On the credibility of punishment in finitely repeated social-dilemma games

Ralph-C Bayer*
University of Adelaide

Abstract

Various experimental studies have shown that the availability of a punishment option can increase the prevalence of cooperative behaviour in repeated social dilemmas. A punishment option is only effective if it is perceived a credible threat. We investigate if credibility of punishment stems from standard strategic equilibrium considerations (Nash Equilibrium or Subgame Perfect NE). We find that punishment is credible due to non-strategic motivations (such as negative reciprocity) and that subgame perfection does not further improve credibility.

- JEL Codes: D03, D62
- Keywords: Cooperation, Punishment, Credible Threats

*School of Economics, Adelaide SA 5005, Australia. Email: ralph.bayer@adelaide.edu.au. Financial support from the Australian Research Council under the Discovery Program Grant DP120101831 is gratefully acknowledged.

1 Introduction

Theory and empirical evidence point to potential future punishment as one of the most effective factors for increasing the frequency of cooperative play in social dilemma situations. In classical game theory possible threats play an important role in supporting equilibria in repeated games where players choose socially efficient actions that are not equilibria of the stage game. In infinitely repeated games the equilibria that yield cooperation in social dilemmas are supported by players' willingness to punish each other in the case of uncooperative behaviour (see e.g. the version of the Folk Theorem by Friedman, 1971).[1]

In (long) finitely repeated games with multiple equilibria cooperation in early periods can be supported as a subgame-perfect equilibrium by the threat of switching to the stage-game equilibrium with lower payoffs if a deviation occurs (shown by Benoit and Krishna, 1985). Experimental studies have shown that punishment opportunities are not only effective if exercising them is part of an equilibrium strategy. In a seminal series of experiments Fehr and Gächter (2000) showed that punishment opportunities can increase cooperation considerably even if actually executing them is never part of a subgame-perfect equilibrium. In their voluntary contribution game with a punishment phase Nash equilibria exist, which are not subgame-perfect and where the threat of non-credible punishment is sufficient to induce cooperation.[2] While this leaves room for equilibrium considerations driving the effectiveness of punishment, in the literature more emphasis is placed on reasons other than equilibrium behaviour (e.g. negative reciprocity) for why punishment opportunities increase the occurrence of cooperation.

Many experimental studies have studied either the behaviour in repeated social dilemma games or the impact of explicit punishment stages in cooperation games. However, to our knowledge, there is no study that compares the effectiveness of punishment opportunities that are part of a subgame-perfect equilibrium and those that are just part of a Nash equilibrium.[3] This study provides such a comparison. A natural hypothesis is that punishment

[1] See Dal Bó and Fréchette (2018) for a comprehensive meta study on the determinants of cooperative play in infinitely repeated prisoner-dilemma games.

[2] For a general result on finitely repeated games see (Benoit and Krishna, 1987).

[3] The study closest to ours is Angelova et al. (2013), who compare subgame-perfect punishment options that are either strict or weak stage-game Nash equilibria.

which is subgame perfect is more effective, because it is only a credible threat in that case. An alternative hypothesis could read as follows: the fact that non-credible punishment has proven effective in many experimental studies shows that credibility (in the sense of subgame perfection) of punishment is not important. Nash equilibrium logic is sufficient. A third hypothesis for why it might not matter if punishment is an equilibrium in a subgame or not is that humans might use punishment in a non-strategic manner. If people punish others as a reaction to perceived unfairness, rather than as a disciplining device, then punishment becomes a credible option even if it is not subgame-perfect.[4] This study reports on experiments that were designed to discriminate between these three hypotheses.

We find that punishment is often non-strategic and therefore becomes credible regardless of the punishment being Nash in the stage game or not. The existence of punishment opportunities increases cooperation frequencies by anticipation of punishment but also by experience. Surprisingly, the increase in cooperation is not greater when punishment is a stage-game Nash equilibrium. Hence, subgame-prefection is not only not required for punishment to be credible, it does not even increase the credibility of the threat of being punished. We further find that the availability of punishment reduces over-all welfare, as its execution is costly and provokes damaging counter-punishment. The occurrence of counter-punishment does not depend on punishment being a stage-game Nash equilibrium. This shows that also counter-punishment is rather emotional than strategic. This is consistent with some findings on feuds and counter-punishment (Nikiforakis and Engelmann, 2011; Nikiforakis, 2010, 2008).

2 Three prisoners' dilemma games

In what follows, we present three versions of a prisoners' dilemma – the standard game and two extended versions that contain a punishment action. In one of the extended games punishment is a dominated strategy, while mutual punishment is a Nash equilibrium in the other. For finitely repeated versions of these games this has strong implications for the condi-

[4] This can be reconciled with standard theory by assuming that there are different types of which one type has a preference for punishment. Then there exist sequential equilibria with cooperation even if meeting a punishment type is rare (shown by Kreps and Wilson, 1982).

tions under which we can expect to observe players choosing the cooperative strategy. With standard preferences there is no Nash equilibrium in the original prisoners dilemma, where the cooperative action is ever played. In the finitely-repeated extended prisoners dilemma with non-equilibrium punishment, Nash equilibria exist that can support cooperative play in early stages. However, these Nash equilibria are not subgame perfect, as they are built on the non-credible threat of punishment. Finally, the extended prisoner's dilemma with punishment that is a stage game Nash equilibrium, has subgame-perfect Nash equilibria, in which cooperative play occurs in early stages.

