Journal of Public Economics 94 (2010) 78–86

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Unobserved punishment supports cooperation

Drew Fudenberg [a], Parag A. Pathak [b,*]

[a] Department of Economics, Harvard University, United States
[b] Department of Economics, MIT, United States

ARTICLE INFO

Article history:
Received 24 May 2009
Received in revised form 28 September 2009
Accepted 14 October 2009
Available online 22 October 2009

Keywords:
Public-goods experiments

ABSTRACT

Costly punishment can facilitate cooperation in public-goods games, as human subjects will incur costs to punish non-cooperators even in settings where it is unlikely that they will face the same opponents again. Understanding when and why it occurs is important both for the design of economic institutions and for modeling the evolution of cooperation. Our experiment shows that subjects will engage in costly punishment even when it will not be observed until the end of the session, which supports the view that agents enjoy punishment. Moreover, players continue to cooperate when punishment is unobserved, perhaps because they (correctly) anticipate that shirkers will be punished: Fear of punishment can be as effective at promoting contributions as punishment itself.

1. Introduction

Costly (or altruistic) punishment occurs when one person pays a cost for another person to incur a cost. In a wide variety of settings, people are willing to use costly punishment against others who have defected. In particular, this is true in public-goods experiments both with anonymous random matching and with fixed groups that play a finite number of times.[1]

The mechanism underlying this costly punishment has been controversial,[2] and understanding when and why it occurs important both for the design of economic institutions[3] and for modeling the evolution of cooperation.[4] Our experimental evidence shows that subjects will engage in costly punishment even when it will not be observed until the end of the session, and moreover that there is as much cooperation in this treatment as in the standard settings where the allocated punishment is observed at the end of each period.

We argue that these findings help differentiate between a "pure preference" explanation of costly punishment and the "repeated-game" theory that cognitive limitations lead subjects to mistakenly treat the one-shot interaction as if it were repeated[5] and thus to punish to increase their own money payoff in future rounds. Our findings also reject simple forms of the "altruistic" theory that subjects punish to benefit the future counterparts of the punished players.[6] To make these points more precisely, note that any behavior at all can be explained by any of these theories without some constraints on how the underlying parameters are allowed to shift from one setting to the next. In particular, if preferences and cognitive errors are allowed to change depending on whether punishment is observed, then the two theories are observationally equivalent, because they are equally devoid of content. Our focus is thus on more constrained versions of these theories: We compare a stable preference for punishment with (1) the repeated-game theory that agents sometimes mistreat a one-shot game as a repeated interaction, but that this mistake is less likely when it is more obvious that the agents will not interact again and (2) the altruism theory that subjects understand they are in a one-shot interaction and punish to influence the future play of the punished player in a way that will help others though not themselves.

In our control, subjects participated in a standard public-goods experiment with anonymous random matching: Each period, subjects first decide how much to contribute to a public project, then observe the contributions of others, and have the option to engage in costly punishment. In our treatment, subjects were not informed of whether and how much they had been punished until after the conclusion of 10 rounds of play. The treatment eliminates some of the possible

* Corresponding author.
E-mail address: ppathak@mit.edu (P.A. Pathak).

[1] For anonymous random matching see e.g. the strangers treatment of Fehr and Gächter (2000), Fehr and Gächter (2002), Egas and Riedl (2008), Nikiforakis and Normann (2008), Anderson and Putterman (2006), and Rockenback and Milinski (2006). For fixed groups, see e.g. Yamagishi (1986), the partner treatment of Fehr and Gächter (2000), and the *P* treatment of Page et al. (2005).

[2] See e.g. Binmore and Samuelson (1999), Fehr et al. (2005), Fudenberg (2006), and Samuelson (2005).

[3] See e.g. Page et al. (2005) and Rockenback and Milinski (2006).

[4] Many authors have shown that evolution can lead to cooperation in repeated interactions, even without a preference for punishment, see for example Axelrod and Hamilton (1981), Fudenberg and Maskin (1990), Nowak and Sigmund (1992) and Imhof et al. (2005). It is less clear how to extrapolate those results to anonymous interactions.

[5] See Andreoni (1995), Ferraro and Vossler (2006) and Houser and Kurzban (2002) for investigations of the role of subject confusion in the public goods experiment without punishment.

[6] Of course the data is consistent with either the combination of either the repeated-game or altruism theory combined with the assumption that subjects somehow expect end-of-session observations to have an impact on play within the session.

0047-2727/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2009.10.007

instrumental effects that subjects might expect from punishment. With observed punishment, even in perfect-stranger treatments, subjects might hope that punishing now could lead to a contagious wave of punishing in the subject population that will benefit them later on, as in the theoretical literature on community enforcement. However, the actual probabilities of repeat interaction in our control and most past work are in fact too small for this to be a good strategy.[7] The delayed information about punishment thus requires that subjects who chose to punish have an even greater mismatch between rules and reality, so the instrumental and mismatch theories suggest that punishment should decrease. This was not the case: In general the treatment led to a small increase in both punishment and payoffs, and the difference is statistically significant when we pool data across periods.

Our findings are consistent with earlier findings that subjects cooperate and punish in the last round of the standard version of these experiments, as there too there is no instrumental rationale for punishment. However, punishment in those settings could be explained with a more limited version of the "preference" hypothesis, where the preference only arises from repeated observations of other agents being punished or from a "warm glow" from past cooperation; our design allows a cleaner separation of the various explanations.[8]

2. Experimental design

The experiment consists of two treatments: one with observed punishment — the standard voluntary contribution mechanism followed by a punishment stage designed to mimic the strangers treatment of designed to mimic the strangers treatment of Fehr and Gächter (2000, 2002) and one where the results of the punishment stage are not revealed to participants. Instructions for all treatments are written in neutral language and are based on Fehr and Gächter (2000).

Each treatment consists of two sets of 10 periods.[9] In each period, subjects played a public good contribution game in groups of 4 which consisted of two stages. In the first stage, each participant was endowed with 20 tokens, of which she could contribute an integer among between 0 and 20 to the public project. Each participant received the amount leftover after contribution plus 0.4 times the sum of the total group contribution to the public project. The stage game payoffs are such that, in a one-shot game, any contribution to the public good is a dominated strategy, but the aggregate benefit of contributing to the public good exceeds the individual benefit of a private investment.

In the second stage of each game, participants were informed of how much each of their three other group members contributed to the public project. Then they were allowed to allocate deduction points to individual group members. Each deduction point cost the participant 1 unit, and caused the recipient of the deduction point to lose 3 units, as in Fehr and Gächter (2002). As the ratio of effect per unit cost (here 3) becomes larger, punishment becomes more common; at low enough ratios the possibility of punishment is not sufficient to maintain cooperation.[10] The two stages together – the first stage public good contribution game followed by the second stage punishment phase – are hereafter denoted a period.

