Democratic punishment in public good games with perfect and imperfect observability*

Attila Ambrus† and Ben Greiner‡

January 29, 2015

Abstract

In the context of repeated public good contribution games, we experimentally investigate the impact of democratic punishment, that is when members of a group decide by majority voting whether to inflict punishment on another member, relative to individual peer-to-peer punishment, on cooperation levels. We find that democratic punishment leads to more cooperation and higher average payoffs than individual punishment, both when members can monitor each other perfectly and in the case of imperfect monitoring, for which scenario previous experimental research found it difficult to maintain cooperation. Democratic punishment achieves this by curbing anti-social punishment and thereby establishing a closer connection between a member's contribution decision and whether subsequently being punished by others. Participating in the democratic punishment procedure makes even non-contributors punishment intentions more pro-social. Additionally, when monitoring is imperfect, democratic punishment discourages future contributions less than individual punishment when a contributor is mistakenly observed as a non-contributor.

Keywords: public good contribution experiments, punishment, voting

JEL Classification: C72, C92, H41

*Financial support through an Australian Research Council Discovery Grant is gratefully acknowledged. We thank Justin Cheong and Johannes Hoelzemann for excellent research assistance.

†Duke University, Department of Economics, Durham, NC 27708, e-mail: aa231@duke.edu

‡University of New South Wales, School of Economics, Sydney, NSW 2052, email: bgreiner@unsw.edu.au

I Introduction

Several papers in the experimental literature, starting from Fehr and Gächter (2000), demonstrated that the availability of a costly punishment option for individuals can increase cooperation in public good contribution games. Gächter, Renner and Sefton (2008) showed that this increases overall net payoffs in the population, provided that the time horizon for interaction is long enough. However, Grechenig, Nicklisch and Thöni (2010) and Ambrus and Greiner (2012) found that the above results hinge on the assumption that individuals can perfectly monitor each others' actions. If there is a small amount of noise in monitoring, then the availability of costly individual punishment does not help the participants' welfare, and in some cases it can even decrease it. The reason is that with imperfect monitoring from time to time a contributor gets punished by fellow team members who received an incorrect negative signal regarding the contribution. This discourages future contributions and can trigger antisocial punishment by the contributor who was "unfairly" punished.[1] Hence even in the long run, contribution levels stay away from the socially efficient levels, and individuals keep on using punishing each other, further decreasing each others' payoffs. Moreover, in a recent paper Fischer, Grechenig and Meier (2013) find that if monitoring is imperfect, centralizing punishment, in the form of delegating punishment rights to a particular individual, does not remedy the issues above, and cooperation levels remain low.

In this paper we find that democratic punishment, in the form of group members after each round of the contribution game deciding which members to punish using simple majority rule, outperforms individual punishment, both in terms of cooperation levels and average payoffs, and in both perfect

[1] In experiments on social dilemma games with imperfect observability and no direct punishment option available, Aoyagi and Fréchette (2009) and Fudenberg, Rand and Dreber (2012) find that players under noise are more forgiving than without noise. On the prevalence of anti-social punishment in public good contribution games with individual punishment, see Cinyabuguma, Page and Putterman (2006), Herrmann, Thöni and Gächter (2008) and Nikiforakis (2008). Hauser, Nowak and Rand (2014) provide a theoretical analysis in the context of a dynamic learning model, explaining why punishment might not promote cooperation when anti-social punishment is possible.

and imperfect monitoring environments. A key reason is that democratic punishment mitigates anti-social punishment, and makes the relationship between one's contribution decision and whether she gets subsequently punished clearer: Specifically, it makes it more likely that contributing members do not get punished, and that non-contributing members get punished. In particular it greatly reduces the opportunities of those who get punished by others for non-contributing to punish back, either preemptively or subsequently. We find that even non-contributors adopt the rule of voting to punish (other) non-contributors. This suggests that participation in a democratic procedure, even if the procedure itself is exogenously given, facilitates pro-social behavior, in the sense of enforcing social norms. This finding complements Dal Bó et al. (2010), who show that endogenous democratic adoption of a policy that automatically fines unilateral non-contributors increases cooperation relative to when the same policy is imposed on the group exogenously, and it is in line with the finding in several papers (Frey, 1994; Frey, Benz and Stutzer, 2004; Pommerehne and Weck-Hannemann, 1996) that there is a positive relationship between direct-democratic participation rights and pro-social behavior.