First, take a version of the classic prisoners dilemma game shown in Table 1. If this game is played repeatedly but finitely many times, then the only Nash equilibrium prescribes that $\langle D, d\rangle$ is being played all the time. We know that for reasons that are unrelated to punishment, the fraction of cooperative behaviour in (one-shot) prisoners' dilemmas is positive. This is typically reconciled by assuming that subjects either have other-regarding preferences (such as in Fehr and Schmidt, 1999; Charness and Rabin, 2002; Cox et al., 2008), that beliefs about intentions are payoff relevant (e.g., Rabin, 1993), or that reputation building is possible as a consequence of some uncertainty about the rationality of players (Kreps et al., 1982).

	c	d
C	$5, 5$	$0, 8$
D	$8, 0$	$2, 2$

Table 1: A Prisoners Dilemma (Game 1)

Now turn your attention to Table 2, which shows an extended prisoners dilemma that also contains a punishment strategy. Playing the punishment strategy costs one monetary unit to the punisher and inflicts a damage of four. If both players punish, then both pay the punishment cost and also bear its damage, which results in a payoff of negative five for both players.[5] If this game is played T times, and we are looking for a subgame-perfect Nash equilibrium, then the only equilibrium entails that $\langle D, d\rangle$ is being played in

[5] Recall that in public goods games the cost-damage ratio of punishment has been shown to have to be below 1/3 for punishment to be effective (Nikiforakis and Normann, 2008). Here the ratio is lower if compared to the Nash outcome (i.e. 3/6) and identical if compared to a reference point of mutual zero profits.

all periods. The only stage-game Nash equilibrium $\langle D, d \rangle$ will be played in the last period regardless of history. Taking into account that play in the penultimate period cannot influence the continuation (if subgame-perfection is assumed), then $\langle D, d \rangle$ is going to be played in the penultimate period as well. This argument can be iteratively applied until the first period is reached.

	c	d	p
C	$5, 5$	$0, 8$	$-4, -1$
D	$8, 0$	$2, 2$	$-4, -1$
P	$-1, -4$	$-1, -4$	$-5, -5$

Table 2: A Prisoners Dilemma with non-Nash punishment (Game 2)

However, there are Nash equilibria that are not subgame-perfect, where in all but the last period $\langle C, c \rangle$ is played. To see this take the following trigger strategy for the row player. Play C in the first period. In all subsequent periods with $t < T$ play C if only $\langle C, c \rangle$ has been observed in the past. Otherwise play P. In the final period T play D if no prior deviation from $\langle C, c \rangle$ occurred. Otherwise play P. Together with the symmetric trigger strategy for the column player this strategy profile is a Nash Equilibrium. To see why this is an equilibrium, observe that if both players follow the equilibrium path, then in the last period they will play $\langle D, d \rangle$, where nobody has an incentive to deviate. In the penultimate period they are supposed to play $\langle C, c \rangle$. The best deviation is playing D and earning 8 instead of 5, a gain of 3. However, according to the strategy profile this will lead to $\langle P, p \rangle$ instead of $\langle D, d \rangle$ resulting in a payoff of -5 instead of 2 in the last round. So there is no incentive to deviate in the penultimate period. The same logic applies for earlier periods, where the one-off gain from a deviation remains the same, while the loss increases as the number of periods where $\langle P, p \rangle$ is played in response to a deviation increases. Hence, the threat of punishment that is not a stage-game Nash equilibrium can be sufficient to uphold cooperation if we do not require that threats are credible in the sense of subgame perfection.

Finally, consider the variant of the extended prisoners' dilemma in Table 3. In this version of the game we have two stage-game equilibria, which are $\langle D, d \rangle$ and now also mutual punishment $\langle P, p \rangle$. In this game the threat of punishment is credible as both players choosing the punishment strategy constitutes a stage-game Nash equilibrium. Consequently, by using credible punishment, cooperation can be implemented in earlier rounds as part of an SPNE. A trigger strategy is now not only Nash in the entire game but also in any subgame. To see this, take the same strategies as a above – C up to period $T-1$ if only $\langle C, c\rangle$ has been played before and otherwise P until the end of the game. Then play D in period T if only $\langle C, c\rangle$ was observed in the past, otherwise play P. The column player plays the corresponding symmetric strategy. Now for any history, in the final period we either observe $\langle D, d\rangle$ or $\langle P, p\rangle$, which both are Nash. So the continuation after $T-1$ periods is subgame-perfect. The best deviation in $T-1$ is to play D, which yields a one-off gain of 3, which will be offset by a loss of the same size, as it would trigger the continuation of $\langle P, p\rangle$ instead of $\langle D, d\rangle$. So there is no incentive to deviate in period $T-1$. Again, the one-off deviation gain remains the same for deviations in earlier periods (i.e. 3), while the loss increases since there are more periods with mutual punishment following a deviation. This strengthens the incentive to stick to cooperation earlier in the supergame.

	c	d	p
C	$5,5$	$0,8$	$-4,-1$
D	$8,0$	$2,2$	$-4,-1$
P	$-1,-4$	$-1,-4$	$-1,-1$

Table 3: A Prisoners Dilemma with Nash punishment (Game 3)

Note that in all the games above action profile $\langle D, d\rangle$ being played in all periods remains a subgame-perfect equilibrium, regardless of the other equilibria described. There are many more Nash equilibria in both games with punishment. While in Game 2 no further subgame-perfect Nash equilibria exist, there are many more in Game 3.

2.1 Hypotheses

In what follows, we develop a set of hypotheses. For a start, one could assume that adding a punishment actions does not alter play at all compared to the game without.

Hypothesis (I)**.** *Adding a punishment option does not increase the frequency of cooperative behaviour.*

Theoretically it is plausible that punishment options do not impact co-operation. In the case where punishment is never stage-game Nash, the refinement of subgame-perfection rules out any impact. Even if either new subgame perfect equilibria emerge with the punishment option, or if we don't require subgame perfection, then the initial defection equilibrium still exists and the impact of adding a punishment option on cooperation depends on equilibrium selection (i.e. coordination of the players). Beyond that, there are also behavioural arguments for why one might expect no impact. As the standard prisoners' dilemma yields already some cooperation with only a cooperative and a non-cooperative action, an additional action that leads to negative stage-game profits for both might be regarded as irrelevant by players. As a consequence punishment is neither executed, nor regarded as a threat and therefore has no impact on cooperation rates. The natural alternative hypothesis is as follows.

Hypothesis (IA). *Adding a punishment option increases the frequency of cooperative behaviour in at least one treatment.*

If we reject Hypothesis I in favour of IA, then we need further hypotheses to determine how behaviour relates to theory.

Hypothesis (II). *The addition of a stage-game Nash punishment option increases the frequency of cooperation by more than a non-Nash option.*

Empirical support for Hypothesis II would count as evidence that at least some of the increased cooperation can be attributed to the punishment option being only credible if it is stage-game Nash. In the extreme case that only in Game 3 (and not 2) increased cooperation frequencies are observed we can conclude that punishment threats are only credible if they are part of a SPNE. A natural alternative hypothesis is the following.[6]

Hypothesis (IIA). *A punishment option increases cooperation to the same extent irrespective of mutual punishment being stage-game Nash or not.*

If people do fear punishment regardless of it being a credible threat in the sense of subgame-perfection or not, then we would expect higher levels of cooperative behaviour in both extended Prisoners' Dilemmas compared

[6]We omit the further possible alternative hypothesis of more cooperation in Game 2, as is does not appear relevant.

to the standard version. Such a finding would give support the view that cooperation does not necessarily arise as the result of subgame-perfect play alone. One possible reason for this could be that subjects cooperate because they follow the Nash logic and still fear punishment despite its lack of credibility (in the sense of subgame perfection.)