In the Observed treatment, at the end of each period participants were shown the total amount of punishment they received from the three other group members, together with the total amount of tokens lost through the allocation of deduction points. In this treatment, this was repeated 10 periods with random group composition and with feedback on the amount of punishment received after each period. In the Unobserved treatment, participants were not informed of whether they were punished until the conclusion of 10 periods. However, as in the control, if a participant allocated any deduction points to another participant, they were shown how much the allocation of each deduction point cost. At the 10th period, subjects were informed of the total amount of deduction points they received, and were notified of their total income from the previous 10 periods, but were not told in which periods they had been punished. This was then repeated for 10 more periods with random group compositions after each period. We will focus on comparisons between sets of 10 periods between Observed and Unobserved.

2.1. Procedures

A total of 132 subjects from Boston area universities participated voluntarily at the Harvard Business School CLER lab. The participants interacted anonymously via the software z-Tree (Fischbacher (2007)). Subjects were not allowed to participate in more than one session of the experiment. A total of 6 sessions were conducted, two in April 2007, one in October 2007, one in April 2008 and two in December 2008. Sessions either had 20 or 24 participants. In each session, the participants were paid a $10 show up fee, plus their earnings from the experiment. The average payment per participant was $23.33, and the sessions averaged approximately 1 h and 15 min.

The details of the sessions are described in Table 1. In Sessions C1 and C2 participants participate in the Observed treatment, while in Sessions T1–T4 subjects participate in the Unobserved treatment. In Observed, we have 3 sets of 10 periods of play, which yields a total of 640 individual observations. In Unobserved, we have 8 sets of 10 periods, which yields a total of 1760 individual observations.

In the beginning of each session, participants read the instructions before completing a control questionnaire. The experimenter then checked and, if necessary, explained the answers to all participants by reading a pre-written explanation. Once all participants' answers were checked, the experiment began. To control for an experimenter effect, all sessions were run by the same individual.

3. Hypotheses

As discussed in the Introduction, we consider three common explanations for costly punishment: The "preference" explanation that subjects act as if they have a stable preference for punishing "shirkers" independent of whether this has an impact on the shirkers' subsequent behavior, the "repeated-game" explanation that participants use

Table 1
Experimental design.

Session	ID	Number of participants	Order of treatments
1	C1	24	Observed
2	C2	20	Observed, Observed
3	T1	24	Unobserved, Unobserved
4	T2	24	Unobserved, Unobserved
5	T3	20	Unobserved, Unobserved
6	T4	20	Unobserved, Unobserved

Notes: All sessions employed a random matching protocol, where subjects were assigned to different groups after each period.
Session 1 had only one set of had 10 periods, while sessions 2–6 had one treatment followed by 10 periods of the second treatment.

[7] Duffy and Ochs (2009) show that subjects play non-cooperatively in a prisoner's dilemma with stochastic end date and anonymous random matching, while they play cooperatively in a control with fixed pairs. See Kandori (1992) and Ellison (1993) for calculations of the discount factor required for contagion to be an equilibrium in infinitely-repeated games with anonymous random matching. Note that the effect of making punishment unobserved is more complex in fixed-group treatments, as here players might worry that observed punishment could lead to retaliation, as in Dreber et al. (2008) and Nikiforakis (2008).

[8] Bochet et al. (2006) find that punishment is significantly higher in round 10 of their treatment *R* (which corresponds to our control except with a fixed-group design instead of re-matching) than in rounds 5–9. However, average contributions in round 10 are markedly lower than in rounds 5–9, and this fall-off could make punishment higher than it would have been with stable contributions. In contrast, we find that unobserved punishment leads to at least as much cooperation as when punishment is observed as it occurs.

[9] The first session we ran consisted of only one set of 10 periods.

[10] See e.g. Egas and Riedl (2008) and Nikiforakis and Normann (2008).

punishment to try to increase their own future payoff, as if they were in a repeated game, and the "altruism" explanation that costly punishment is altruistic is intended to benefit the future partners of the punished agent. The treatment eliminates the altruistic benefit, and makes the cognitive error required for the repeated-game theory even larger and thus less plausible. Thus, the repeated-game and instrumental explanations suggest the following hypothesis:

Hypothesis 1. Unobserved punishment will decrease the punishment of a defection, so punishment will be higher in Observed than Unobserved.

In addition, since punishment has been shown to increase cooperation in VCM games, if subjects expect Unobserved to have less punishment, we would also expect it to have less cooperation. This suggests a second hypothesis:

Hypothesis 2. If subjects believe that hypothesis 1 is correct, then contributions will be higher in Observed than Unobserved.

4. Experimental results

We begin by describing the time pattern of contributions to the public good averaging across the treatments and then report summary statistics by session averaging over periods. Sessions C2 and Sessions T1–T4 have two sets of 10 periods; we treat each of these as separate observations in most of our analysis. While participants did receive feedback after the first set of 10 periods in Sessions T1–T4, we show below that this did not have a significant effect.

Fig. 1 displays the evolution of average contributions over the 10 periods by pooling data for each Observed and Unobserved sessions. The figure shows that contributions in the Unobserved treatments begin at about 60% of subjects' endowments, while contributions in the Observed treatments begin at about 40% of subjects' endowments. The error bars in the figure are 95% confidence intervals assuming that the underlying distribution of observations in the sample is normally distributed. A comparison of the level of contributions in Unobserved in the first period to Observed in the first period indicates a statistically significant difference between contribution levels.

Across periods, the level of contributions in the Unobserved sessions is relatively stable with a mean of between 60 and 65% of the endowment, and a small downward trend in the last period. Likewise, the level of contributions in Observed is relatively stable with a mean of between 35 and 45% of the endowment, and a small downward trend in the last period.

Fig. 2 reports the period-by-period evolution of the fraction of participants who punish at least one other subject. The difference in this fraction between Unobserved and Observed was largest in the first period (35% vs 20%) but the confidence intervals are wide and overlapping, suggesting that the difference is not statistically significant

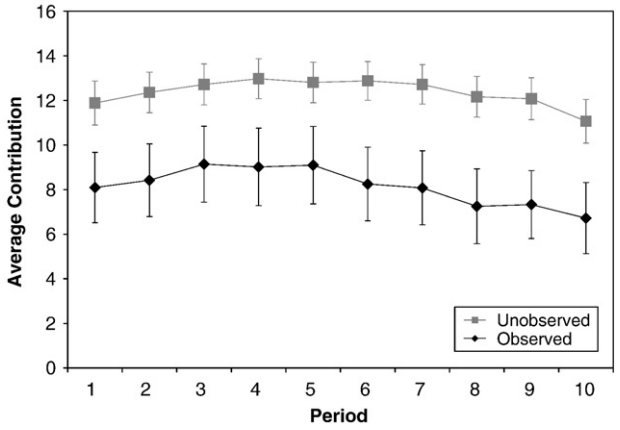


Fig. 1. Evolution of average contribution with 95% confidence intervals.

