American Economic Review 2012, 102(7): 3317–3332
http://dx.doi.org/10.1257/aer.102.7.3317

Imperfect Public Monitoring with Costly Punishment: An Experimental Study†

By ATTILA AMBRUS AND BEN GREINER*

This paper experimentally investigates the effects of a costly punishment option on cooperation and social welfare in long, finitely repeated public good contribution games. In a perfect monitoring environment, increasing the severity of the potential punishment monotonically increases average net payoffs. In a more realistic imperfect monitoring environment, we find a U-shaped relationship. Access to a standard punishment technology in this setting significantly decreases net payoffs, even in the long run. Access to a severe punishment technology leads to roughly the same payoffs as with no punishment option, as the benefits of increased cooperation offset the social costs of punishing. (*JEL* C92, H41, K42)

A large and growing experimental literature in economics, starting with Fehr and Gächter (2000), demonstrates that the possibility of costly punishment facilitates increased cooperation in finite-horizon social dilemma situations such as prisoner's dilemma and public good contribution games.[1] A recent paper by Gächter, Renner, and Sefton (2008) shows that if the game horizon is long enough, the possibility of punishment also increases average net payoffs in the population.[2] That is, while in early rounds of the game (roughly the first 10 rounds of the 50-round game investigated) the welfare-improving effect of increased cooperation is more than counterbalanced by the welfare-reducing effect of relatively frequent use of the punishment option, in the rest of the game a high level of cooperation is maintained with little explicit use of the punishment option. This result is consistent with group selection models of cooperation and punishment (see, for example, Boyd et al. 2003, and Chapter 13 in Bowles 2004).

*Ambrus: Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham, NC 27708 (e-mail: aa231@duke.edu); Greiner: School of Economics, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: bgreiner@unsw.edu.au). We thank Scott French, Drew Fudenberg, Jeffrey Miron, Andreas Nicklisch, Ori Weisel, and three anonymous referees for helpful comments and suggestions. Financial support through an Australian School of Business Research Grant is gratefully acknowledged.

† To view additional materials, visit the article page at http://dx.doi.org/10.1257/aer.102.7.3317.

[1] For the original references in social sciences, see Yamagishi (1986); Ostrom, Walker, and Gardner (1992); and the theoretical contribution of Boyd and Richerson (1992). For empirical evidence for the relevance of costly punishment outside the lab, see Krueger and Mas (2004) and Mas (2008).

[2] An earlier string of papers (Fehr and Gächter 2002; Gürerk, Irlenbusch, and Rockenbach 2006; Dreber et al. 2008; Egas and Riedl 2008; Herrmann, Thöni, and Gächter 2008) shows that in repeated games with a shorter time horizon, the social costs of punishment tend to outweigh the benefits coming from increased cooperation. Rand et al. (2009) investigates the effect of access to punishment versus reward options in long (50-round) contribution games. For a theoretical investigation of the potential social costs and benefits of punishment, see Hwang and Bowles (2010).

In this paper we investigate how the option of costly punishment affects welfare in a more realistic environment, in which subjects observe each others' decisions with a small amount of noise. In particular, we investigate a public good contribution game in which, after each contribution decision, the public record of a player—that is, the information on the subject's contribution announced publicly to all players—might differ from the true contribution of the subject. Specifically, even if the subject contributed to the public good, with 10 percent probability the public record indicates no contribution. This design corresponds to partnership situations in which even if a member of the partnership contributes to a joint project, the others do not recognize the contribution, at least not until some later time. In our design, such mistakes in the public record only influence the subjects' information, but not their payoffs, which are determined by their true actions.[3]

Our design is mostly similar to that of Gächter, Renner, and Sefton (2008). In particular, we examine 50-round public good contribution games, and we adopt the same mapping between contributions and payoffs.[4] The only difference is that in our experiments subjects can only choose between contributing all or none of their endowments in each round. This was implemented in order to simplify the noise structure, with the intent that subjects understand better how their public records depend probabilistically on their decisions. Because of this change, we also ran a control design in which subjects observed each others' contributions perfectly. The other dimension in which we varied the design was the amount and effectiveness of costly punishment subjects could inflict on each other. We employed (i) a no punishment environment; (ii) a standard punishment technology used in Gächter, Renner, and Sefton (2008) and other experimental papers, in which a subject can inflict a damage of three tokens for every token spent on punishment, and there is an upper limit on the amount of damage that could be inflicted; and (iii) a strong punishment technology, in which a subject can inflict a damage of six tokens for every token spent on punishment, and there is no upper limit on the amount of punishment. Hence, our experiment facilitated investigating the effects of increasing the severity of punishment in both perfect and imperfect monitoring environments.

We found that in the benchmark perfect monitoring environment, increasing the severity of punishment increased both the amount of contributions and the average net payments (that is, payments net of the costs of imposed and received punishments) monotonically. This reinforces the findings in Nikiforakis and Normann (2008) and Egas and Riedl (2008), the first papers in the literature to investigate the effects of varying the severity of punishment.[5] In the presence of either of the punishment options subjects learned to cooperate. In the strong punishment design this

[3] The realized payoffs were revealed to subjects at the end of the experiment.

[4] As expressed in Gächter, Renner, and Sefton (2008), there is an assertion in the experimental literature that play in long finitely repeated games, aside from the last few rounds, is similar to play in indefinitely repeated games with a large continuation probability. We are not aware of a formal test of this claim. Our results are relevant for infinite-horizon situations to the extent that the above assertion is adopted. In the real world there are both situations that are well approximated by a finite-horizon model (if there is a highlighted point of time after which the probability of continued interaction is very small), and ones that are better approximated by an infinite-horizon model.

[5] Nikiforakis and Normann (2008) investigated punishment effectiveness ratios 1:1, 1:2, 1:3, and 1:4 in ten-times repeated public good contribution games with perfect monitoring. Egas and Riedl (2008) studied punishment effectiveness ratios 3:1, 1:1, and 1:3 in a perfect strangers design.

learning quickly led to almost full cooperation in the public good game and virtually no use of the punishment option after a few initial rounds.