We also find evidence that individuals react differently, with respect to subsequent contributions to the public good, when they are punished democratically by group members versus when they get punished individually by fellow members. In both cases getting punished after not contributing increases expected contribution in the next round. The difference is that when an individual gets punished even though he contributed (but others observed an incorrect negative signal about the contribution), this punishment discourages her to contribute in the next round in the individual punishment treatment, but not in the democratic punishment treatment.

At the aggregate level, in the individual punishment treatment the above effects in case of imperfect monitoring result in more and more groups where cooperation ceases to exist, while we do not observe such convergence towards no-contribution groups in the group punishment treatment.

Our experimental design involves groups of five subjects, playing twenty times repeated public good contribution games. In the individual punish-

ment treatment, after each round each group member decides independently which other members to punish. In the democratic punishment treatment, after each round members simultaneously cast votes which members should be punished, and punishment is inflicted on those members who received at least three votes. In order to put the two punishment schemes on an equal footing, we set payoffs in the democratic punishment treatment such that if the group votes to punish a member, the punishment inflicted is the same as when all four other members punish the member in the individual punishment treatment. Similarly, the cost of a group punishment on each of the other members is the same as the cost of punishing in the individual punishment treatment.

In our setting, members of a group cannot commit ex ante to a particular punishment rule, instead in each round a majority decides on whether to punish someone or not. There are several papers in the literature taking a different approach, in which there is a democratic group decision at the beginning of the game, deciding on whether to adopt a punishment scheme (either the option of individual punishment or an automated punishment rule) and in some cases on features of the punishment scheme (how severe punishment is allowed to be, or who can be punished): see Andreoni and Gee (2012); Bó, Foster and Putterman (2010); Ertan, Page and Putterman (2009); Kamei, Putterman and Tyran (forthcoming); Markussen, Putterman and Tyran (2014); Sutter, Haigner and Kocher (2010); Tyran and Feld (2006). Other studies allow the punishment to be delegated to a specific subject, who carries them out without commitment: see for example Baldassarri and Grossman (2011); Fehr and Fischbacher (2004); Leibbrandt and López-Pérez (2011, 2012).[2] Lastly, Cinyabuguma et al. (2006) investigate a setting in which after each round group members can vote whether to expel certain members of the group, and show that the threat of expulsion can facilitate more cooperation. All of the above papers only consider settings with perfect monitoring, as opposed to our study.

[2]For a related theoretical analysis, see Aldeshev and Zanarone (2014).

II Experimental Design

We implemented four treatments in a 2×2 factorial design. Our main comparison is between a repeated 5-person public good game that allows for individual punishment and a public good game in which a majority of group member votes is required in order to punish another group member. We employ both games in two different environments, one with perfect observation of other group members' contributions, and one in which the signal about other group member's contribution is noisy, such that there is a small chance of 10 percent that a contribution is displayed to others as a defection.[3]

At the beginning of the experiment, participants were matched to groups of five, which stayed constant for all 20 rounds. Within each group, participants were assigned IDs from 1 to 5, which also stayed constant for the course of the experiment. Each round consisted of 2 stages, a public good contribution stage and a punishment stage.In the public good contribution stage, each group member was endowed with 50 points, and decided whether she wanted to contribute these 50 points to a "project" or not. If the endowment was kept, it increased the participant's payoff by 50 points. If the endowment was contributed, it benefitted each of the five group members by 0.3 times 50 = 15 points. Thus, if no group member contributed, each would earn 50 points, while the symmetric efficient outcome of 75 points for each could be reached if all contributed their endowment.

Our treatments differ only in the second stage of each round. First, after their simultaneous decisions in Stage 1, participants were informed about the contribution of each group member in their group. In our *No noise* treatments, the actual contribution of the respective participant was displayed. In the *Noise* treatments, the display showed a "public record" of each group member's contribution. Participants were informed that if a group member did not contribute his endowment, then the public record would always indicate "no contribution". If the group member contributed, however, then there was a 10 percent chance that the public record showed "no contribution" rather than "contribution".

[3]The same design of imperfect monitoring was used in Ambrus and Greiner, 2012.