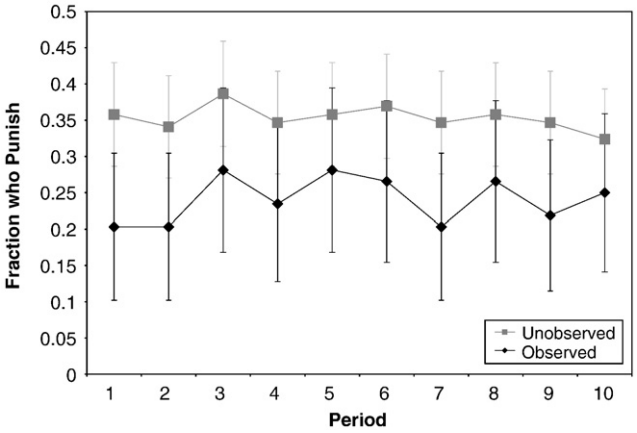


Fig. 2. Evolution of fraction who punish with 95% confidence intervals.

when the comparison is made period by period.[11] Across periods, the average fraction of subjects who punish in Unobserved is larger than that in Observed. Fig. 3 reports the period-by-period expenditures on punishment. The fraction of subjects who punish was somewhat higher in Unobserved, but the total amount of punishment received by an average participant was very similar in both treatments.

The finding that subjects contribute and punish in the Observed treatments is expected even though this is a finitely repeated game whose unique subgame perfect equilibrium is for all agents to never punish and never contribute. Cooperation is also observed in infinitely-repeated games (see, e.g., Dal Bo (2005)) but in that setting it is less of a puzzle. However, it is a surprise that contribution levels and the fraction of participants who punish are no lower in the Unobserved treatment.

Fig. 4 shows the average income of participants, where income in Unobserved is computed after stage 2. This income is not revealed to participants in the experiment until the conclusion of the 10th period. In the figure, the mean income in Unobserved is larger than that of Observed for each period, though the confidence intervals each period overlap. Income with unobserved punishment brackets but tends higher than the average income for the control, which is consistent with past findings on the ambiguous effect of punishment on average income.[12]

Table 2 reports some summary statistics across the sessions averaging across periods. The columns of the table correspond to the sessions and the set of games considered. In the table, we see that the average contribution in Session C1 and in both sets of ten periods in Session C2 is smaller than the averages for Sessions T1–T3. The contributions in session T4 are within the range of those in C1 and C2. Given the pattern that average contributions are higher in Fig. 1, this should come as no surprise. It is worthwhile pointing out that average contribution in sessions C2 and Session T4 are lower than in the other sessions. However, both C2 and T4 were run on the same day, and contributions in these sessions were roughly similar to each other.

We also see that the average across periods of the fraction of subjects who punish in a period is smaller in all of the Observed sessions than in 5 out of the 8 observations of Unobserved, a pattern mirroring that of Fig. 2. Table 2 also reports the fraction of participants who never punish, and who punish in at least two periods. Within each session, there is considerable heterogeneity in subjects' behavior in both treatments. There is overlap in the fraction who never punish between Observed and Unobserved, and likewise in the fraction who punish at least two periods. Punishment is also usually directed

[11] We verify this formally in Table 3.

[12] Egas and Riedl (2008) report that our 1:3 punishment technology leads to lower net payoffs than in the no punishment control, while Nikiforakis and Normann (2008) finds that this punishment technology leads to higher net payoffs. Similarly, Dreber et al. (2008) find that adding a punishment action to an infinitely-repeated prisoners dilemma leads to more cooperation but no increase in payoffs.

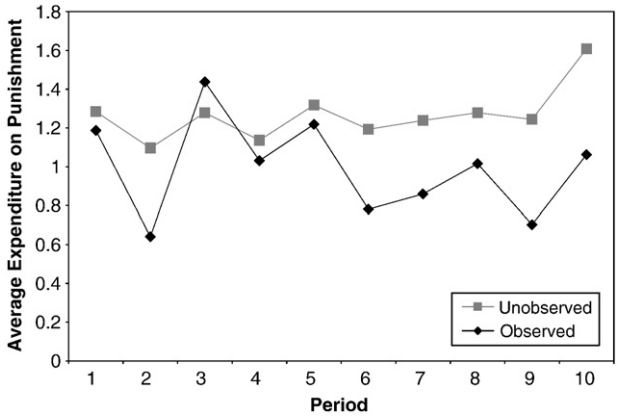


Fig. 3. Evolution of average expenditure on punishment.

towards a group member who contributed less than the average of the other three group members in both the Observed sessions and the Unobserved sessions. Across Observed and Unobserved sessions, the likelihood that the participant who punishes is a group member who contributed more than the average of the other group members is higher in the Unobserved sessions in 6 out of the 8 comparisons. Finally, Fig. 4 demonstrated that income in the Unobserved sessions was slightly higher on average than the Observed sessions, and this is borne out in 6 out of the 8 observations. The findings in the Observed sessions are similar to Fehr and Gächter (2002) who demonstrate that an observed punishment technology facilitates cooperation and that most but not all punishment is directed at low contributors; Herrmann et al. (2008) confirm that most punishment is directed at low contributors in Boston, Melbourne, and a number of Western European cities (but not in other parts of the world.).

We have focused on comparisons of the two treatments pooling treatments across periods, and pooling periods across treatments. In Table 3, we formally test for differences between treatments in a regression model. The model includes interactions for the period and whether the game is in the first set of 10 periods or second set of 10 periods. The first four columns report specifications where we do not allow the difference between the two sessions to depend on the period. The coefficient in column (1) of Table 3 implies that the average contribution levels in the Observed treatments are a statistically significant 4.27 lower than in the Unobserved treatment, which is consistent with the difference between the averages in Fig. 1. The next three columns indicate that the fraction who punish is about 12% lower in Observed, the expenditure on punishment is lower by −0.34, and overall income is lower by −1.16; each of these effects is significant at the 1% level.