In the imperfect monitoring environment, the observed patterns are very different. The possibility of using the standard punishment option, while increasing contributions by a modest amount, significantly decreased average net earnings. Contribution levels stayed far away from full cooperation, and subjects kept using the punishment option regularly throughout the whole game. In fact, average per round net earnings stabilized in the second half of the experiment, suggesting that the same qualitative conclusions would hold even over longer time horizons.

In contrast to standard punishment, the strong punishment option does increase average contributions significantly, even in the imperfect monitoring environment. The use of the punishment technology remains relatively frequent throughout the game, however. In our experiment, these contrasting effects on the payoffs cancel each other out, and average net earnings with the strong punishment option are about the same as with no punishment option.

To summarize, in a noisy environment, it is not clear whether the costly punishment option is beneficial for society, even in the long run. Moreover, we find a U-shaped relationship between the severity of possible punishment and social welfare: the possibility of an intermediate level of punishment significantly decreases social welfare relative to when no punishment is available, while the possibility of severe punishment has a roughly zero net benefit for society.

A closer look at the data provides hints for why costly punishment is less effective in a noisy environment in establishing cooperation. First, subjects who were punished "unfairly," in the sense that the punishment followed a contribution by the subject, were less likely to contribute in the next round.[6] Such unfair punishment happens more often in the imperfect monitoring environment, following a wrong public record. The above effect gets curtailed in the design with strong punishment, but at the cost that when punishment occurs (and it does occur from time to time) then it inflicts heavy damage. Second, in the case of regular punishment, the positive effect of punishing noncontributors on their subsequent contributions is reduced by noise. This suggests either that noncontributors do not believe that others will keep on punishing them for public records of not contributing in a noisy environment, or that they keep on not contributing because of the possibility that even if they contribute, they could receive a wrong public record and get punished anyway.

Our paper complements findings in a number of recent papers. Bereby-Meyer and Roth (2006) show that players' ability to learn to cooperate in a repeated prisoner's dilemma game is substantially diminished when payoffs are noisy, even though in their experiment players could monitor each other's past actions perfectly.[7] In contrast, we find that a small noise in monitoring, albeit decreasing contributions in all conditions, does so significantly only in the strong punishment treatment. Abbink and Sadrieh (2009) find that if contributions are observed perfectly but there is noise

[6] This is consistent with the findings of Hopfensitz and Reuben (2009) in that punishment facilitates future cooperation, but only when it evokes shame and guilt, not when it evokes anger. The paper uses information on players' emotions captured through a questionnaire during the experiment. Herrmann, Thöni, and Gächter (2008) also find that (antisocial) punishment of contributors lowers their subsequent contributions.

[7] See also Gong, Baron, and Kunreuther (2009) on repeated prisoner's dilemma games with stochastic payments, in a group versus individual decision-making context.

in observing punishment then subjects punish each other more, reducing overall efficiency. Bornstein and Weisel (2010) and Patel, Cartwright, and van Vugt (2010), using different designs, show that the benefits of costly punishment are diminished when there is uncertainty regarding the realized endowment of subjects (but contributions are perfectly observed). Most closely related to our investigation is Grechenig, Nicklisch, and Thöni (2010), who in a work independent from ours also point out that in a noisy environment punishment can reduce welfare. They do not investigate the effects of increasing the severity of punishment technology, which is the main focus of our paper, and instead examine the effects of varying the level of noise in observations. Furthermore, like all the above papers, Grechenig, Nicklisch, and Thöni (2010) focus on relatively short repeated games, in which the welfare benefits of costly punishment are ambiguous even without noise (see footnote 2).

We also contribute to the small but growing experimental literature on repeated games with imperfect public monitoring (Miller 1996; Aoyagi and Fréchette 2009; Fudenberg, Rand, and Dreber 2012) although these papers investigate issues largely unrelated to ours.[8]

I. Experimental Design

We implemented six treatments in a 3×2 factorial design. In the punishment dimension we used no punishment, regular punishment, and strong punishment options, and in the noise dimension we employed either no noise in the information about other group members' contributions, or a small amount of noise. In our baseline experimental design, the instructions and procedures follow closely those of Gächter, Renner, and Sefton (2008). Namely, experimental subjects participated in a 50-round repeated public good game. At the beginning, participants were randomly and anonymously matched to groups of 3 that stayed constant over all 50 rounds. In each round, each of the 3 participants in a group was endowed with 20 tokens and asked to either contribute all or none of these tokens to a group account.[9] If the endowment was kept, it benefitted the participant by 20 points, while if the endowment was contributed, it benefitted each of the 3 group members by $0.5 \times 20 = 10$ points.

After all group members made their choice simultaneously, they were informed about the outcome of the game. In the *no noise* treatments, participants were informed about the choices in their group, while in the *noise* treatments only a "public record" of each group member's choice was displayed. If a group member did not contribute, then the public record would always indicate "no contribution." If the group member contributed, there was a 10 percent chance that the public record showed "no contribution" rather than "contribution." Participants were fully informed about the structure of the noise.[10]

[8] Earlier experimental papers that investigate manipulating players' information in repeated games in less standard ways (such as presenting information with delay, or in a cognitively more complex manner) include Kahn and Murnighan (1993), Cason and Khan (1999), Sainty (1999), and Bolton, Katok, and Ockenfels (2005).

[9] This binary choice differs from Gächter, Renner, and Sefton (2008), as we aimed to implement a simple noise structure.

