in circumstances where punishment is discouraged. We have established that punishments are discouraged in each of our treatments with retaliation and therefore expect lower contributions in these

treatments (Hypothesis 3). Surprisingly, there are no significant differences in average contribution levels between the Punishment Only treatment and any of the treatments with retaliation (Model 1).

Table 3.5. Maximum likelihood estimates of multilevel linear regression model for contributions

	Model 1	*Model 2*
Treatment variables		
Anonymous	-0.024 (0.894)	-0.080 (0.953)
Non-Anonymous	-0.269 (0.811)	0.171 (0.868)
Reminder	-0.036 (0.825)	0.393 (0.882)
Control variables		
TreatmentPos:2	1.284 (0.622)*	1.284 (0.622)*
TreatmentPos:3	1.731 (1.016)*	1.731 (1.016)*
TreatmentPos:4	1.112 (2.020)	1.112 (2.020)
Period	0.141(0.021)***	0.181 (0.040)***
Interactions		
Period x Anonymous		0.010 (0.060)
Period x Non-Anonymous		-0.080 (0.056)
Period x Reminder		-0.078 (0.056)
Constant	15.004 (0.666)	14.786 (0.692)
Random effects (SD)		
Subject	2.096	2.097
Group	2.504	2.504
Case	3.481	3.478
N	3480	3480
Deviance	19107.3	19103.2

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

We also find no difference in trend over time (Model 2) between the Punishment Only treatment and the Retaliation treatments. Overall the addition of interactions between the period in which a contribution was made and the three counter-punishment treatments does not significantly improve the model ($\chi^2 = 4.15$, $d = 3$, $p = 0.26$). Re-estimating Model 2 with each of the counter-punishment treatments as reference category shows that the trend over time is still positive and significant in all treatments. Overall, we conclude that contribution levels are not meaningfully different between any of our four treatments.

Result 3: There is not less cooperation in the treatments where there is less punishment. This result does not support Hypothesis 3.

Prosocial and antisocial punishment

The results so far show that retaliation does occur when people are given the opportunity. The results also show that retaliation occurs frequently whether it can be done anonymously or not. We find, as expected, that (the threat of) retaliation effectively deters punishment. However, this does not seem to have an adverse impact on the contributions participants make to the public good. One reason why the impact of retaliation on contributions may be smaller than one would expect based on its overall frequency is that retaliation has a different effect on antisocial punishment than on prosocial punishment.

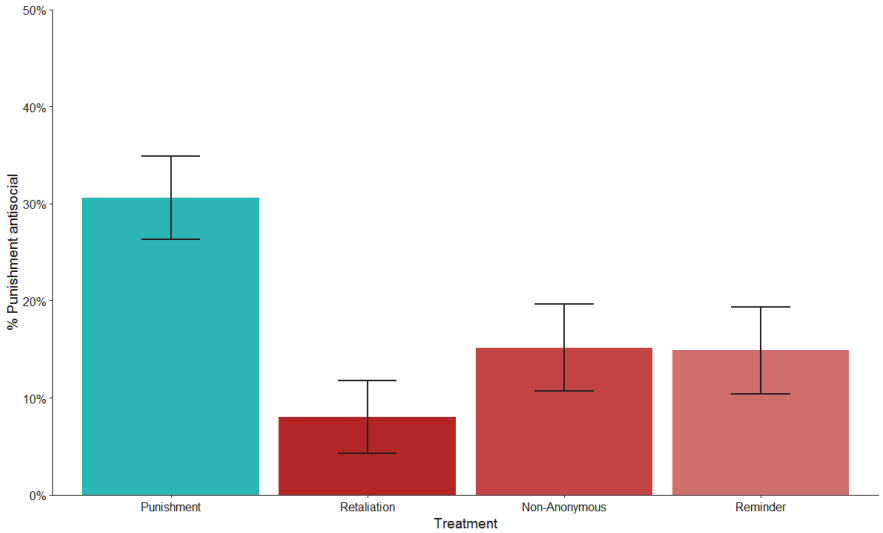
First, we note that antisocial and prosocial punishment are subject to retaliation at different rates. Participants retaliate in 45.5% of the cases if they were punished antisocially (i.e. even though they contributed more than the average contribution of the others in their group), and only retaliate in 35.8% of the cases if they were punished prosocially. This distinction is important since antisocial punishment is likely to lower contributions (punishing high contributors and thereby discouraging high contributions) while prosocial punishment results in higher contributions (punishing low contributors and thereby discouraging low contributions). We find that the socially harmful antisocial punishment is more likely to trigger retaliation than the socially beneficial prosocial punishment. The finding that antisocial punishment is subject to retaliation at a higher rate than prosocial punishment is not unique to our participants. The same was found by Nikiforakis (2008), who also notes that even though antisocial punishment is more prone to being retaliated against, the bulk of retaliations is aimed at prosocial punishers. This is because antisocial punishment is far less common than prosocial punishment. The same is true in our experiments: 76.0% of retaliations are responses to prosocial punishment.