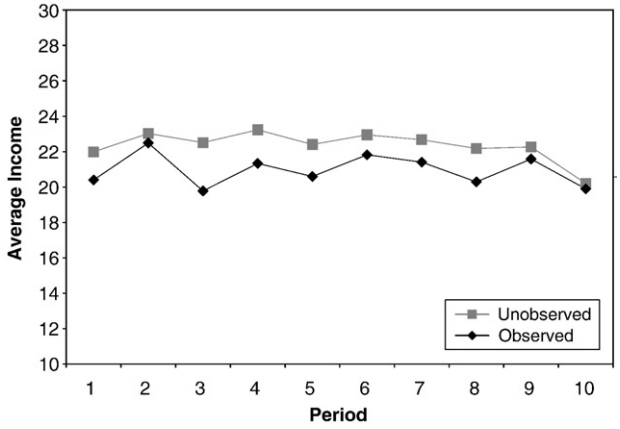


Fig. 4. Evolution of average income.

Columns (5)–(8) allow the difference between the two treatments to differ by period; here the reported estimate is the difference in the dependent variable while *p*-values are for the Wilcoxon two-sample test of the hypothesis that the two samples are from the same distribution are reported in brackets. Column (5) shows that contributions are higher in each period, and that the differences are all significant at the 1% level; contributions are between 3.5 and 5.0 higher in Unobserved than Observed. Many of the coefficients in column (6) are also significant, suggesting that the overall significant difference in the fraction who punish in column (2) does not extend to a stronger result that there is a significance difference in each period. Columns (7) and (8) show that the effect of the treatment on punishment expenditures and income is not statistically significant in a period-by-period comparison, even though it is significant if the periods are pooled.[13]

Result 1. Contribution levels are higher when punishment is unobserved for each period.

Result 2. Averaging across periods, participants are also more likely to punish, spend more on punishment, and earn less when punishment is observed.

4.1. Contribution and punishment

In Table 2 we saw that the vast majority of punishment was directed towards subjects who contribute less than the average of the other three group members. Table 4 relates whether the participant was punished to the difference between the participant's contribution and the average contribution of the other three group members. The models in columns (1)–(4) are linear probability models, column (5) reports the estimated average marginal effects from a logit. Column (6) reports a specification where the deviation from the group average consists of two terms: a term for an undercontribution and a term for contributing more than the group average, denoted overcontribution. The undercontribution is always a negative number, while the overcontribution is a positive number. The specifications include a different set of controls with indicators for subjects and periods as indicated.

In the Observed sessions for the specification in column (1), the coefficient on the difference is −0.025, while the coefficient in the Unobserved sessions is −0.045, and a statistically significant difference. This means that for each token less a participant contributes to the public good relative to the three other group members, that subject is 2.5% more likely to be punished in the Observed treatment, while in the Unobserved treatment this participant is 4.5% more likely to be punished.

The difference is sensitivity between Observed and Unobserved remains present across the different sets of controls in columns (1)–(5), and is also present when we allow for separate terms for under- and over-contribution. The estimates in column (6) show that when subjects contribute more than the group average, they are less likely to be punished.

To investigate the possibility of session effects, we also estimate models broken down by session. The results broken down by session display a similar pattern: in most specifications, the coefficient on the deviation in the Sessions T1–T4 has a larger magnitude than in Sessions C1 and C2. The main exception is the model with subject controls and the model with subject and period controls, where the coefficient in Session C1 is bracketed by the coefficients in the Session T1–T4. This suggests that we can conclude that there is more sensitivity to contributing less than the group mean comparing sessions pooled together.

These parameter estimates suggest whether a subject is punished is qualitatively similar across Observed and Unobserved, while the extent of the dependence of punishment on the deviation from the group is at

[13] We have also estimated tobit models for contribution, punishment expenditure, and income and a logit model for fraction who punish. Since the estimates are very similar to those from the OLS estimates in Table 3, we do not report them.

Table 2
Experimental design.

	Observed					Unobserved					
Session	C1	C2		T1		T2		T3		T4	
Period set	First (1)	First (2)	Second (3)	First (4)	Second (5)	First (6)	Second (7)	First (8)	Second (9)	First (10)	Second (11)
Average contribution	10.07	8.45	5.55	14.35	13.05	10.58	13.03	14.66	15.35	7.97	9.42
Average fraction who punish in a period	24%	33%	16%	38%	36%	45%	35%	43%	29%	29%	26%
Fraction who											
Never punish	46%	20%	30%	33%	36%	29%	38%	15%	30%	50%	50%
Punish at least two periods	50%	80%	40%	63%	50%	67%	58%	80%	65%	45%	40%
How often is punishment directed at someone who contributed less than average of other group members?	92%	78%	83%	92%	98%	93%	85%	88%	94%	95%	95%
How often is punisher someone who contributed more than average of other group members?	75%	67%	74%	85%	92%	73%	85%	88%	93%	68%	83%
Average income	21.53	20.00	21.25	24.02	23.11	18.45	22.08	23.60	25.14	20.58	22.13

least as large in the Observed as Unobserved. This helps explain why punishment need not be observable to support cooperation.

Result 3. When the sessions of each treatment are pooled, whether a participant receives punishment is more sensitive to whether she contributed less than the average of the other three group members when punishment is unobserved.

In Table 5, we report estimates where the dependent variable is the total amount of punishment received by a participant; the table shows that punishment expenditure is more sensitive to a participant's deviation from the group average in the treatment. The estimate in column (4), which includes controls for both period and subject, suggests that if a participant contributes 10 tokens less than the group average, she receives about 1 more punishment point (which deducts her income by 3 tokens) in Observed, while in Unobserved, she receives 2.2 more punishment points (which deducts her income by 6.6 tokens). This pattern persists across each of our controls. Moreover, the estimated coefficients for Sessions T1–T4 are larger in magnitude than both of the coefficients for Session C1 and C2 across each set of controls.

Result 4. When punishment is unobserved, the amount of punishment a participant receives is more sensitive to whether she gave less than the average of the other three group members.

Note that the way the punishment depends on individual and group contribution support the hypothesis that the desire to punish is intrinsic, or is a characteristic of preference, rather than an attempt to induce other subjects to contribute more.

4.2. The role of feedback

After the first ten games of the Unobserved treatment, participants were given feedback on the total amount of punishment they received before they play the next set of ten games. Fig. 5 reports the average contribution (on the left hand *y*-axis) and the fraction of subjects who punish (on the right hand *y*-axis) over the periods of the game before (first set of ten periods) and after feedback (second set of ten periods). The average contribution before feedback for the first period is lower than the average contribution in the first period immediately after feedback. However, the evolution of average contribution after feedback

Table 3
Comparisons of contributions, fraction who punish, expenditure on punishment, and income across treatments.