[10] In choosing the noise structure, we were mainly inspired by the contract theory literature on hidden action, starting with Mirrlees (1974, 1975), where it is standardly assumed that if an agent does not make any effort then the observed outcome is failure for sure, while making costly effort implies that the observed outcome is success

In the *no punishment* treatments, a round ended after the above information was displayed, and the experiment proceeded to the next round. In the *punishment* treatments, subjects participated in a second stage in each round. Here they were asked whether they wanted to assign up to five deduction points to the other two members of their group.[11] Each assigned reduction point decreased the payoff of the subject by one point. In the *regular punishment* treatments, each assigned deduction point implied a reduction of three points from the punished group member's income. Received punishment was capped at the earnings from the public goods game in the same round, however, while a punisher always had to pay for assigned punishment points. This punishment technology mimics the one used in Gächter, Renner, and Sefton (2008) and many other public good experiments in the literature. In the *strong punishment* treatments, each assigned reduction point reduced the income of the punished group member by six points, and that income reduction was not capped, therefore negative round incomes were allowed.[12]

The experimental sessions took place in February and March 2010 and 2011 at the Australian School of Business Experimental Research Laboratory at the University of New South Wales. Experimental subjects were recruited from the university student population using the Online Recruitment System for Economic Experiments (Greiner 2004). Overall, 339 subjects participated in 12 sessions, with between 24 and 30 subjects per session. Upon arrival, participants were seated in front of computers at desks that were separated by dividers. Participants received written instructions and could ask questions that were answered privately.[13] The experiment started after participants completed a short comprehension test at the screen. The experiment was programmed in zTree (Fischbacher 2007). At the end of the experiment, participants filled out a short demographic survey. They were then privately paid their cumulated experimental earnings in cash (with a conversion rate of AU$0.02 per point) plus a AU$5 show-up fee. The average earning was AU$28.94, with a standard deviation of AU$5.31.

II. Results

A. *Aggregate Results*

As groups stay constant over all 50 rounds, each group in our experiment constitutes one statistically independent observation. To test for treatment differences nonparametrically we apply two-sided Wilcoxon rank-sum tests, using group averages as independent observations.

Table 1 lists the average contributions, punishments, and net profits observed in our six treatments. Figures 1 and 2 display the evolution of public good contributions and net profits over time.

with positive probability. We note that most of the game theoretic literature on repeated games with imperfect public monitoring imposes a full support assumption on the public signal, which our noise structure does not satisfy.

[11] As in Gächter, Renner, and Sefton (2008), decisions/public records of the other two group members were always displayed anonymously in random ordering, in order to mitigate reputation effects across rounds. Punishment choices were elicited on that same ordering, though, so that punishment could be dedicated.

[12] The overall experiment income of a subject was capped at zero, however.

[13] See the online Appendix for a copy of the instructions.

TABLE 1—AVERAGE CONTRIBUTIONS, PUNISHMENT, AND NET PROFITS IN TREATMENTS

	N participants	Average contribution	Average punishment	Average net profits
No noise				
No punishment	57	5.59		22.80
Regular punishment	57	12.40	0.64	23.66
Strong punishment	54	17.61	0.48	25.45
Noise				
No punishment	57	4.04		22.02
Regular punishment	60	9.60	1.45	19.10
Strong punishment	54	16.04	0.65	23.48

As Table 1 reveals, *noise* leads to lower contributions in all three punishment conditions. This is, however, only statistically significant for *strong punishment* ($p = 0.011$) and not significant for *no* and *regular punishment* ($p = 0.511$ and $p = 0.144$, respectively).

The effects of punishment on contributions are more significant. Contributions increase monotonically from *no punishment* to *regular punishment* to *strong punishment*, both under no noise (p-values of 0.005, 0.030, and 0.001 for *regular punishment* versus *no punishment*, *strong punishment* versus *regular punishment*, and *strong punishment* versus *no punishment*, respectively) and noise (p-values of 0.004, 0.001, and 0.001, respectively).

With respect to the average number of assigned punishment points, Table 1 seems to suggest that there are less punishment points assigned when their effect is more severe.[14] This, however, is only significant in the *noise* treatments ($p = 0.001$), while statistically no such effect can be established when there is *no noise* ($p = 0.385$). On the other hand, both *regular* and *strong punishment* are more likely when there is *noise* than if there is *no noise* ($p = 0.001$ and $p = 0.068$, respectively).

Finally, while *noise* does not have a measurable effect on net payoffs when there is no punishment option available ($p = 0.511$), it (weakly) significantly decreases net payoffs when punishment is available ($p = 0.024$ and $p = 0.069$ for *regular* and *strong punishment*, respectively). Along the punishment dimension, when there is *no noise*, only *strong punishment* has a significant positive effect on payoffs compared to the baseline with *no punishment* ($p = 0.035$), while the differences between *regular punishment* and both other punishment treatments are insignificant ($p = 0.737$ and $p = 0.352$ when compared to *no punishment* and *strong punishment*, respectively). If there is noise then the picture looks different: the *regular punishment* condition yields lower net payoffs than both the baseline and the *strong punishment* condition, though this effect is only significant for the latter ($p = 0.319$ and $p = 0.033$, respectively). The robustness of these results is confirmed by further tests applied to data from only the last 30 or last 20 rounds.

Figures 1 and 2 suggest that after some initial volatility, contributions and net profits in the different treatments tend to stabilize over time, aside from relatively small endgame effects in the very last rounds (similar to Gächter, Renner, and Sefton

[14] This observation is closely related to the endogenously lower number of noncontributions. For a more in-depth analysis of punishment, see Table 3 and the discussion in Section IIB below.