Second, while the overall frequency of punishment is not all that high, *prosocial* punishment is quite frequent. Overall, across all treatments, 10.8% of punishment opportunities are taken. In comparison, again across all treatments, 27.6% of prosocial punishment opportunities are taken. Given that for every low contributor three peers will have an opportunity to punish, a low contributor's chance of being punished at least once is thus substantial. Even in the retaliation treatments, where Table 3.4 shows that punishment is deterred, there is substantial prosocial punishment. The overall rate of prosocial punishment in the treatments with retaliation is 25.0%, and

even in the treatment with the lowest rate (Reminder), 23.2% of prosocial punishment opportunities are taken. In the Punishment Only treatment, 34.7% of prosocial punishment opportunities are taken. This means that, rather than being nearly halved as the overall reduction in punishment suggested (Figure 3.2 & Table 3.4), prosocial punishment is only about 28% (on average) or at most 33% (in the Reminder treatment) less frequent when the threat of retaliation is present.

Figure 3.4 shows the percentage of all punishments given out which was antisocial, per treatment. In the Punishment Only treatment, without retaliation, 69.4% of punishments are prosocial. Averaged across the three treatments with retaliation, we find that 87.0% of all punishments are prosocial. It appears that antisocial punishment is more effectively deterred by (the threat of) retaliation than prosocial punishment. To test if the deterrent effect of retaliation is indeed stronger on antisocial punishment than on prosocial punishment we re-estimate the Poisson model of punishment presented in Table 3.4. The number of punishment points assigned is again the dependent variable. This time we include a dummy variable indicating whether punishing would be antisocial or not (0 = not antisocial, 1 = antisocial).

Figure 3.4. Percentage of punishments given which are antisocial



Note. Error bars represent 95% confidence intervals

We test the differential effect of retaliation on antisocial and prosocial punishment by including interactions between the treatment variables and this dummy variable indicating whether the punishment would be antisocial. We expect to see that the difference in the likelihood of punishment between the Punishment Only treatment and the retaliation treatments is greater for antisocial punishment than for prosocial punishment.

Table 3.6 shows the results of this Poisson model. Overall, antisocial punishment is less likely to occur than prosocial punishment (Model 1, $b = -1.760$, $p < 0.001$ one-sided). Antisocial punishment is also deterred more strongly than prosocial punishment by the presence of retaliation, as is evidenced by the significantly negative interactions between antisocial punishment and the treatments with retaliation (Model 2, all $p < 0.001$ one-sided). With the inclusion of these interactions, the main effects of the treatment variables in Model 2 show the decrease in *prosocial punishment* compared to the Punishment Only treatment. The decrease in prosocial punishment is smaller than the decrease in antisocial punishment (cf. the significantly negative interactions between Antisocial and the treatments in Table 3.6), and in the Anonymous retaliation treatment there is not significantly less prosocial punishment than in the Punishment Only treatment.

Result 4: Antisocial punishment is deterred more strongly than prosocial punishment by the possibility of retaliation. In the Anonymous retaliation treatment, prosocial punishment is not significantly less prevalent than in the Punishment Only treatment.

Table 3.6. Maximum likelihood estimates of a multilevel Poisson model for punishment (antisocial)

	Model 1	*Model 2*
Treatment variables		
Anonymous	-0.609 (0.200)**	-0.280 (0.208)
Non-Anonymous	-0.528 (0.172)**	-0.301 (0.182)*
Reminder	-0.709 (0.186)***	-0.480 (0.195)**
Control variables		
TreatmentPos:2	-0.364 (0.126)**	-0.350 (0.128)**
TreatmentPos:3	-0.382 (0.247)	-0.382 (0.251)
TreatmentPos:4	-0.447 (0.446)	-0.367 (0.460)
Group_Neg_Diff	0.039 (0.017)*	0.042 (0.017)**
Own_Neg_Diff	0.109 (0.016)***	0.105 (0.016)***
Period	-0.042 (0.010)***	-0.042 (0.010)***
Antisocial	-1.760 (0.090)***	-1.252 (0.120)***
Interactions		
Antisocial x Anonymous		-1.446 (0.262)***
Antisocial x Non-Anonymous		-0.715 (0.197)***
Antisocial x Reminder		-0.730 (0.199)***
Constant	-1.623 (0.178)	-1.768 (0.182)
Random effects (SD)		
Subject	1.147	1.134
Group	0.408	0.418
Case	0.830	0.797
N	10440	10440
Deviance	7059.6	7016.9***