Second, participants were asked to indicate their willingness to monetarily punish ("reduce the earnings of") each other group member. In our *Individual Punishment* treatments, each group member could directly reduce the earnings of another group member by 15 points, at a cost of 5 points. In the *Democratic punishment* condition, group members simultaneously cast votes for each group member whether to punish that group member or not. Thus, for each group member, votes from all four other group members were collected. If three or more group members voted to punish a participant, then the earnings of that group member were reduced by 60 points, and each of the other four group members (independent of how they voted) incurred a cost of 5 points for this punishment. If no majority was reached (because two or less group members voted for punishment), then no points are reduced and no costs incurred. Thus, the equivalent of a punishment by a group (when majority is reached) in the *Democratic Punishment* treatments is being punished by each other group member in the *Individual Punishment* treatments, and the equivalent of no group punishment (because there was no majority to punish) is not being punished at all in the *Individual Punishment* treatments.

After all participants simultaneously made their punishment decisions, they were informed about the punishments and votes in their group, and the consequences for their round payoffs. In the *Noise* treatments, any payoff information was provisional based on public records; participants were informed about their true earnings in each round at the end of the experiment.

The experimental sessions took place in March and April 2014 at the Business School Experimental Research Laboratory at the University of New South Wales. Experimental subjects were recruited from the university student population using the online recruitment system ORSEE (Greiner, 2004). Overall, 325 subjects participated in 12 sessions, with either 20, 25, or 30 subjects per session. Upon arrival participants were seated in front of a computer at desks which were separated by dividers. Participants received written instructions and could ask questions which were answered privately. The experiment was programmed in zTree (Fischbacher, 2007). Sessions

lasted about one hour. 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. Participants could incur losses in a particular round, but session losses were capped at the show-up fee. No participant incurred losses over the whole session. The average earning was AU$ 27.81 (including showup-fee), with a standard deviation of AU$ 4.05, a minimum payoff of AU$ 16.20 and a maximum payoff of AU$ 35.30.

TABLE 1: Average contributions, punishment and net profits in treatments

	N part.	N groups	Avg. contr.	Avg. punishm.	Avg. net profits
No noise					
Individual punishment	75	15	23.33	5.96	53.72
Democratic punishment	80	16	36.75	2.40	65.18
Noise					
Individual punishment	80	16	18.78	6.36	50.92
Democratic punishment	90	18	27.58	4.37	57.97

III Results

III.A Aggregate results

Table 1 lists the average contributions, punishments, and net profits observed in our four treatments. Figures 1, 2 and 3 display the evolution of public good contributions, punishment, and net profits over time. As groups stay constant over all 20 rounds, each group in our experiment constitutes one statistically independent observation. To test for treatment differences non-parametrically, we apply 2-sided Wilcoxon rank-sum tests, using group averages as independent observations. Table 2 reports the results.

In both the perfect monitoring and the noisy environment, we observe higher contributions, less punishment (only significant for the *No Noise* condition), and consequently higher net profits when groups vote over punishment compared to when group members can punish individually. Introduc-

TABLE 2: P-values from Non-parametric Wilcoxon Ranksum tests across treatment dimensions

	Contributions	Received Punishment	Net profits
Individual Punishment vs. Democratic punishment			
with No noise	0.012**	0.005***	0.000***
with Noise	0.055*	0.137	0.003***
No noise vs. Noise			
with Individual punishment	0.489	0.874	0.385
with Democratic punishment	0.023**	0.030**	0.005***
Diagonal treatment comparisons			
No noise-Ind. vs. Noise-Democratic	0.469	0.192	0.036**
Noise-Ind. vs. No noise-Democratic	0.002***	0.003***	0.000***

ing *Noise* in the observation of other group members' contribution behavior lowers contributions and net profits, and increases observed punishment for both when punishment is individual as well as when punishment is a group decision, but statistically significantly so only for the latter environment.[4,5]

The regressions reported in Table 3 confirm and further detail these results. We estimate the likelihood of contribution (Model 1), the amount of punishment points received in a round (Models 2-5), as well as the net profits in a round (Model 6) using treatment dummies and a *Round* control. The dummy *Noise* equals 1 in the Noise treatments and 0 otherwise; the dummy *Democratic Punishment* is 1 for the treatments with voting over punishment and 0 in the individual punishment treatments; and the interaction effect *Noise* $\times$*Democratic punishment* equals 1 only in the respective treatment with democratic punishment under noise. For each estimation we ran additional post-estimation F-tests in order to determine the total effect of *Noise* under democratic punishment (Noise + N$\times$DP) and the total effect of *Democratic punishment* under noise (DP + N$\times$DP).