Hypothesis (III)**.** *The observed increased cooperation in both extended games can be attributed to play following a Nash equilibrium that does not have to be subgame perfect.*

However, this is not the only reason why Hypothesis IIA might be supported by the data. Note that the Nash logic that leads to cooperation in Game 1 requires that punishment is never actually administered. From the literature in public goods games we know that punishment is actually observed. This hints at punishment that is actually credible due to factors other than standard equilibrium considerations.

Hypothesis (IIIA)**.** *Punishment is credible in both extended games, as punishment is triggered by factors other than standard equilibrium considerations.*

Suppose punishment is triggered by reciprocity considerations or by an emotion following perceived unfair behaviour. Similarly, one could think of the existence of a type, who enjoys punishing. Then punishment becomes a credible threat also in Game 2, where it is otherwise not credible (in the sense of subgame perfection). Note that a small fraction of players who punish for behavioural reasons might be sufficient to lead to a high number of punishers, as there might be an incentive for standard types to imitate the behavioural types (such as in Kreps and Wilson, 1982; Kreps et al., 1982).

Unfortunately, it is not possible to directly test Hypothesis III versus IIIA. For this purpose, we will use two different indicators drawn from observed behaviour. Firstly, we will look at the frequency of punishment in the last stage of the supergame. According to Hypothesis III we should only observe subjects playing the defection strategy (and no punishment) in Game 2, while both defection and punishment can be observed in Game 3. Non-strategic punishment, as postulated in Hypothesis IIIA, should lead to similar levels of punishment in the last period in both games. Similarly, we can compare the fraction of players that use the punishment strategy at least once in Game 2 and in Game 3. Similar and positive rates are again evidence for Hypothesis IIIA.

3 Experimental Design

In our experiments the participants played five supergames consisting of six repetitions of a stage game. Every supergame was played with a different opponent, which was clearly indicated to the subjects. An on-screen message read "This is a new phase. You are randomly matched with a new person. You will play six periods of the game with this person." In the *Baseline* treatment the stage game underlying the six supergames was the standard prisoners' dilemma (Game 1). We call the second treatment, where Game 2 was played repeatedly *Non-Nash,* as here punishment is not a stage-game Nash equilibrium. The final treatment, where subjects faced Game 3, will be referred to *Nash,* indicating that punishment is a stage-game Nash equilibrium.

Over-all 118 subjects, recruited via the online-recruiting system ORSEE (Greiner, 2015) from the student body of the three universities in Adelaide, participated in our six sessions. The experimental sessions were computerized and conducted at the Adelaide Laboratory for Experimental Economics (AdLab) using z-Tree (Fischbacher, 2007). Subjects earned experimental Dollars, which were converted to real Australian Dollars at the rate of one Australian Dollar for five Experimental Dollars. A session lasted about 60 minutes on average, for which subjects earned on average AUD 17.85.

4 Results

We start reporting our results by assessing the impact of punishment opportunities on cooperation rates. Next, we identify the underlying motivation for punishment behaviour and finish with assessing the welfare implications of punishment .

4.1 Cooperation rates

We start by comparing the frequencies of subjects choosing the cooperative action across treatments. Figure 1 shows the evolution of the fraction of cooperative actions chosen in the different stages of the supergames. In all three treatments the fraction of cooperative actions declines within a supergame. Cooperation rates in the standard repeated prisoners' dilemma are quite high in the first stage game (59 percent) but fall continuously to end at 11 percent in the last period. These dynamics are consistent to findings in other repeated prisoners' dilemma studies (e.g. Selten and Stoecker, 1986; Andreoni and Miller, 1993; Cooper et al., 1996; Normann and Wallace, 2012; Angelova et al., 2013). The fraction of cooperative actions in the two games with punishment is slightly higher but follows a very similar path.

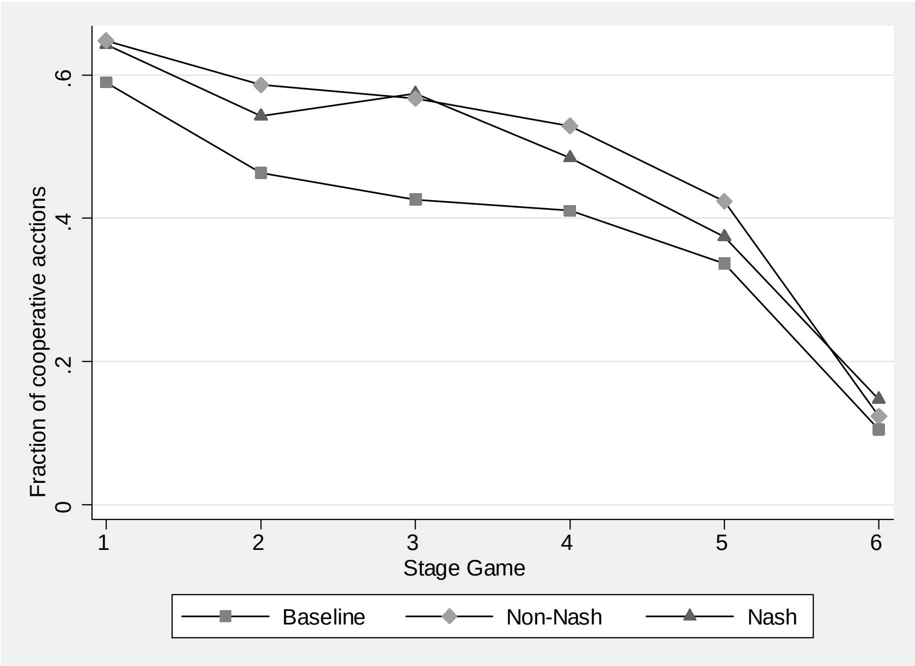


Figure 1: Average cooperation rates by period and treatment

In order to be able to test if there are significantly different rates of cooperation across the treatments we use the number of cooperative actions chosen per supergame. As a supergame lasts for six periods and there are two players, this measure ranges from zero to twelve. Figure 2 shows the distributions of the measure across the three treatments.