*Note. One-sided p-values * < 0.05 ** < 0.01 *** < 0.001*

DISCUSSION

Our aim was to investigate whether retaliation becomes less frequent, and therefore less of a risk to the effectiveness of peer punishment as a means of enforcing contributions to a public good, when embedded in ongoing social interactions with less anonymity. We find that when the opportunity to retaliate exists, people will take advantage of it. Retaliation is common. It is particularly common in response to punishment which can be perceived as undeserved, by being directed from a low(er)

contributor to a high(er) contributor. We introduced potential future consequences gradually lifting the veil of anonymity behind which retaliators could hide. Step by step, we gave participants more of the necessary information to retaliate back and forth for the entire duration of the experiment. This only started to matter when we reminded participants explicitly that they were retaliated against and by whom. It is possible that the future consequences we introduced are simply not strong enough, and that more severe future consequences would deter retaliation. It is almost certainly possible to design an experiment in which retaliating carries such great risk that it would be deterred. The question is then whether such a design is still a valid representation of real social interactions. Future research might consider investigating how risky retaliating needs to be to significantly reduce the amount of retaliation that occurs. We believe that the design we chose, in which people are informed about retaliations they receive and can then use existing sanctioning institutions to retaliate in future interactions, is a reasonable starting point as it approximates how chains of punishment and retaliation could occur in reality.

Although our main manipulation has only a small effect on the prevalence of retaliation, our results raise some interesting points about the conditions under which retaliation hampers the effectiveness of peer punishment. In stark contrast to previous research on this topic (Denant-Boemont et al., 2007; Nikiforakis, 2008; Nikiforakis & Engelmann, 2011), we find that introducing the possibility of retaliation has no impact whatsoever on the contributions participants make. We find this even though retaliation is common and peer punishment is consequently deterred, conditions which resulted in strongly decreasing contributions in previous research (Denant-Boemont et al., 2007; Nikiforakis, 2008). We see two contributing factors which, together, may explain this difference.

First, antisocial punishment is more strongly deterred than prosocial punishment. Several explanations may be given for the different effects of (the threat of) retaliation on antisocial and prosocial punishment. One is that antisocial punishment in the Punishment Only treatment may in part be an attempt at retaliation in the absence of a formal mechanism to do so. Low contributors who are the recipients of punishment may suppose, accurately, that this punishment likely originates from high contributors in their group. They may then pre-emptively or in response to received punishment direct their own punishment at high contributors. When retaliation mechanisms are introduced the need for this imprecise attempt at retaliation is removed. Another explanation is that antisocial punishers may accurately

suspect that they are highly likely to be retaliated against. The threat of retaliation thus looms more menacingly over their heads than over those of prosocial punishers.

This suggests that we overestimate the negative effect of retaliation on the effectiveness of peer-sanctioning institutions as enforcers of cooperation when we look only at how strongly punishment is deterred overall. The deterrent effect on prosocial punishment, the kind of punishment which enforces high contributions, is much smaller. While overall punishment is only about half as frequent when retaliation is possible than when it is not, the decrease in prosocial punishment is only about one-third even in the worst case (the Reminder treatment). On balance, when retaliation is possible prosocial punishments make up a greater proportion of all punishments than when it is not.

That there is more retaliation towards antisocial punishers than towards prosocial punishers may in part be explained by the differential legitimacy of the different types of punishment. Contributing to collective action is valued, as evidenced by the attribution of status to high contributors (Willer, 2009). What's more, contributing is even considered moral, as evidenced by positive moral judgments given to high contributors and negative moral judgments given to low contributors (Simpson, Willer, & Harrell, 2017). While punished free-riders may be motivated to retaliate out of anger or out of a desire to discourage future punishment, they are likely to recognize that they deviated from what is considered desired and moral behavior and that this to some extent justifies punishment. Recipients of antisocial punishment, on the other hand, are likely to consider the punishment they received to be in violation of a social norm. The fact that prosocial punishment is not as strongly deterred as antisocial punishment by the possibility of retaliation may be explained both by the lower level of retaliation towards prosocial punishers and by the fact that the normative motivations for prosocial punishment may make it harder to deter by retaliation.