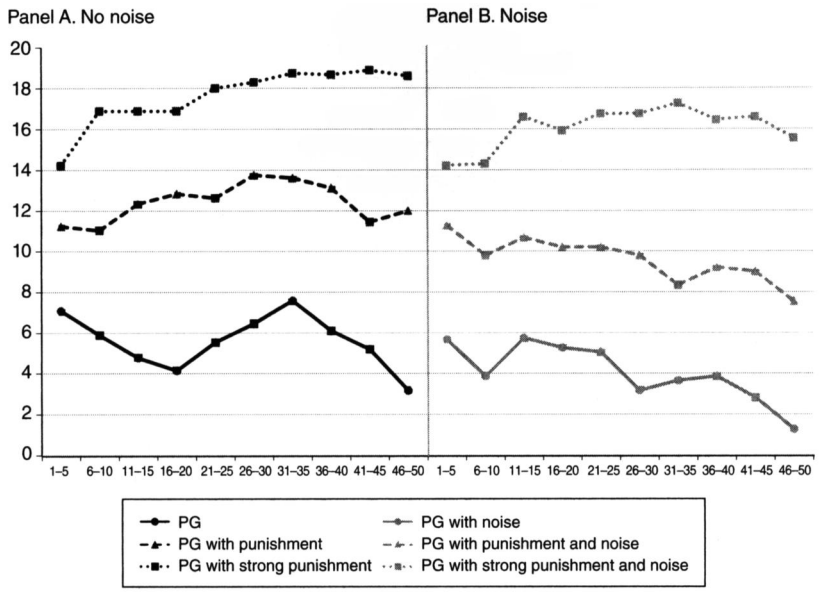

FIGURE 1. AVERAGE CONTRIBUTIONS OVER TIME

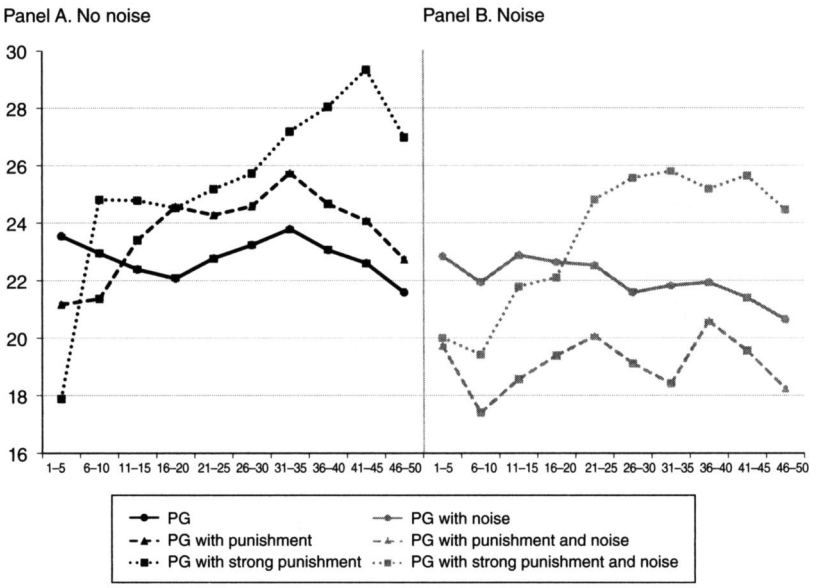

FIGURE 2. AVERAGE NET PROFITS OVER TIME

2008). This observation is corroborated by a battery of 2-sided Wilcoxon matched-pairs signed-rank tests comparing the average contributions and net profits in rounds 11 to 30 to those in rounds 31 to 50, which all yield p-values larger than 0.132, with the following exceptions: contributions increase over time with no noise and strong punishment ($p = 0.052$) and decrease with noise when there is no punishment or it is weak ($p = 0.016$ and $p = 0.011$, respectively), and net profits in the *no-punishment-noise* treatment are lower in later rounds ($p = 0.010$).

TABLE 2—PROBIT/TOBIT/OLS ESTIMATIONS OF CONTRIBUTIONS, PUNISHMENTS, AND NET EARNINGS BASED ON TREATMENT DUMMIES

Model dependent	Probit public good contribution	Tobit received punishment			OLS net earnings
		All	Defectors	Cooperators	
Intercept					21.54***
					[0.70]
Round	−0.001	−0.07***	−0.05***	−0.09***	0.05***
	[0.001]	[0.01]	[0.02]	[0.03]	[0.02]
Regular punishment	0.332***				0.86
	[0.096]				[1.36]
Strong punishment	0.576***	−1.37	0.33	0.04	2.65**
	[0.072]	[1.39]	[2.20]	[2.03]	[1.33]
Noise	−0.099	3.13***	2.42**	0.89	−0.78
	[0.101]	[1.22]	[1.14]	[1.93]	[0.80]
Noise × regular punishment	−0.041				−3.77*
	[0.150]				[2.01]
Noise × strong punishment	−0.030	−1.44	−1.48	−0.71	−1.19
	[0.169]	[1.66]	[2.45]	[2.40]	[1.65]
N	16,950	11,250	3,815	7,399	16,950
Pseudo R^2	0.195	0.035	0.017	0.018	
N left-censored		8,621	1,995	6,636	
N right-censored		75	67	8	
Adjusted R^2					0.053

Notes: For the probit/Tobit estimations on contributions/punishment, we report marginal effects rather than coefficients. Punishment points are not multiplied with the factor 3/6 yet, and are considered censored at 0 and 10. For all estimations, robust standard errors are clustered at group level and given in brackets.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

To complement the nonparametric analysis we ran probit, Tobit, and ordinary least square (OLS) regressions controlling for interaction effects between our treatments. In particular, we regressed contributions, received punishments and net earnings on the treatment dummies *regular punishment* and *strong punishment*, the dummy *Noise* (being 1 in all noise treatments), and interaction effects of *noise* with the two punishment dummies. All regressions also control for trends over time. As the groups of three participants are our units of statistically independent observations, we cluster standard errors by group.

Table 2 lists the results from this analysis. We find a strong positive effect of punishment on contributions to the public good, which is almost twice as large if punishment is more severe. Noise, on the other hand, has no significant effect on how much participants contribute. The number of assigned punishment points is not significantly affected when punishment is more severe, but noise increases this number significantly.[15] A split of the analysis into punishment of defectors versus punishment of cooperators shows that noise mostly affects the punishment toward

[15]This, however, is only the case with regular punishment. The total effect of noise under strong punishment; i.e., the joint effect of *Noise* and *Noise* × *Strong Punishment* is statistically not different from zero ($p = 0.1248$).