It is very instructive that the distributions are bimodal in all three treatments. Sustaining cooperation among two players in a supergame either works very well or not at all. While the distributions of the number of cooperative actions look almost indistinguishable between the two treatments with punishment opportunities, in the *Baseline* treatment the density is higher at the less cooperative end. More than 35 percent of supergames in the *Baseline* treatment result in none or only one cooperative action, while for the other two treatment less than 20 percent fall into this category. The average number of cooperative actions taken is similar in the *Nash* and *Non-Nash* treatments (5.53 vs. 5.75) and lower in the Baseline treatment (4.15). The differences between the two punishment treatments and the *Baseline* are statistically significant (*Nash* vs. *Baseline* $p < .06$, *Non-Nash* vs. *Baseline* $p < .03$; Mann-Whitney U-Tests)[7].

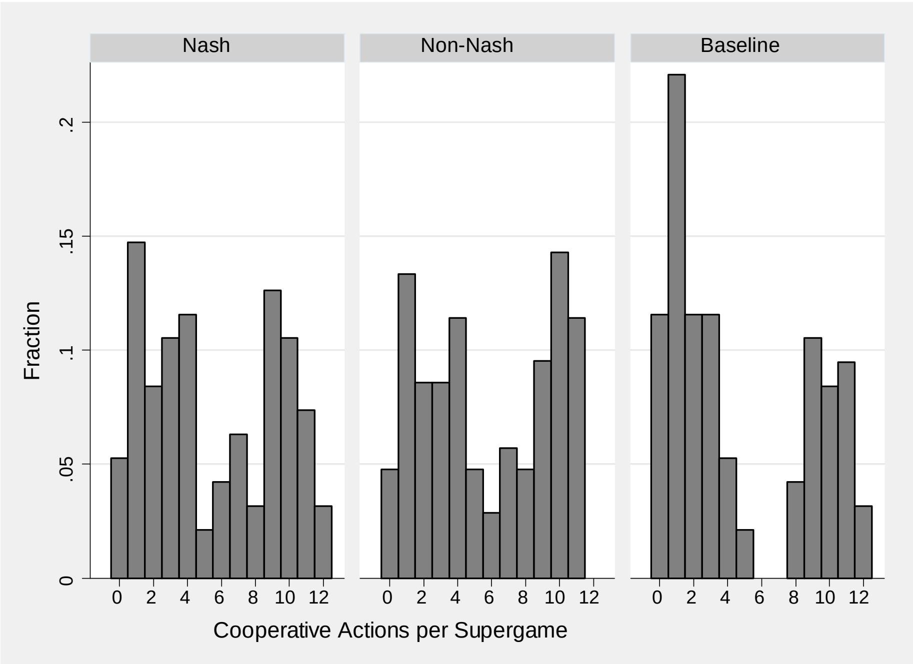


Figure 2: Distributions of the number of cooperative actions by treatment

Result 1. *Punishment increases the fraction of cooperative choices to the same extent regardless of it being credible (in the sense of subgame perfection) or not.*

This result supports the two alternative Hypotheses IA and IIA, which makes it necessary to look for evidence that allows us to discriminate between Hypotheses III and IIIA.

[7]The M-W test is the preferred test here, as it performs quite well for bimodal distributions, while the t-test lacks power. Kolmogorov-Smirnov tests on the equality of the distributions and t-tests with unequal variance correction lead to similiar results.

4.2 The Role of Punishment

One might expect that punishment behaviour differs depending on the game played (i.e. *Nash* or *Non-Nash*) and the behavioural assumptions made. If we assume that subjects play selfish equilibrium (subgame-perfect Nash), then we would only expect to see punishment in the *Nash* treatment. In a disequilibrium world, where punishment is used as a strategic tool, in order to induce others to cooperate, one might expect punishment to occur in both the treatments where it is available. However, one would expect punishment to be more prominent in earlier periods, as then the future cooperation induced is more likely to outweigh the cost. In contrast to this intuition, punishment frequencies are increasing in both treatments, with the fraction of punishment choices being highest in the last period in both treatments. Figure 3 documents this.

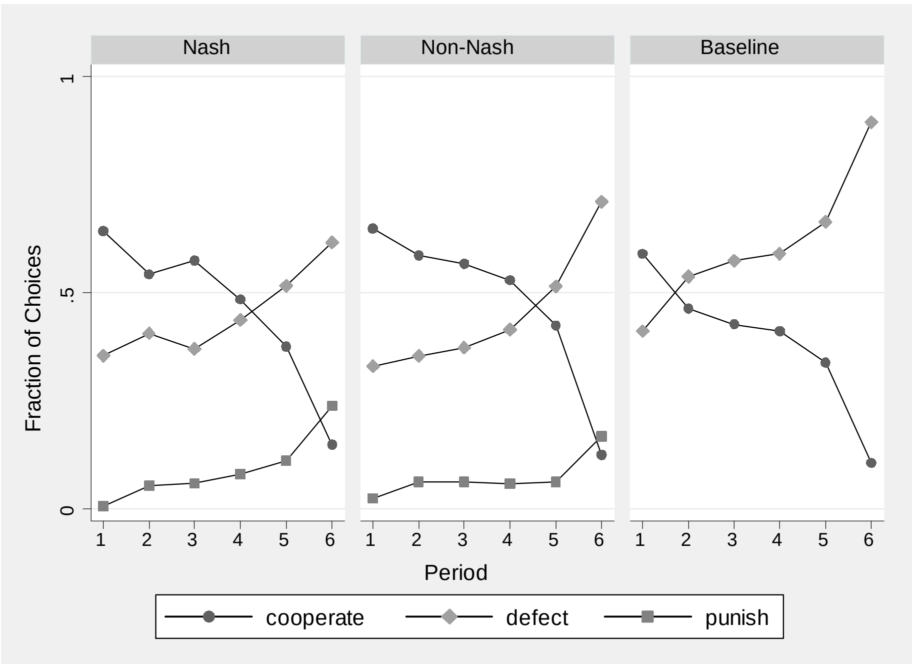


Figure 3: Fraction of Choices by Treatment over Time

The finding that punishment rates are highest in the last period in the *Non-Nash* treatment (and at comparable levels as in the *Nash* treatment – 16.7 vs. 23.7 percent) suggests that a large proportion of punishment is non-strategic and stems from negative reciprocity. Punishment in the final stage of the supergame in the *Non-Nash* treatment cannot be the result of equilibrium logic, as no Nash equilibrium exists where anybody actually punishes in the last stage. Moreover, off-equilibrium strategic teaching motives cannot be the cause, since there are no future periods. Hence, non-strategic motivation such as reciprocity are likely drivers of punishment behaviour in this treatment. This result is consistent with findings that punishment occurs in one-shot public goods games with punishment opportunities (Walker and Halloran, 2004).