This tells us that the effect of retaliation on the effectiveness of peer punishment is not as disastrous as it at seems at first. This already suggests that deterring punishment does not necessarily result in a decrease in contributions. However, this cannot be the whole story. Since we exactly replicated Nikiforakis' (2008) counter-punishment treatment we have no reason to suspect that this differential deterrence of prosocial and antisocial punishment should occur in one experiment and not the other. We must, therefore, assume that this observation cannot explain the difference in results.

Given the fact that there is no difference in treatments, but a clear difference in contribution levels and their trend over time, a plausible second factor in the explanation is the subject population involved. We note that average contribution levels in the first period of any treatment are lower by several points in the experiments by Denant-Boemont et al. (2007) and Nikiforakis (2008) than in our experiments. This may be a rough indication that the percentage of individuals inclined towards cooperative behavior is lower in subject groups used in previous research than in ours. If this is the case, then it may well be that in our experiments there is enough prosocial punishment remaining to enforce high contribution levels even after some of it is deterred, while in the experiments by Denant-Boemont et al. (2007) and Nikiforakis (2008) the frequency of prosocial punishment drops below the level which is sufficient to deter free riding.

This raises a general point about the scope of situations in which the possibility to retaliate against peer punishments will negatively affect contributions made towards a public good. The introduction of retaliation will be particularly detrimental in situations where the peer punishment institution is already only just effective enough to sustain high levels of cooperation. These situations may be caused by low impact and high cost of the punishment institution itself, or by the balance between prosocial punishers and targets of prosocial punishment. When there are few prosocial punishers and the group contains a relatively large number of free-riders, a small amount of deterrence may be sufficient for peer punishment to become ineffective. However, in a group with many individuals willing to contribute and punish others for not contributing, a small set of free-riders will not be able to avoid punishment even if they deter some prosocial punishers by retaliating. In circumstances in which antisocial punishment is relatively common compared to prosocial punishment, the introduction of opportunities for retaliation may even result in higher contributions towards the public good. Given that we find antisocial punishment to be more effectively deterred by retaliation than prosocial punishment is, this may be true even in some situations where prosocial punishment is more common than antisocial punishment.

All in all, the results presented in this paper suggest that under the right circumstances peer punishment can be effective despite retaliation. This is evidenced by the high and increasing contributions in our experiment even though retaliation was not just possible but even common. Peer punishment institutions which are widely supported, with effective punishment mechanisms and many willing punishers, may be very robust against the threat of retaliation. At the same time, the impact of

retaliation on more fragile peer punishment institutions can be disastrous (as has been demonstrated by previous research). Future research should consider investigating properties of a peer punishment institution which make it more or less vulnerable to retaliation, one such property being the composition of the group to which this institution is applied.

Another question for future research is whether our results about the deterrent effect of retaliation on antisocial sanctions apply similarly if sanctions are not based on punishment but on (withholding) rewards. Earlier research suggested that the effectiveness of a sanctioning regime may suffer from the possibility of retaliation for reward-based as well as punishment-based regimes. When sanctions are reward-based, retaliation means that a sanctioned player withholds rewards from someone who withheld rewards from him in the past. Both theoretical work and experiments showed how this may reduce the effectiveness of sanctions because players are deterred from using reward for enforcing contributions to the collective effort (Flache, 1996, 2002; Flache & Bakker, 2012; Flache & Macy, 1996). However, results are as inconclusive as they are for retaliation of peer punishment (Flache et al., 2017). Future work could explore whether there may be a similar differentiation in the effects of retaliation between pro-social and antisocial sanctions for reward-based peer sanctioning as we found here for punishment-based peer-sanctioning.

In conclusion, our study suggests that the detrimental effects of retaliation in peer punishment institutions may have been overestimated by previous research. We implemented an experimental situation that captures an important feature of peer punishment in many real-life cooperation problems. Cooperation decisions, punishment decisions and retaliation decisions are all embedded in ongoing social interactions and can be responded to in future encounters. Our study suggests that this can deter specifically retaliation against pro-social punishments, resulting in no less contributions to a collective good than are given if there is no possibility to retaliate against punishment.