[4]Ambrus and Greiner (2012) only study an individual punishment environment and find a significant effect of noise on all three observables. However, in Ambrus and Greiner (2012) the game was repeated 50 times (while only 20 times here) and featured smaller (3-person) groups.

[5]The statistical comparison across diagonal treatment cells of our 2$\times$2 design is consistent with that, showing strong difference between our least efficient treatment with Individual Punishment under Noise compared to our most efficient treatment condition employing Democratic Punishment in a No Noise environment, and no differences (except for profits) between the other two – in terms of efficiency "intermediate" – treatments.

FIGURE 1: Average contributions over time

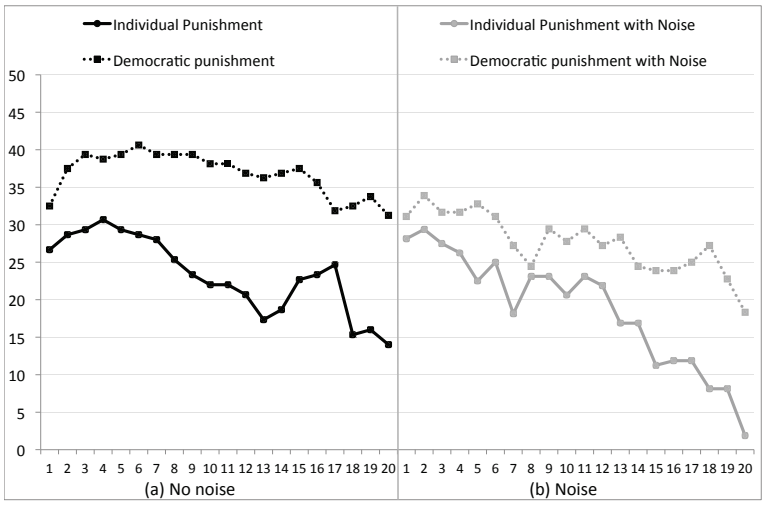
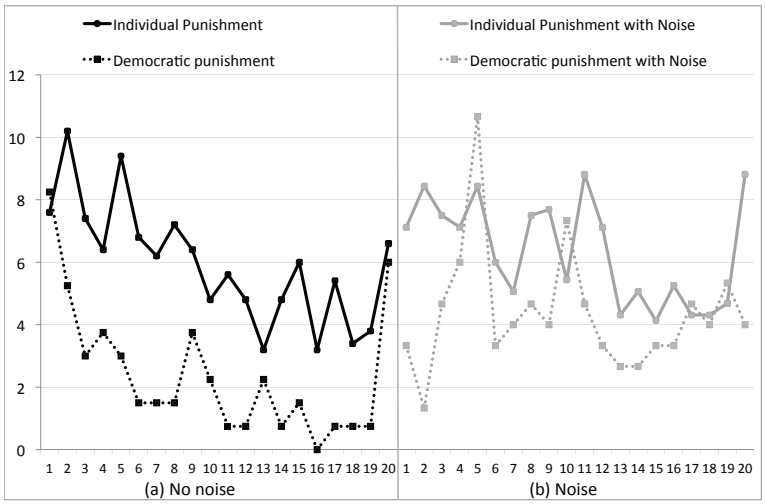
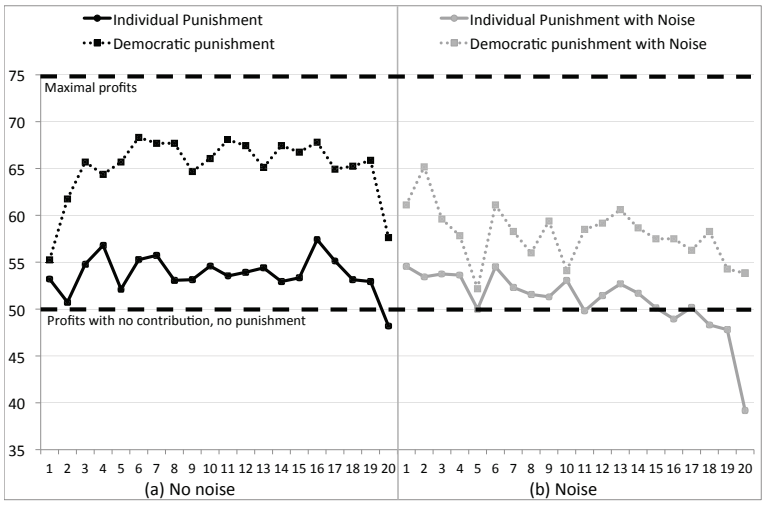


FIGURE 2: Average received punishment over time



The results of Model 1 and 6 in Table 3 replicate the non-parametric tests, in that we observe a significant increase in contributions and net profits when the group votes to punish compared to individual punishment (both when there is perfect and imperfect monitoring), and that noise has a statistically significant detrimental effect on contributions and net profits only in the democratic punishment condition.