Dependent variable	Contribution (1)	Fraction who punish (2)	Expenditure on punishment (3)	Income (4)	Contribution (5)	Fraction who punish (6)	Expenditure on punishment (7)	Income (8)
Observed	−4.271***	−0.128***	−0.338***	−1.158***				
	(0.294)	(0.022)	(0.108)	(0.333)				
Observed, differences by period								
Period 1					−3.694	−0.163	−0.148	−1.348
					[$p<0.001$]	[$p=0.023$]	[$p=0.794$]	[$p=0.137$]
Period 2					−3.829	−0.148	−0.512	−0.249
					[$p<0.001$]	[$p=0.041$]	[$p=0.536$]	[$p=0.346$]
Period 3					−3.500	−0.110	0.139	−2.598
					[$p<0.001$]	[$p=0.134$]	[$p=0.358$]	[$p=0.002$]
Period 4					−3.939	−0.126	−0.182	−1.587
					[$p<0.001$]	[$p=0.099$]	[$p=0.890$]	[$p=0.013$]
Period 5					−3.822	−0.100	−0.123	−1.778
					[$p<0.001$]	[$p=0.268$]	[$p=0.782$]	[$p=0.012$]
Period 6					−4.761	−0.123	−0.518	−0.785
					[$p<0.001$]	[$p=0.135$]	[$p=0.124$]	[$p=0.108$]
Period 7					−4.779	−0.171	−0.471*	−0.984
					[$p<0.001$]	[$p=0.034$]	[$p=0.075$]	[$p=0.148$]
Period 8					−5.000	−0.118	−0.356	−1.569
					[$p<0.001$]	[$p=0.181$]	[$p=0.599$]	[$p=0.053$]
Period 9					−4.938	−0.150	−0.631	−0.440
					[$p<0.001$]	[$p=0.060$]	[$p=0.248$]	[$p=0.377$]
Period 10					−4.444	−0.075	−0.577	−0.244
					[$p<0.001$]	[$p=0.273$]	[$p=0.633$]	[$p=0.465$]

Notes: Columns 1–4 report coefficient on indicator for Observed with period*set interactions where set is an indicator of the first set of 10 periods.
Columns 5–8 report mean of regression coefficient of indicator for Observed interacted with the period with *p*-value of Wilcoxon two-sample test of the hypothesis that the two samples are equal in brackets. Number of observations is equal to 2400 in specifications (1)–(4), and 240 for each paired comparison in columns (5)–(8).
*Significant at 10%;***significant at 1%.

Table 4
Estimates of punishment on deviation from group mean.

	LPM estimates					LPM estimates		
	Deviation	Deviation	Deviation	Deviation	Logit deviation	Undercontribution	Overcontribution	*N*
	(1)	(2)	(3)	(4)	(5)	(6)		(7)
Observed (C1–C2)	−0.025***	−0.028***	−0.025***	−0.028***	−0.027***	−0.037***	−0.018**	640
	(0.002)	(0.003)	(0.002)	(0.003)	(0.011)	(0.005)	(0.005)	
Unobserved (T1–T4)	−0.045***	−0.041***	−0.045***	−0.041***	−0.040***	−0.061***	−0.020***	1760
	(0.001)	(0.002)	(0.001)	(0.002)	(0.027)	(0.003)	(0.003)	
By session								
Session C1	−0.037***	−0.041***	−0.037***	−0.041***	−0.037***	−0.056***	−0.022***	240
	(0.003)	(0.005)	(0.003)	(0.005)	(0.024)	(0.008)	(0.009)	
Session C2	−0.017***	−0.020***	−0.018***	−0.020***	−0.020***	−0.023***	−0.018**	400
	(0.003)	(0.004)	(0.003)	(0.004)	(0.006)	(0.007)	(0.007)	
Session T1	−0.045***	−0.042***	−0.045***	−0.042***	−0.037***	−0.058***	−0.025***	480
	(0.002)	(0.003)	(0.002)	(0.003)	(0.038)	(0.005)	(0.005)	
Session T2	−0.044***	−0.033***	−0.044***	−0.033***	−0.034***	−0.050***	−0.018**	480
	(0.003)	(0.005)	(0.003)	(0.005)	(0.014)	(0.008)	(0.008)	
Session T3	−0.048***	−0.043***	−0.048***	−0.043***	−0.038***	−0.072***	−0.009	400
	(0.026)	(0.004)	(0.003)	(0.004)	(0.026)	(0.006)	(0.007)	
Session T4	−0.042***	−0.043***	−0.042***	−0.043***	−0.044***	−0.063***	−0.024***	400
	(0.003)	(0.004)	(0.003)	(0.004)	(0.031)	(0.007)	(0.007)	
Subject controls	N	Y	N	Y	Y	Y		
Period controls	N	N	Y	Y	N	Y		

*Dependent variable is an indicator variable if the subject is punished by another group member, while independent variable is the amount that the subject contributes less than the average of the other three group members (undercontribution). Columns labelled LPM report linear probability model estimates of coefficient on undercontribution. All specifications include indicators for the first 10 periods. Subject and period controls are fixed effects for subjects and periods, respectively. Logit specifications are average marginal effects. Specification (6) includes a separate independent variable for when undercontribution is positive, "overcontribution", and when it is negative, "undercontribution." Column (7) reports number of observations.
** Significant at 5%; ***significant at 1%.

closely tracks the evolution before feedback, suggesting that the average contribution pattern does not change in response to feedback. Likewise, the fraction of participants who punish in any period before feedback is slightly higher than the fraction who punish after feedback, but the evolution of these two fractions closely track one another. By the last period, the fraction who punish either before or after feedback is virtually the same. Indeed, the fact that nearly a third of participants punish even in the last period of the experiment, a fact which has been documented by others, is further evidence in favor of the view that agents have an intrinsic punishment for punishment.

Taken together, these patterns suggest that feedback after the first 10 periods did not significantly influence the contribution levels or willingness to punish in the Unobserved treatment.

5. Conclusion

Our data shows that players spend resources on punishment even when it will not be observed until the end of ten periods. Recent studies (e.g., Andreoni et al. (2003)) have shown that costly rewards can also

Table 5
Estimates of punishment expenditure as a function of deviation from group average.