TABLE 3—AVERAGE PUNISHMENT POINTS SPENT, CONDITIONAL ON RECEIVER'S CONTRIBUTION AND PUBLIC RECORD

	All rounds		Only first round	
	Punishment	Strong punishment	Punishment	Strong punishment
No noise				
After contribution decision				
Contribution	0.212	0.316	0.114	0.771
Defect	1.338	1.681	1.636	3.000
Noise				
After public record				
Contribution	0.411	0.262	0.742	0.583
Defect	2.236	1.666	1.414	1.444

Note: Punishment points are not multiplied with factor 3 or 6 yet.

(thought-to-be) defecting group members. In the next subsection we analyze punishment patterns in more detail.

With respect to net earnings, punishment has a significant positive effect only when it is strong. When there is noise in addition to punishment, net payoffs are significantly reduced, but only under the regular punishment technology.[16] This leads to a U-shape of net earnings along the severity of punishment dimension under noise: regular punishment has a negative effect on net earnings, but with strong punishment this negative effect is mitigated by an additional positive earnings effect.

B. *Punishment Pattern*

Table 3 displays the average number of received punishment points conditional on the published contribution of a subject. Obviously, punishment received following a public record of no contribution is considerably higher than otherwise.[17] Even for cooperators, however, punishment levels are greater than zero. This might stem from antisocial punishment (see, for example, Herrmann, Thöni, and Gächter 2008), or could be an effect of some subjects punishing for older offenses (for indirect evidence for the presence of this effect, see Fudenberg and Pathak 2010). With regular punishment we observe higher punishment levels under noise (but significantly so only for punishment toward contributors, $p = 0.030$), while punishment levels are either unaffected by noise or slightly smaller when punishment is strong ($p = 0.331$ and $p = 0.033$ for punishment after contribution and no contribution records, respectively).

Comparing regular to strong punishment, we observe that punishment toward contributors is not affected by the punishment technology, neither with nor without noise, and neither in terms of assigned nor (multiplied) received punishment points (all p-values larger than 0.266). With respect to defectors, however, the number of received (multiplied) punishment points (the eventually resulting income reduction)

[16] Hypothesis F-tests confirm at the 5 percent level that the joint effect of *Regular Punishment* and *Noise* × *Regular Punishment* is negative, but cannot reject that the joint effect of *Strong Punishment* and *Noise* × *Strong Punishment* is different from zero ($p = 0.133$).

[17] This is strongly significant in all four punishment treatments, with all p-values smaller than 0.006. These and the following tests are based on the corresponding averages on the independent group level.

is larger if punishment is more severe, both with and without noise ($p = 0.027$ and $p = 0.001$, respectively), while the number of assigned points is only different if there is no noise ($p = 0.015$, versus $p = 0.827$ with noise). As a result, a stronger punishment technology leads to greater discrimination between contributors and defectors: while the former attract (not significantly) fewer punishment points, the latter are punished more heavily.

All these effects already exist in the very first round of the game (see the right part of Table 3), and are statistically significant except for the differences between regular and strong punishment. Since in the first round subjects cannot punish for older offenses, this provides clearer evidence that a public record of not contributing in a given round attracts more punishment points in the same round than does a public record of contributing. Contributors do receive some punishment even in the first round, though, indicating the existence of purely antisocial punishment.

C. *Reactions to Received Punishment*

We employ probit regression analysis to analyze reactions to received punishment and other previous experiences of subjects. In Model 1 of Table 4, we estimate the current-round contribution of a participant based on the number of punishment points she received in the previous round ($RecPnmt_{PR}$, not yet multiplied with the punishment factor). We control for the previous-round contribution of this participant ($Contr_{PR}$), and interact previous punishment and contribution with treatment dummies on whether noise was present (*Noise*), whether the strong punishment technology was present (*StrPnmt*), or both (*Noise* × *StrPnmt*).

We find that the $Contr_{PR}$ dummy has a large and significant effect. Our main interest, however, lies in the interaction terms. For noncontributors, the higher the received punishment, the more likely they are to contribute in the next round. This effect is significantly increased when the punishment has a stronger impact. When, on the other hand, contributors get punished, they are likely to decrease their contribution in the next round, and more so the higher the punishment. The punishment technology effect discussed above now works in the other direction, softening this discouraging effect when punishment is strong. In both cases, noise does not seem to play a role.

The probit Model 2 reported in Table 4 concentrates on choices under *Noise*, and explores whether having been a contributor with a *wrong* public record in the previous round ($PRwrong_{PR}$) has an effect on how the participant reacts to being punished by her group members. While in the new model any other effects are robust, the lack of significance for $Contr_{PR} \times PRwrong_{PR}$ suggests that having had a wrong public record does not influence contributions by itself, the significant positive effect on the interaction term with the received punishment indicates that the above contributors are less likely to reduce their contribution when being punished, and the effect is similar in both punishment regimes. Nevertheless, the net effect of received punishment on the next-round contribution of a subject with a wrong public record ($Contr_{PR} \times RecPnmt_{PR} + Contr_{PR} \times PRwrong_{PR} \times RecPnmt_{PR}$) is still negative.

Finally, Model 3 includes the average contribution of the other two group members ($OtherContr_{PR}$, scaled to [0,1]) as a control in the estimation equation of Model 1.