FIGURE 3: Average net profits over time



The Models 2 to 5 in Table 3 explore effects of treatment conditions on punishment behavior. Model 2 predicts all punishments (independent of towards whom they were directed), and shows that introducing democratic punishment significantly reduces overall punishment in both non-noisy and noisy environments. *Noise* increases punishments when groups punish but not when individuals punish, which is related to the observation that democratic voting seems to be less effective in reducing punishments under noise than when there is no noise. Models 3 and 4 regress punishment of defectors (as identified by their public record) and cooperators, respectively. The results show that democratic punishment leads to a significant decrease of punishment of cooperators in both environments, but to a decrease of punishment of defectors only in the noise environment and there only weakly significantly. Model 5 serves the purpose to show that due to the relatively low likelihood of "noise" in public records, the punishment patterns towards "true cooperators" (some of which might have a wrong public record of no defection) are very similar to those towards the subset of cooperators who are clearly identified as such by their public record.

TABLE 3: PROBIT/TOBIT/OLS ESTIMATIONS OF CONTRIBUTIONS, PUNISHMENTS AND NET EARNINGS BASED ON TREATMENT DUMMIES

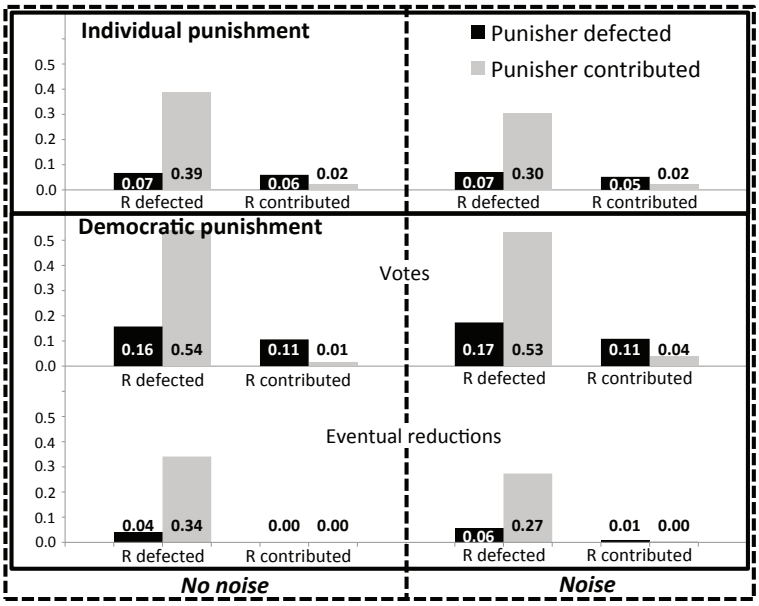
Dependent	Public Good Contribution	Received Punishment				Net Profits
		All	PR Defect	PR Coop.	True Coop.	
Model	Probit	Tobit	Tobit	OLS	OLS	OLS
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept		-34.22***	-3.07	2.42***	2.21**	55.40***
		[11.44]	[9.70]	[0.84]	[0.86]	[1.98]
Round	-0.014***	-0.97***	-1.87***	-0.04	-0.01	-0.16*
	[0.003]	[0.34]	[0.43]	[0.03]	[0.03]	[0.08]
Noise	-0.089	5.24	-0.67	0.00	1.28	-2.80
	[0.084]	[8.23]	[9.86]	[1.04]	[1.00]	[2.48]
Democratic punishment	0.266***	-52.90***	-19.48	-2.05***	-2.07***	11.46***
	[0.102]	[10.04]	[19.42]	[0.69]	[0.69]	[2.72]
Noise ×	-0.097	15.29	-1.28	0.18	0.71	-4.40
Democratic punishment	[0.130]	[12.74]	[22.79]	[1.04]	[1.04]	[3.34]
P-values from post-estimation F-tests						
Noise + N×DP = 0	0.067	0.040	0.924	0.098	0.000	0.002
DP + N×DP = 0	0.021	0.000	0.067	0.017	0.076	0.000
N	6500	6500	3190	3310	3470	6500
Pseudo R-squared	0.072	0.030	0.012			
N left-censored		5410	2258			
N right-censored		219	216			
Adjusted R-squared				0.058	0.028	0.071