As punishment is a stage game equilibrium in the *Nash* treatment, we cannot conclude the same for the *Non-Nash* treatment from the use of punishment in the last period. For this purpose, we will assess how similar punishment behaviour (with respect to frequency and dynamics) was in the *Nash* and *Non-Nash* treatment. The closer punishment behaviour in the *Nash* treatment resembles that in the *Non-Nash* treatment, where we have been able to attribute it to non-strategic behaviour, the more confident we can be that this is also the case in the *Nash* treatment. First, we test if the fraction of games where at least one player punishes in the last period differs across treatments. The raw percentages are 29.5 (*Non-Nash*) and 37.8 (*Nash*) percent. This difference is not significant ($p = .21$, test of proportions, two-sided). Secondly, a visual inspection of the punishment rate dynamics in Figure 3 gives the first clue that also punishment dynamics might not be very different across treatments.

With a similar aim, we estimate a logit model on whether a player ever used the punishment strategy in a supergame. In order to allow for correlation across opponents within a super-game, we clustered errors on the supergame level. As we were only interested in the impact of the treatment and on the predicted margins with respect to the treatment, we relegate the estimation results to the appendix. The coefficient on the treatment dummy for the *Non-Nash* was highly insignificant ($p = .704$). Figure 4 plots the predicted probabilities of a player punishing at least once in a supergame by treatment, which result from the regression. The predicted probabilities are virtually identical (*Nash* .31; *Non-Nash* .33) and clearly statistically not significantly different ($p = .70$).

With this we reach the preliminary conclusion that punishment behaviour is quite similar in the *Nash* and *Non-Nash* treatment.

In our final piece of assessment we compare the history dependence of play across treatments. If we cannot find any difference between *Nash* and

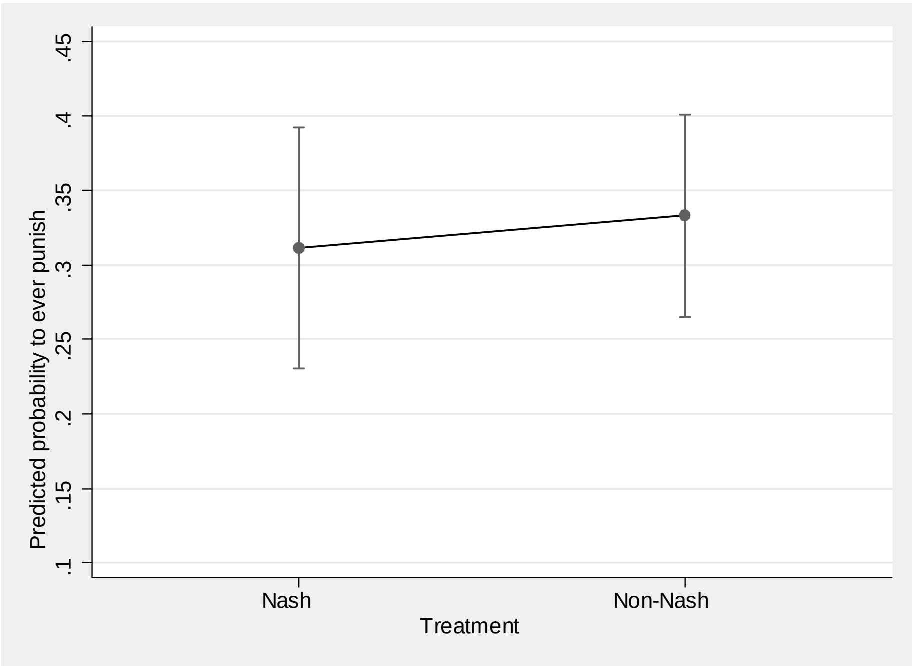


Figure 4: Predicted margins to ever punish by treatment

Non-Nash, then this together with the findings above is strongest evidence for no differences in behaviour in the two treatments with punishment options. For this purpose, we ran a multinomial logit regression. We estimate how previous play of the opponent, the treatment and the previous-play-treatment interaction influence the likelihood that a player cooperates, defects or punishes. We control for demographics of the subjects as well as for time trends within and across supergames.[8] Table 4.2 reports the results. The base behaviour is"defect." On the left of the Table we report the estimated coefficients for the influence of the independent variable on the likelihood to play "cooperate" instead, while we report the coefficients for choosing the punishment strategy on the right. The standard errors are shown in parentheses and stars denote significance on the five (single star) and one-percent level (double star).[9]

We are mainly interested in how past choices of the opponent impacts current play. More precisely, we want to know if there differences across

[8]We allow for clustering of errors on the subject level.

[9]We report the average marginal effects of primary variables only, as the average marginal effects of the interaction terms are hard to interpret.

decision	Coefficient (base: defect) cooperate		punish	
otherdecision$_{t-1}$×treatment (base: *cooperate*×*Nash*)				
cooperate×*NoNash*	.010	(.183)	.178	(.452)
defect×*Nash*	-2.794^{**}	(.258)	1.418^{**}	(.397)
defect×*NoNash*	-2.918^{**}	(.251)	1.327^{**}	(.376)
punish×*Nash*	$-.869^{*}$	(.439)	2.641^{**}	(.530)
punish×*NoNash*	-1.339^{**}	(.394)	2.306^{**}	(.532)
stage	$-.489^{**}$	(.050)	$.197^{**}$	(.068)
male	.211	(.145)	$.601^{**}$	(.230)
university level (base: *pg coursework*)				
pg research	1.028^{**}	(.322)	$-.349$	(.430)
undergraduate	$-.001$	(.279)	.011	(.305)
Controls (age, course, maths level, supergame)				
included	*yes*		*yes*	
Log Pseudo$\mathcal{L}$	-1366.820			
Observations	2000			

Table 4: Multinomial logit explaining choices in the punishment treatments

treatments how players react to the past play of the opponent. We first test if the differences between the six coefficients for the *Nash* treatment and their corresponding coefficients for the *Non-Nash* treatments are jointly zero. The test does not allow us to reject this hypothesis ($p > .97$, F-Test with the null hypothesis that the coefficients are jointly equal). Testing all differences separately does not yield a single significant difference even without any p-value correction for multiple hypothesis testing. This allows us to conclude that the dynamics of play are not different across treatments.