TABLE 4—PROBIT ESTIMATIONS OF CURRENT CONTRIBUTION BASED ON PREVIOUS ROUND BEHAVIOR

	Model 1	Model 2	Model 3
$RecPnmt_{PR}$	0.041***	0.022**	0.012
	[0.011]	[0.011]	[0.008]
$RecPnmt_{PR}$ × Noise	−0.009		0.004
	[0.013]		[0.010]
$RecPnmt_{PR}$ × StrPnmt	0.039**	0.0571***	0.023**
	[0.016]	[0.018]	[0.012]
$RecPnmt_{PR}$ × Noise × StrPnmt	0.012		−0.003
	[0.023]		[0.015]
$Contr_{PR}$	0.794***	0.705***	0.535***
	[0.025]	[0.037]	[0.046]
$Contr_{PR}$ × $RecPnmt_{PR}$	−0.144***	−0.124***	−0.073***
	[0.040]	[0.023]	[0.024]
$Contr_{PR}$ × $RecPnmt_{PR}$ × Noise	0.046		0.024
	[0.041]		[0.026]
$Contr_{PR}$ × $RecPnmt_{PR}$ × StrPnmt	0.076*	0.025	0.051*
	[0.043]	[0.026]	[0.027]
$Contr_{PR}$ × $RecPnmt_{PR}$ × Noise × StrPnmt	−0.070		−0.031
	[0.046]		[0.030]
$Contr_{PR}$ × $PRwrong_{PR}$		0.016	
		[0.057]	
$Contr_{PR}$ × $PRwrong_{PR}$ × $RecPnmt_{PR}$		0.085***	
		[0.021]	
$Contr_{PR}$ × $PRwrong_{PR}$ × $RecPnmt_{PR}$ × StrPnmt		−0.021	
		[0.034]	
$OtherContr_{PR}$			0.386***
			[0.051]
$Contr_{PR}$ × $OtherContr_{PR}$			0.121*
			[0.065]
Observations	11,025	5,586	11,025
Pseudo R^2	0.454	0.353	0.535

Notes: We report marginal effects rather than coefficients. $Contr_{PR}$ and $RecPnmt_{PR}$ refer to contribution and punishment received in the previous round, respectively, while $PRwrong_{PR}$ indicates whether the public record of a contributor in the previous round was wrong, and $OtherContr_{PR}$ represents the average contribution (scaled [0,1]) of the other two group members in the previous round. *Noise* and *StrPnmt* are dummies indicating whether noise or the strong punishment technology were present. Robust standard errors, clustered at group level, are given in brackets.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

We find that current contributions are indeed highly positively correlated with the other group members' previous-round contributions. This might be interpreted as an alternative type of punishment by reducing future payoffs (though such punishment cannot be targeted toward an individual), or as evidence for coordination on and convergence to a group norm. The inclusion of these controls reduces the positive effect of punishment on subsequent contributions of noncontributors, but the effect remains significantly positive in the strong punishment treatment. The negative effects of punishment on contributors' subsequent choices are robust to including the controls. These results, however, have to be interpreted with care due to multicollinearity, as the relation between own and others' contributions in the previous round ($Contr_{PR}$ and $OtherContr_{PR}$) is highly correlated with the subsequently received punishment ($RecPnmt_{PR}$).

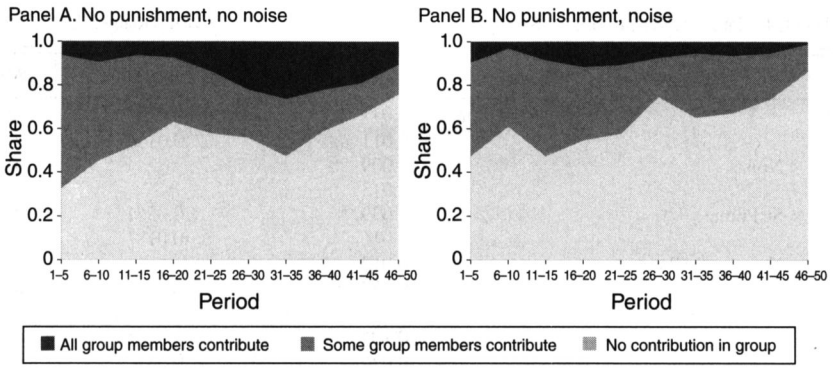

FIGURE 3. NO PUNISHMENT TREATMENTS—GROUP COOPERATION OVER TIME

D. *Evolution of Cooperation and Punishment in Groups*

In Figures 3 and 4 we classify the groups in the different treatments by whether there was full, partial, or no contribution to the public good in different rounds, and study the emergence of such groups over time. Figure 4 additionally includes the pattern of punishment over time for groups that started and ended with full public good contributions, groups that started low but converged to full contributions after some time, and groups that did not manage to reach full contributions.

Figure 3 shows that when there is no punishment available, groups who started out with at least some contributions become no-contribution groups over time. As we observe on the left side of Figure 4, under regular punishment and if there is no noise, most groups polarize such that either all or none of the group members contribute. When we add noise to the information about others' contributions, we observe higher dispersion of contributions within groups such that there is no convergence to polarized groups but there is an increase in the number of no-cooperation groups. Under a severe punishment regime, groups quickly converge to homogenous full-contribution groups. This general tendency stays intact with noise in the public information.

We statistically confirm these observations with a battery of Fisher's exact tests, comparing the modal behavior of groups in the first ten rounds and in the last ten rounds of a treatment.[18] We find that when there is no punishment available, then the share of groups who at least partly contribute shrinks and the share of groups with no contributions at all increases over time, both with and without noise (all p-values smaller than 0.10). On the contrary, under the strong punishment regime, the share of partly contributing groups decreases, too, but is accompanied by a significant increase in full contribution groups (all p-values smaller than 0.05), also irrespective of the presence of noise. With regular punishment, however, we observe a significant decrease in the share of partly contributing groups when there is no

[18] If there was no unique modal behavior in a group over the ten rounds, the group was assigned to the class of "some contributed," which represents the median in these cases. The results reported below are largely robust when using the median rather than the modal behavior, or using only rounds 1–5 and 46–50, or only the first and the last round, respectively.