	OLS estimates				OLS estimates		
	Deviation	Deviation	Deviation	Deviation	Undercontribution	Overcontribution	*N*
	(1)	(2)	(3)	(4)	(5)		(6)
Observed (C1–C2)	−0.092***	−0.010***	−0.092***	−0.101***	−0.169***	−0.025	640
	(0.010)	(0.013)	(0.010)	(0.011)	(0.022)	(0.024)	
Unobserved (T1–T4)	−0.228***	−0.220***	−0.228***	−0.220***	−0.429***	−0.003	1760
	(0.007)	(0.009)	(0.007)	(0.009)	(0.015)	(0.015)	
By session							
Session C1	−0.135***	−0.162***	−0.135***	−0.162***	−0.248***	−0.056	240
	(0.016)	(0.021)	(0.016)	(0.021)	(0.036)	(0.042)	
Session C2	−0.065***	−0.064***	−0.065***	−0.064***	−0.116***	−0.011	400
	(0.012)	(0.016)	(0.012)	(0.016)	(0.029)	(0.029)	
Session T1	−0.211***	−0.200***	−0.211***	−0.200***	−0.372***	−0.016	480
	(0.010)	(0.013)	(0.010)	(0.013)	(0.020)	(0.021)	
Session T2	−0.306***	−0.285***	−0.306***	−0.285***	−0.599***	−0.003	480
	(0.016)	(0.025)	(0.016)	(0.025)	(0.037)	(0.035)	
Session T3	−0.255***	−0.243***	−0.255***	−0.243***	−0.465***	0.019	400
	(0.015)	(0.020)	(0.015)	(0.020)	(0.031)	(0.035)	
Session T4	−0.153***	−0.178***	−0.153***	−0.178***	−0.316***	−0.041	400
	(0.013)	(0.018)	(0.013)	(0.018)	(0.032)	(0.032)	
Subject controls	N	Y	N	Y	Y		
Period controls	N	N	Y	Y	Y		

*Dependent variable is the amount of punishment received by a subject from another group member, while independent variable is the amount that the subject contributes less than the average of the other three group members (undercontribution). Columns labelled OLS report least squares estimates of coefficient on undercontribution. All specifications include indicators for the first 10 periods. Subject and period controls are fixed effects for subjects and periods, respectively. Column (5) includes a separate independent variable for when undercontribution is positive, "overcontribution", and when it is negative, "undercontribution." Column (6) reports the number of observations.
*** Significant at 1%.

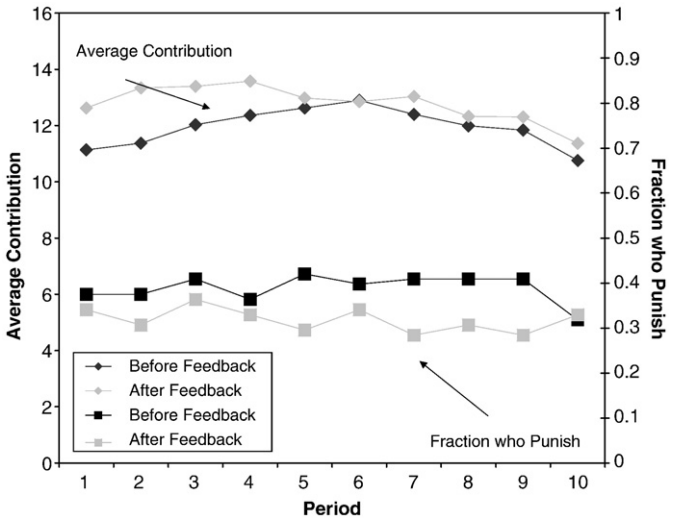


Fig. 5. Evolution of contribution and fraction punishing before and after feedback.

support cooperation. We predict that as with costly punishments, these rewards will still be used even when they are not observed.

Our results are consistent with the view that agents enjoy punishment, where 'enjoyment' includes anger and a desire for retribution, and poses severe problems for strategic interpretations of costly punishment. Moreover, players continue to cooperate when punishment is unobserved, perhaps because they (correctly) anticipate that shirkers will be punished: Fear of punishment can be as effective at promoting contributions as punishment itself.[14] At this point we do not have an explanation for why punishment tends to if anything increase when it is unobserved; the answer may become more clear if this effect is found in other contexts.

Acknowledgments

We thank Enst Fehr and Simon Gächter for sharing their data and z-Tree code with us, Pedro Dal Bo, Anna Dreber Almenberg, Ernst Fehr, Nikos Nikiforakis, Martin Nowak, and David Rand for helpful discussions, and NSF grant SES 0646816 for financial support.

Appendix A. Instructions

You are now taking part in an economic experiment. If you read the following instructions carefully, you can, depending on your decisions, earn a considerable amount of money. Please read these instructions carefully.

We have distributed the same instructions to all participants of this experiment. These instructions are solely for your private information. *You are prohibited from communicating with other participants during the experiment.* If you have any questions, please raise your hand, and we will come to you immediately. If you violate this rule, we will have to exclude you from the experiment and all payments.

During the experiment, we will not speak in terms of US Dollars, but instead in tokens. Your entire earnings from the experiment will be calculated in tokens. At the end of the experiment, the total amount of tokens you have earned will be converted to dollars at the following rate:

$$1\,token = 0.03\,dollars.$$

The experiment is divided into different periods. In each period, you will be divided into groups of four. Therefore, there will be three other members in your group. The composition of your group members will change at random after each period. Each other player in the experiment is equally likely to be in your group in the next period, i.e. If there are 32 players, the probability that any specific player is in your group the next period is $1/32 = 0.03$.

In each period, the experiment consists of *two stages*. In the first stage, you have to decide how many tokens you would like to contribute to a project. In the second stage, you are informed of the contributions of the three other group members to the project. You then decide whether or how much to reduce their earnings in the first stage by distributing deduction points to them. The following pages describe the experiment in detail.

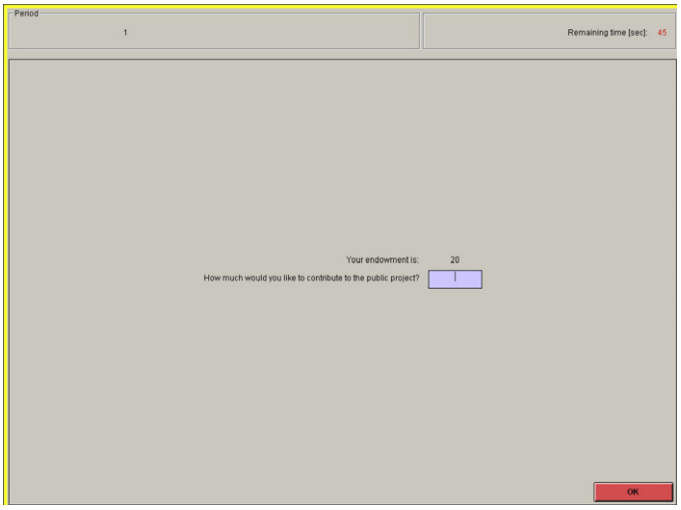
Appendix B. Detailed information on the experiment

B.1. The first stage

At the beginning of each period, each participant receives 20 tokens. We will call this his or her endowment. Your task is to decide how much of your endowment you wish to contribute to the project and how much you wish to keep for yourself. The consequences of your decision are explained in detail below.