In order to get a better feel for the dynamics, in Figure 5 plot the predicted probabilities (from the multinomial logit) and the confidence intervals of playing a certain action conditional on the past play of the opponent. It is easy to see that the likelihoods of playing a certain action conditional on the opponent's past action are virtually identical for the two punishment treatments. As expected pairwise tests yield no significant differences even without any correction for multiple hypotheses testing.

This implies that the way subjects react to past behaviour of their oppon-

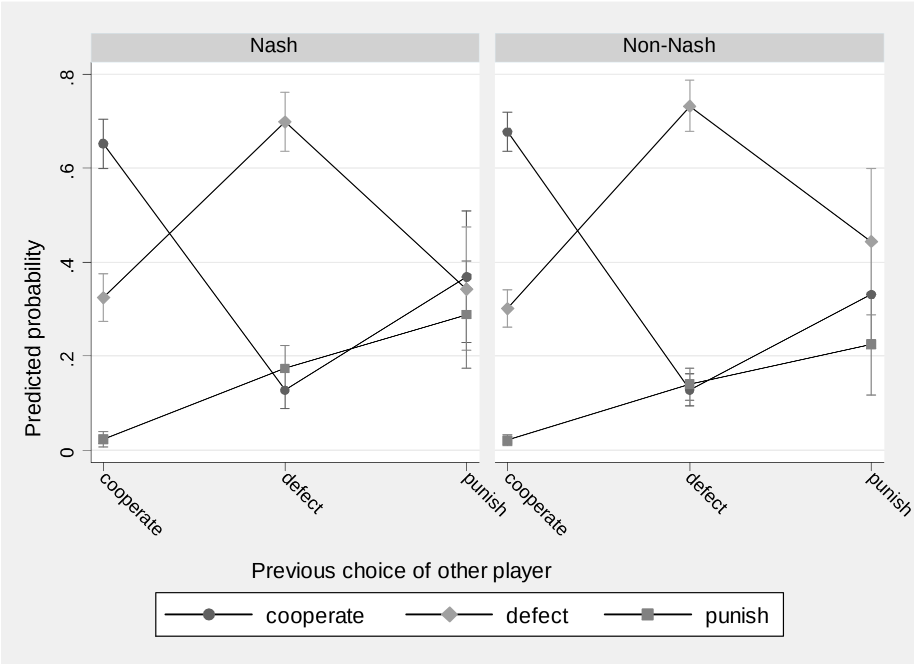


Figure 5: Predicted probabilities for actions played depending on opponent's past play

ent does not differ with respect to whether punishment is a Nash equilibrium of the stage game or not. Therefore, the likelihood of punishing conditionally on the action of the opponent in the period before is independent from punishment being a stage-game Nash strategy. This finding provides further support that the actual punishment is motivated by other factors than equilibrium considerations. This further implies that the threat of punishment is credible in both treatments.

As expected, we find that the likelihood of choosing the punishment action is greater after the opponent defected than after the opponent cooperated ($p < .001$, Wald Test, jointly for both treatments). Less intuitive is that being punished in the period before makes it even more likely for a subject to choose the punishment strategy ($p < .05$, Wald Test, jointly for both treatments) than after defection. The fact that choosing the punishment strategy is most likely after having been punished, documents the occurrence of counter-punishment and explains a large fraction of the upwards trend of punishment over time.[10]

[10] There still remains an unexplained upwards trend of about two percentage points per

Moreover, having been punished the period before, reduces the probability of cooperating significantly ($p < .001$, Wald Test, jointly for both treatments) compared to the case where the opponent cooperated in the period before. Punishment is at least more effective in inducing cooperation of the opponent than defecting ($p < .001$, Wald Test, jointly for both treatments). This implies that executed punishment is effective, to a certain extent, at inducing future cooperation. However, in many cases, punishment leads to counter-punishment. This can potentially be attributed to the absence of a feeling of guilt of the punished as shown by Hopfensitz and Reuben (2009). Below we will show that this escalation of punishment is partly responsible for a negative effect of punishment opportunities on average welfare, despite the increase in the fraction of cooperative actions (see Nikiforakis and Engelmann, 2011, who obtain similar results in a public goods setting).[11] We briefly summarize our results.

Result 2. *The use of punishment in the Non-Nash treatment cannot be rationalized by Nash equilibrium logic and has to be attributed to non-strategic motivations.*

Result 3. *The use of punishment does not differ across treatments.*

The two results above provide evidence for the alternative Hypothesis IIIA. Punishment is credible in both treatments because subjects use it non-strategically.

Result 4. *Punishment is effective at increasing future cooperation (compared to playing defect) but often also provokes counter-punishment.*

4.3 Efficiency

The higher fraction of cooperative choices in the punishment treatments does not necessarily imply that social welfare is higher in those treatments. The addition of a punishment option, might not only lead to more cooperation due to the threat of punishment but also to welfare losses due the actual use of the punishment action. Depending on which effect dominates punishment options increase or decrease social welfare.

period.

[11] Some interesting results that are not related to our immediate research question are that PhD students are more cooperative than undergraduate and coursework masters students. Economics and science students are less cooperative than medicine, law and engineering students and males tend to punish more often than females.