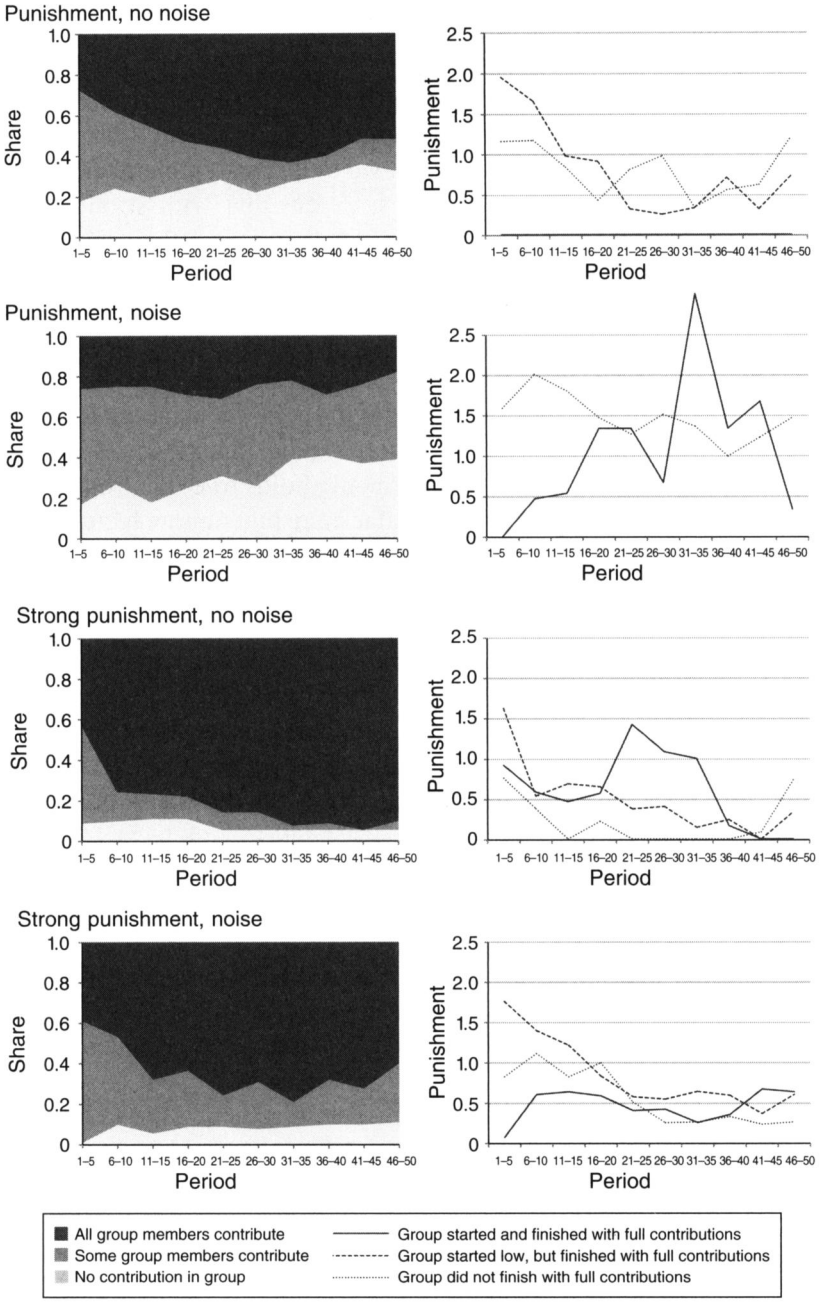

FIGURE 4. PUNISHMENT TREATMENTS—GROUP COOPERATION OVER TIME AND AVERAGE PUNISHMENT IN DIFFERENT COOPERATION CLASSES

noise ($p = 0.005$), and we find a weakly significant increase in the share of groups who do not contribute at all when there is noise ($p = 0.065$), in both cases with no significant effect on the individual shares of the other two group types.

With respect to the effects of noise on observed group types, we do not find any significant differences in the first ten rounds of all punishment regimes (all p-values larger than 0.18), and we find differences for the last ten rounds only for regular punishment,

for which the share of groups with full contributions in the last ten rounds is significantly lower ($p = 0.048$) and the share of partially contributing groups weakly significantly higher ($p = 0.065$) when there is noise than when there is no noise.

When comparing between punishment conditions in the very first round (before players start to respond to others' actions), we find that treatments do not start out with different distributions of group types, except that with strong punishment there are fewer no-contribution groups in the first round than without punishment ($p = 0.020$ and $p = 0.008$ for no noise and noise, respectively).[19] In the last ten rounds, however, we find significant differences in the distribution of group types across punishment conditions. When there is no noise, we have a monotone increase of the last-round share of full-contribution groups and a monotone decrease of the share of no-contribution going from no punishment to regular punishment and strong punishment (all p-values smaller than 0.05). Comparing the end of the treatments with noise, this pattern only holds true for strong punishment (all p-values smaller than 0.05), while regular punishment now features more partial-contribution groups than no punishment and strong punishment ($p = 0.020$ and $p = 0.067$, respectively).

The right side of Figure 4 displays average punishment in different classes of groups. If there is no noise, then groups that start with full contributions and end with full contributions experience no punishment at all during the game. While we do not observe such groups under noise and regular punishment, we observe a positive but small amount of punishment in such groups under noise and a strong punishment regime (potentially indicating successful but costly coordination on cooperation).

III. Conclusion

This paper finds that while in a perfect monitoring public good contribution environment increasing the severity of a costly punishment option unambiguously increases average net payoffs, in an imperfect monitoring environment the relationship is nonmonotonic. Moreover, at least for some punishment technologies, the presence of costly punishment can be detrimental to society. This weakens the case that group selection evolutionary procedures lead to emotional responses like anger and revenge, inducing individuals to punish cheaters.

A possible direction for future research is reexamining the questions addressed in this paper using data from real-world environments in which dissatisfied participants can punish each other, such as feedback scores in electronic commerce, or grades and teacher evaluations in higher education.

[19] Within the first ten rounds, some further differences emerge. The share of no contribution groups also becomes significantly different when comparing no punishment and regular punishment in the noisy environment ($p = 0.006$). Under noise, more groups fully cooperate in the strong punishment regime than with regular or without punishment ($p = 0.099$ and $p = 0.001$, respectively), while with noise this difference is only significant comparing between strong and no punishment ($p = 0.019$).