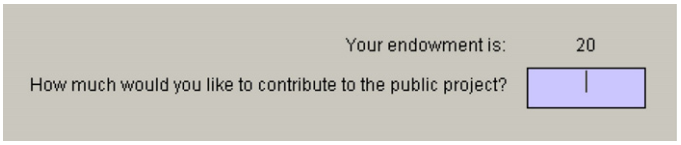
At the beginning of each period, the following input screen for the first stage will appear:

First stage input screen



The number of the period appears in the top left corner of the screen. In the top right corner of the screen, you can see how many more seconds remain for you to decide on the distribution of your endowment. Your decision must be made before the time displayed is 0 s.

First stage input screen (enlarged)



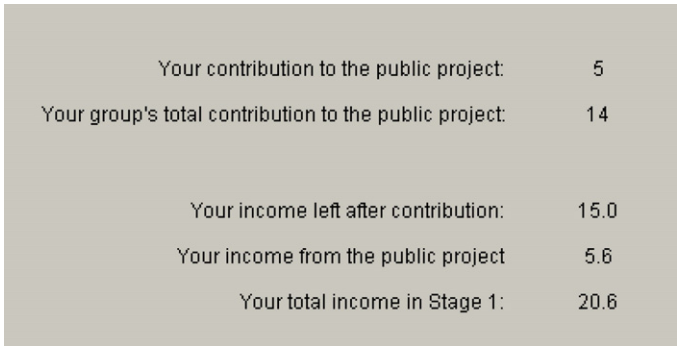
Your endowment in each period is 20 tokens. You have to decide how many tokens you want to contribute to the project by typing a number between 0 and 20 in the input field. The field can be reached by clicking with the mouse. As soon as you have decided how many tokens to contribute to the project, you have also decided how many points to keep for yourself. This is what is left of your endowment: (20 − your

[14] This same effect was observed in a fixed-partners treatment by Vyrastekova et al. (2008).

contribution) tokens. After entering your contribution, you must press the OK button (either with the mouse, or by pressing the Enter key) in the bottom left corner of the screen. Once you have done this your decision can no longer be revised.

After all members of your group have made their decision, the following income screen will show you the total amount of points contributed by all four group members to the project (including your contribution). This screen also shows you how many tokens you have earned at the first stage.

First stage income screen



Once you have read this screen, you will be asked to click OK in the bottom right hand corner.

Your income consists of two parts:

1) the tokens which you have kept for yourself ("Your income left after contribution"):
2) the "income from the public project". This profit is calculated as follows:

Your income from the project = 0.4*the total contribution of all 4 group members to the project.

Your income in tokens at the first stage of a period is therefore:

(20 − *your contribution to the project*)
+ 0.4*(*total contributions to the project*)

The income of each group member from the project is calculated in the same way. This means that each group member receives the same income from the project.

Example: Suppose the sum of the contributions of all group members is 60 tokens. In this case, each member of the group receives an income from the project of: 0.4*60 = 24 tokens. If the total contribution to the project is 9 points, then each member of the group receives an income of 0.4*9 = 3.6 tokens from the project.

For each point that you keep for yourself you earn an income of 1 token. Suppose you gave this token to the project instead. Then the total contribution to the project would rise by one unit. Your income from the project would rise by 0.4*1 = 0.4 points, so your total income would decrease by 1–.4 = .6. The total income of the other group members would also rise by 0.4 each, so that the total income of the group from the project would rise by 1.6. Your contribution to the project therefore raises the income of the other group members. Likewise, you earn 0.4 for each point contributed by the other members to the project.

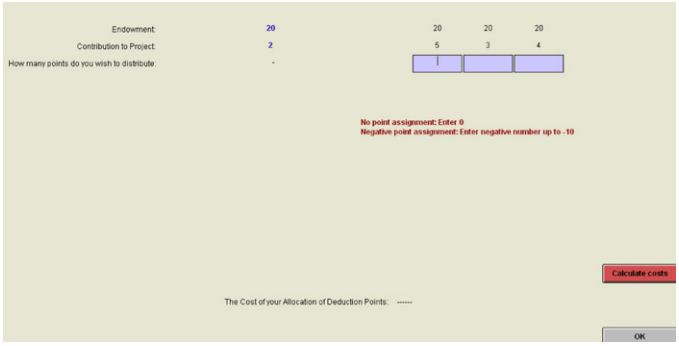
Once you have reviewed your income and the group's total contributions from the first stage, please click OK. The first stage is then over and the second stage commences.

B.2. The second stage

In the second stage, you will see how much each of the other group members contributed to the project. At this stage, you can also *reduce or leave equal* the income of *each* group member by *distributing deduction points*. The other group members can also reduce your income if they wish to.

The input screen in the second stage is:

Second stage input screen



On this screen, you see how much the three other group members contributed to the project in the first stage. Your contribution is displayed in blue in the first column, while the contributions of the other group member are shown in the remaining three columns. In each column, you will see the endowment and the contribution of each of the group members.

For instance, in the screen above, you contributed 2 out of your endowment of 20 tokens. Another member of your group contributed 5 out of 20 tokens, the group member displayed in the next column contributed 3 out of 20 tokens, while the group member in the last column contributed 4 out of 20 tokens.

Appendix C. Allocating deduction points

You must now decide how many deduction points you wish to give each of the other three group members. You must enter a number for each of them. If you do not wish to change the income of a specific group member, then enter 0. You can allocate between −10 and 0 deduction points to each group member. Each deduction point that you allocate to a group member will reduce their income from the first stage by three times the number of deduction points you allocate. For instance, if you assign −3 points to group player in the second column, then his or her income from the first stage will be reduced by 9.

Each deduction point you allocate comes at a cost to you. The more deduction points you allocate the higher your cost. Your total costs are equal to the number of deduction points you allocate.

Cost of deduction points
= *Total deduction points allocated to the other players*

For instance, suppose you give 2 deduction points to one member of the group, this costs you 2 tokens; if you give 9 deduction points to another member, this costs you a 9 tokens; and if you give the last group member 0 deduction points, this has no cost for you. In this case, your total cost of distributing deduction tokens would be 2 + 9 = 11 tokens. Your total costs of distributing deduction points are displayed on the input screen. If you click on the "Calculate costs" button, then you will see the total costs at the bottom of the screen. Once you have decided your deduction points, and calculated its costs, click OK. You must click "calculate costs" before clicking OK. As long as you have not clicked OK, you can revise your decision.