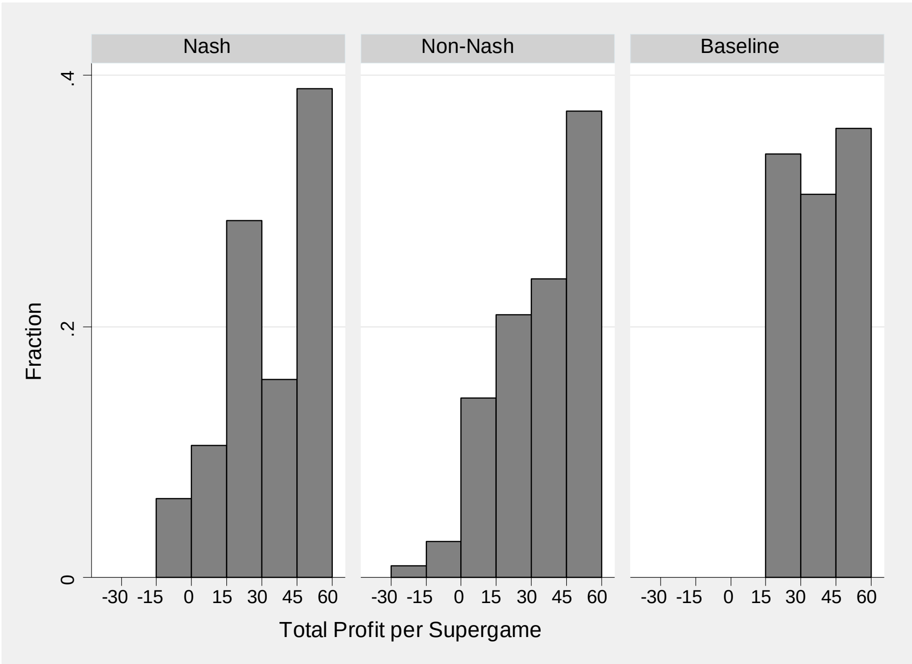


Figure 6: Total Welfare Distributions in the Different Treatments

Figure 6 plots the distribution of welfare per supergame (i.e. the sum of total profits for the two players in a supergame). Note that the minimum welfare in the Baseline treatment is 24, while it can be much lower (and even negative) in the two other treatments if subjects choose the punishment strategy. Actual punishment had a negative influence on the welfare in many supergames. In both treatments, where punishment was possible about 30 percent of the supergames led to welfare lower than the minimum welfare in the standard Prisoners' Dilemma. In other words, in the punishment treatments about 30 percent of the supergames ended with lower payouts than if the players had just played the defect equilibrium in all rounds. This welfare-destroying impact of actual punishment is not offset by the slightly higher proportion of supergames with near maximum welfare (45 to 60) in the punishment treatment.[12]

Average welfare is very similar in the punishment treatments (*Nash* 33,67, *Non-Nash* 34.2), while it is significantly higher in the *Baseline* treat-

[12] This is similar to the observation in stranger-matching public goods games with punishment, where the increased level of contributions is not sufficient to outweigh welfare lossed from exectuted punishment (Fehr and Gächter, 2000).

ment (39.36, $p < .018$ vs. *Nash* and $p < .026$ vs. *Non-Nash*: two-sided, two-sample t-tests).[13] One important reason for the differences in welfare stems from the observation that punishment is used non-strategically. In particular punishment and counter-punishment in the last period as observed are welfare damaging, as they cannot have any positive effect on future cooperation.

Result 5. *Even though punishment options induce more cooperative choices, they decrease over-all welfare, since punishment is actually executed. This effect does not depend on punishment being a stage-game equilibrium.*

5 Conclusion

Punishment opportunities have been shown to be very effective in public goods games, despite of being non-credible in the sense of subgame-perfection (Fehr and Gächter, 2000).[14] It was unclear, if the effectiveness of punishment there has to do with Nash equilibrium logic in the supergame, with punishment being credible as a strategic teaching tool or as a means of exerting negative reciprocity. This paper uses variants of prisoners' dilemma games in order to investigate this question. We find that the effectiveness of punishment is neither related to Nash or subgame-perfect Nash equilibrium logic. Punishment is credible, as it is motivated by reciprocity. While punishment opportunities increase the fraction of cooperative play, welfare decreases, as the cost for punishment exceeds the welfare gains from slightly more cooperation. The welfare damaging effect of punishment opportunities is stronger than expected, since some subjects react to punishment with further welfare-damaging counter-punishment.

[13] Here a non-parametric Mann-Whitney test is not appropriate, as the null-hypothesis of equal distributions can never be satisfied due to the different domains, which makes the interpretation of the p-values impossible. Despite the well-known skewness issues, the Welch-Satterthwaite corrected t-test for unequal variances, which we chose, performed best in a recent Monte-Carlo study on data with similar characteristics (inhomogenous skewness and sample size of about 100) as ours (Fagerland and Sandvik, 2009).

[14] The effectiveness seems to depend on factors such as the feedback format though (Nikiforakis, 2010)

A Logit regression on punishment in a supergame

Here we report the panel logit regression that estimates how the treatment and other factors impact on the likelihood that a subject punishes at least once in a supergame. In the text we referred to the absence of differences across treatments.

	ever punish coefficient	*ever punish* marginal effect
Treatment (base: Nash)		
Non-Nash	0.111	0.022
	(0.292)	(0.057)
Age (base: under 26)		
over 26	-0.560	-0.105
	(0.365)	(0.063)
Subject of Study (base: Arts)		
Commerce/Finance	-1.885***	-0.393***
	(0.498)	(0.106)
Economics	-0.079	-0.019
	(0.753)	(0.181)
Engineering	-0.582	-0.140
	(0.483)	(0.115)
Law	-0.807	-0.193
	(0.587)	(0.136)
Medicine	-1.814**	-0.383**
	(0.598)	(0.120)
Science	-0.497	-0.120
	(0.469)	(0.112)
Higher maths (base: No)		
Yes	-0.061	-0.012
	(0.306)	(0.061)
Degree (base: Postgraduate Coursework)		
Postgraduate Research	-1.011	-0.198
	(0.538)	(0.101)
Undergraduate	-0.486	-0.102
	(0.399)	(0.085)
Gender (base: Female)		
Male	0.075	0.015
	(0.240)	(0.047)
Constant	0.800	–
	(0.540)	
Observations	400	400

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B Sample Instructions

(*Here are the instructions for the baseline treatment. The instructions for the other treatments are identical up to the added action in the screenshot and the reference table at the end.*)

Experimental Instructions

Welcome to the experiment. Before we start, please read the instructions carefully.