REFERENCES

Abbink, Klaus, and Abdolkarim Sadrieh. 2009. "The Pleasure of Being Nasty." *Economics Letters* 105 (3): 306–08.

Ambrus, Attila, and Ben Greiner. 2012. "Imperfect Public Monitoring with Costly Punishment: An Experimental Study: Dataset." *American Economic Review*. http://dx.doi.org/10.157/aer.102.7.3317.

Aoyagi, Masaki, and Guillaume Frechette. 2009. "Collusion as Public Monitoring Becomes Noisy: Experimental Evidence." *Journal of Economic Theory* 144 (3): 1135–65.

Bereby-Meyer, Yoella, and Alvin E. Roth. 2006. "The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation." *American Economic Review* 96 (4): 1029–42.

Bolton, Gary E., Elena Katok, and Axel Ockenfels. 2005. "Cooperation among Strangers with Limited Information about Reputation." *Journal of Public Economics* 89 (8): 1457–68.

Bornstein, Gary, and Ori Weisel. 2010. "Punishment, Cooperation, and Cheater Detection in "Noisy" Social Exchange." *Games* 1 (1): 18–33.

Bowles, Samuel. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton, NJ: Princeton University Press.

Boyd, Robert, and Peter J. Richerson. 1992. "Punishment allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and Sociobiology* 13 (3): 171–95.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences of the United States* 100 (6): 3531–35.

Cason, Timothy N., and Feisal U. Khan. 1999. "A Laboratory Study of Voluntary Public Goods Provision with Imperfect Monitoring and Communication." *Journal of Development Economics* 58 (2): 533–52.

Dreber, Anna, David G. Rand, Drew Fudenberg, and Martin A. Nowak. 2008. "Winners Don't Punish." *Nature* 452: 348–51.

Egas, Martijn, and Arno M. Riedl. 2008. "The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society B* 275 (1637): 871–78.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90 (4): 980–94.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137–40.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–78.

Fudenberg, Drew, and Parag A. Pathak. 2010. "Unobserved Punishment Supports Cooperation." *Journal of Public Economics* 94 (1–2): 78–86.

Fudenberg, Drew, David G. Rand, and Anna Dreber. 2012. "Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World." *American Economic Review* 102 (2): 720–49.

Gächter, Simon, Elke Renner, and Martin Sefton. 2008. "The Long-Run Benefits of Punishment." *Science* 322: 1510.

Gong, Min, Jonathan Baron, and Howard Kunreuther. 2009. "Group Cooperation under Uncertainty." *Journal of Risk and Uncertainty* 39 (3): 251–70.

Grechenig, Kristoffel, Andreas Nicklisch, and Christian Thöni. 2010. "Punishment Despite Reasonable Doubt—A Public Goods Experiment with Sanctions under Uncertainty." *Journal of Empirical Legal Studies* 7 (4): 847–67.

Greiner, Ben. 2004. "An Online Recruitment System for Economic Experiments." In *Forschung und wissenschaftliches Rechnen 2003*, edited by Kurt Kremer and Volker Macho, 79–93. Göttingen, Germany: Ges. für Wiss. Datenverarbeitung.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach. 2006. "The Competitive Advantage of Sanctioning Institutions." *Science* 312 (5770): 108–11.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science* 319 (5868): 1362–67.

Hopfensitz, Astrid, and Ernesto Reuben. 2009. "The Importance of Emotions for the Effectiveness of Social Punishment." *Economic Journal* 119 (540): 1534–59.

Hwang, Sung-Ha, and Samuel Bowles. 2010. "Is Altruism Bad for Cooperation?" Unpublished.

Kahn, Lawrence M., and J. Keith Murnighan. 1993. "Conjecture, Uncertainty, and Cooperation in Prisoner's Dilemma Games: Some Experimental Evidence." *Journal of Economic Behavior and Organization* 22 (1): 91–117.

Krueger, Alan B., and Alexandre Mas. 2004. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy* 112 (2): 253–89.

Mas, Alexandre. 2008. "Labour Unrest and the Quality of Production: Evidence from the Construction Equipment Resale Market." *Review of Economic Studies* 75 (1): 229–58.

Miller, John H. 1996. "The Coevolution of Automata in the Repeated Prisoner's Dilemma." *Journal of Economic Behavior and Organization* 29 (1): 87–112.

Mirrlees, James A. 1974. "Notes on Welfare Economics, Information and Uncertainty." In *Essays on Economic Behavior under Uncertainty*, edited by M. Balch, D. McFadden, and S. Wu, 243–58. Amsterdam: Elsevier Science, North Holland.

Mirrlees, James A. 1975. "The Theory of Moral Hazard and Unobservable Behavior: Part I." Unpublished.

Nikiforakis, Nikos, and Hans-Theo Normann. 2008. "A Comparative Statics Analysis of Punishment in Public-Good Experiments." *Experimental Economics* 11 (4): 358–69.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review* 86 (2): 404–17.

Patel, Amrish, Edward Cartwright, and Mark van Vugt. 2010. "Punishment Cannot Sustain Cooperation in a Public Good Game with Free-Rider Anonymity." University of Gothenburg Working Paper in Economics 451.

Rand, David G., Anna Dreber, Tore Ellingsen, Drew Fudenberg, and Martin A. Nowak. 2009. "Positive Interactions Promote Public Cooperation." *Science* 325 (5945): 1272–75.

Sainty, Barbara. 1999. "Achieving Greater Cooperation in a Noisy Prisoner's Dilemma: An Experimental Investigation." *Journal of Economic Behavior and Organization* 39 (4): 421–35.

Yamagishi, Toshio. 1986. "The Provision of Sanctioning Systems as a Public Good." *Journal of Personality and Social Psychology* 51 (1): 110–16.