If you choose 0 points for a particular group member, you do not change his or her income. However if you give a member 1 deduction point (by choosing 1), you reduce his or her income from the first stage by 3 tokens. If you give a group member 2 points (by choosing 2) you reduce his or her first stage income by 6 tokens, etc. The amount

of points you distribute to each member determines therefore how much you reduce their income from the first stage.

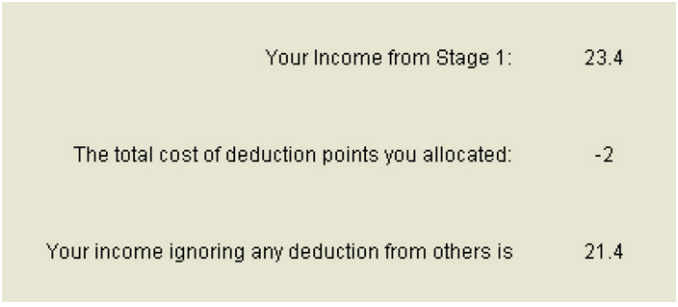
Whether or by how much the income from the first stage is reduced depends on the total of the received deduction points. If somebody received a total of 3 points (from all of the other group members in this period), his or her income would be reduced by 9 tokens. If someone receives a total of 4 points, his or her income would be reduced by 12 tokens.

Income reduction = 3**Total deduction points allocated to you*

Total income (in tokens) at the end of the 2nd stage = *period income*
= (*income from the 1st stage–income reduction*)
−*costs of your deduction points*

After all participants have made their decision, you will be notified of your income from the first stage, the cost of the deduction points you allocated, and your income minus the cost of deduction points. You will not be informed of the number of deduction points that you received. Your income from the period will be displayed on the following screen:

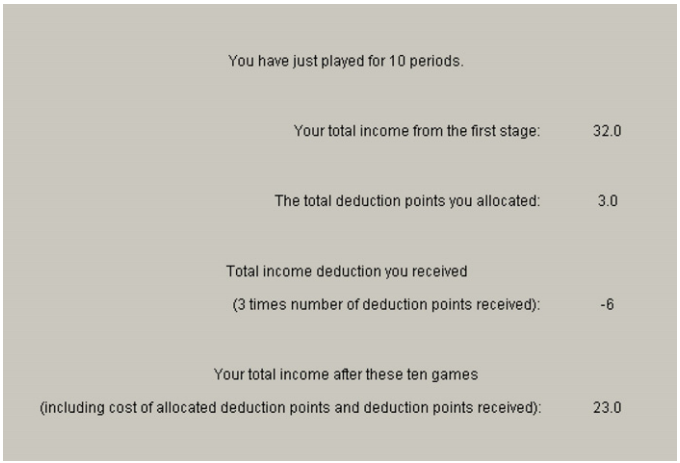
Second stage income screen



The calculation of your income from the first period, the costs of your distribution of points and your income in the period ignoring any deduction from others are as explained above.

After the tenth game, you will be notified of your total income. This is the sum of your first stage income in each round, minus the costs you incurred by allocating deduction points to other players, minus your income deduction due to the deduction points that other players gave to you. However, your total income cannot be negative; if your costs exceed your income, your payoff will be 0.

This information will be displayed as follows:



Note that while you will be informed of the total deduction points that other players gave you, you will not be told which periods these deductions occurred in or which player gave them to you.

Please sit quietly until the rest of the participants have read the instructions. At this point, we will review the instructions and then if there any questions, please raise your hand.

References

Anderson, C., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games Economic Behavior 54, 1–24.
Andreoni, J., 1995. Cooperation in public-goods experiments: kindness or confusion. American Economic Review 85 (4), 891–904.
Andreoni, J., Harbaugh, W., Vesterlund, L., 2003. The carrot or the stick: rewards, punishments, and cooperation. American Economic Review 93, 893–902.
Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.
Binmore, K., Samuelson, L., 1999. Evolutionary drift and equilibrium selection. Review. Economic Studies 66, 363–394.
Bochet, O., Page, T., Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior and Organization 60, 11–26.
Dal Bo, P., 2005. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. American Economic Review 95, 1591–1604.
Dreber, A., Rand, D., Fudenberg, D., Nowak, M., 2008. Winners don't punish. Nature 452, 348–351.
Duffy, J., Ochs, J., 2009. Cooperative behavior and the frequency of social interaction. Games and Economic Behavior 66, 785–812.
Egas, M., Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. Proceedings of the Royal Society: Biological Sciences 275 (1637).
Ellison, G., 1993. Cooperation in the prisoner's dilemma with anonymous random matching. Review of Economic Studies 61, 567–588.
Fehr, E., Gächter, S., 2000. Cooperation and punishment in public good experiments. American Economic Review 90, 980–994.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.
Fehr, E., Fischbacher, U., Kosfeld, M., 2005. Neuroeconomic foundations of trust and social preferences: initial evidence. American Economic Review 96, 1611–1630.
Ferraro, P. and C. Vossler, 2006. The Source and Structure of Confusion in Public Goods Experiments. Unpublished mimeo, University of Tennessee-Knoxville
Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.
Fudenberg, D., 2006. Advancing beyond advances in behavioral economics. Journal of Economic Literature 44, 94–711.
Fudenberg, D., Maskin, E., 1990. Evolution and cooperation in noisy repeated games. American Economic Review 80, 274–279.
Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.
Houser, D., Kurzban, R., 2002. Revisiting kindness and confusion in public goods experiments. American Economic Review 92 (4), 1062–1069.
Imhof, L., Fudenberg, D., Nowak, M., 2005. Evolutionary cycles of cooperation and defection. Proceedings of the National Academy of Sciences 102, 10797–10800.
Kandori, M., 1992. Social norms and community enforcement. Review of Economic Studies 59, 63–80.
Nikiforakis, N., 2008. Punishment and counter-punishment in public good games: can we really govern ourselves? Journal of Public Economics 92, 91–112.
Nikiforakis, N., Normann, H.T., 2008. A comparative statics analysis of punishment in public good experiments. Experimental Economics 11, 358–369.
Nowak, M.A., Sigmund, K., 1992. Tit for tat in heterogeneous populations. Nature 355, 250–253.
Page, T., Putterman, L., Unel, B., 2005. Voluntary association in public goods experiments: reciprocity, mimicry, and efficiency. Economic Journal 115, 1032–1053.
Rockenback, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444, 718–723.
Samuelson, L., 2005. Economic theory and experimental economics. Journal of Economic Literature 43, 65–107.
Vyrastekova, J., Funaki, Y., and A. Takeuchi (2008) Strategic vs. Non-Strategic Motivations of Sanctioning. mimeo.
Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51, 110–116.