During the experiment, your earnings will be calculated in points rather than dollars. Accumulated points will then be converted to Dollars at the following exchange rate at the end of the session to determine your payment:

5 points = AUD $1.00

You will be paid in cash immediately after the experiment. You are not allowed to communicate with other participants during the experiment. Should you have any questions, please raise your hand and we will attend to you individually. Failure to comply with the outlined rules will result in exclusion from the experiment and we reserve the right to forfeit your payment.

Summary

You will be playing 5 identical games consecutively.

Each game consists of 6 rounds and you will be asked to select one action per round. You will be playing this game with another participant who will be randomly assigned by a computer.

After every 6 rounds, you will be randomly paired with another participant until you have played a total of 5 games.

The Game

This is a 2-player game. After you have been randomly assigned to another participant by a computer, you and this other player will play a game consisting of identical rounds. In each round, you will be asked to choose an action. Similarly, the other player will also be asked to choose an action at the same time.

You will be presented with two actions (A or B) to choose from. The other player will also be presented with two actions (X or Y) to choose from.

Your payoffs for every possible combination of actions that you and the other player may make are shown on the same screen in a table. The other player's payoffs will be displayed in a similar fashion in a separate table beneath your payoffs table.

You then indicate your choice of action at the bottom of the screen and finalize your decision by clicking the "OK" button.

Payoffs

Both yours and the other player's choice of action, and respective payoffs for the current round will then be revealed after you have both finalized your decisions.

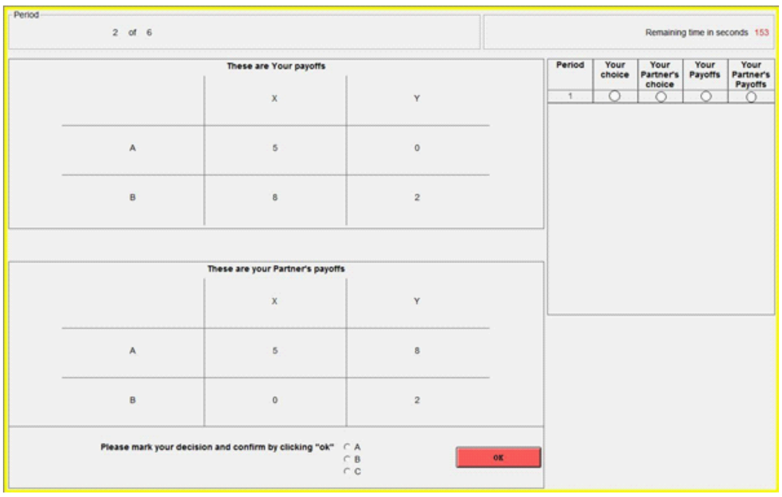
The final payoff you receive in each round depends on:

1. The action that you have selected; and
2. The action that the other player has selected.

The payoff the other player receives depends on:

1. The action he/she has selected; and
2. The action you have selected.

The following is a screenshot to familiarize you with what to expect during each round.



The header on the top left hand corner of the screen indicates the current round you and the other player are playing. The table beneath the header shows your payoffs for all possible combinations of yours and the other player's actions. The box on the right side of the screen records your payoffs and your partner's payoffs for every round played. [*Note: Every game will be the same throughout the whole experiment. As a guide, please refer to the table attached at the back of these instructions which tells you your payoffs corresponding to all possible combinations of actions that you and the other player may choose.*]

A new game commences and you will be randomly paired with another participant after every 6 rounds. This process repeats until 5 games have been played. After all 5 games have been played, your total profit will be recorded and you will be paid in cash.

[*Note: Please refer to the table attached at the back of these instructions which tells you your payoffs corresponding to all possible combinations of actions that you and the other player may choose*]

-End of Instructions-

Your Partner's Payoff / Your Payoff	X	Y
A	5 5	0 8
B	8 0	2 2

The table above illustrates payoffs in the game. Your payoffs are denoted by numbers within the shaded triangles whereas the other player's payoffs are denoted by numbers within the un-shaded triangles. For example, if you chose 'B' and the other player chose 'X' in particular round, payoffs for that particular round are:

Your Payoff = 8

Your Partner's Payoff = 0

References

Andreoni, J. and J. H. Miller (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal 103*(418), pp. 570–585.

Angelova, V., L. V. Bruttel, W. Güth, and U. Kameke (2013). Can subgame perfect equilibrium threats foster cooperation? an experimental test of finite-horizon folk theorems. *Economic Inquiry 51*(2), 1345–1356.

Benoit, J.-P. and V. Krishna (1985). Finitely repeated games. *Econometrica 53*(4), pp. 905–922.

Benoit, J.-P. and V. Krishna (1987). Nash equilibria of finitely repeated games. *International Journal of Game Theory 16*(3), 197–204.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1996). Cooperation without reputation: Experimental evidence from prisoner's dilemma games. *Games and Economic Behavior 12*(2), 187 – 218.

Cox, J. C., D. Friedman, and V. Sadiraj (2008). Revealed altruism. *Econometrica 76*(1), 31–69.

Dal Bó, P. and G. R. Fréchette (2018, March). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature 56*(1), 60–114.

Fagerland, M. W. and L. Sandvik (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials 30*(5), 490 – 496.

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

Fehr, E. and K. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

Fischbacher, U. (2007). Z-tree - Zurich toolbox for readymade economic experiments. *Experimental Economics 10*(2), 171–178.

Friedman, J. (1971). A non-cooperative equilibrium for supergames. *Review of Economic Studies 38*, 1–12.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125.

Hopfensitz, A. and E. Reuben (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal 119*(540), 1534–1559.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory 27*, 245–252.

Kreps, D. M. and R. Wilson (1982). Reputation and imperfect information. *Journal of Econimic Theory 27*, 253–279.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics 92*(1-2), 91–112.

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior 68*(2), 689 – 702.

Nikiforakis, N. and D. Engelmann (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization 78*(3), 319 – 332.

Nikiforakis, N. and H.-T. Normann (2008). A comparative statics analysis of punishment inăpublic-good experiments. *Experimental Economics 11*(4), 358–369.

Normann, H.-T. and B. Wallace (2012). The impact of the termination rule on cooperation in a prisoner's dilemma experiment. *International Journal of Game Theory 41*(3), 707–718.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review 83*(5), 1281–1302.

Selten, R. and R. Stoecker (1986). End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization 7*(1), 47 – 70.

Walker, J. and M. Halloran (2004). Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics 7*(3), 235–247.