Note: For the Probit estimation on contributions, we report marginal effects dy/dx rather than coefficients. Received punishment points are censored at 0 and 60, but Models 4 and 5 do not converge as Tobit models, so we report results from OLS regressions in these cases. For all estimations, robust standard errors are clustered at group level and given in brackets. *, **, and *** indicate significance at the 10%, 5%, and 1%-level, respectively.

III.B Punishment pattern

Figure 4 shows the punishment pattern in our four treatments. It displays the frequency of punishment conditional on whether the punisher contributed or not and whether the punishment receiver contributed or not. For the democratic punishment treatments, the figure distinguishes between votes for punishment and eventual punishment (when votes for punishment

FIGURE 4: Frequency of (vote for and eventual) punishment, conditional on punisher's own contribution and receivers' public record



reached the required majority). Table 4 displays results from non-parametric tests comparing the results reported in Figure 4 along treatment dimensions and punishment source and target characteristics.[6]

In general, contributors are much more likely than non-contributors to punish defectors (highly significant except for eventual democratic punishment under *No Noise*), but are less likely than non-contributors to punish contributors (but not significantly so). As one would expect, defectors attract more punishment than contributors, significantly so from contributors and under democratic punishment also from defectors.

[6]Since we are employing a full battery of tests here, we decided to adjust the p-values required for a particular significance level with a Bonferroni correction. We assume each set of four tests in Table 4 to belong the the same 'family' of hypotheses, and correspondingly divide the required p-value for a particular significance level by 4. As a result, a Null hypothesis is rejected at the 10% level when the p-value is 0.025 or below, and it is rejected at the 5% level (1% level) when the p-value is 0.0125 (0.0025) or below, respectively. Table 4 reports the original p-values obtained from the tests, but the stars represent the corrected significance level. As before, group-level averages serve as independent observations.

TABLE 4: P-values from non-parametric tests comparing results reported in Figure 4

Individual punishment vs. Democratic punishment (votes)					
No noise				*Noise*	
P defect, R defect	0.029			P defect, R defect	0.000***
P defect, R contr	0.791			P defect, R contr	0.171
P contr, R defect	0.093			P contr, R defect	0.000***
P contr, R contr	0.511			P contr, R contr	0.049
Democratic punishment (votes) vs. Democratic punishment (eventual)					
No noise				*Noise*	
P defect, R defect	0.728			P defect, R defect	0.001***
P defect, R contr	0.003***			P defect, R contr	0.000***
P contr, R defect	0.030			P contr, R defect	0.000***
P contr, R contr	0.005**			P contr, R contr	0.000***
Individual punishment vs. Democratic punishment (eventual)					
No noise				*Noise*	
P defect, R defect	0.187			P defect, R defect	0.730
P defect, R contr	0.000***			P defect, R contr	0.010**
P contr, R defect	0.373			P contr, R defect	0.796
P contr, R contr	0.001***			P contr, R contr	0.051
No noise vs. Noise					
Individual punishment		*Democratic punishment (votes)*		*Democratic punishment (eventual)*	
P defect, R defect	0.874	P defect, R defect	0.894	P defect, R defect	0.401
P defect, R contr	0.791	P defect, R contr	0.648	P defect, R contr	0.092*
P contr, R defect	0.206	P contr, R defect	0.091	P contr, R defect	0.051*
P contr, R contr	0.343	P contr, R contr	0.046	P contr, R contr	0.176
Punisher defected vs. Punisher contributed					
Individual punishment		*Democratic punishment (votes)*		*Democratic punishment (eventual)*	
No noise, R defect	0.001***	No noise, R defect	0.001***	No noise, R defect	0.071
No noise, R contr	0.728	No noise, R contr	0.023*	No noise, R contr	no diff
Noise, R defect	0.001***	Noise, R defect	0.000***	Noise, R defect	0.004**
Noise, R contr	0.074	Noise, R contr	0.045	Noise, R contr	0.084
Receiver defected vs. Receiver contributed					
Individual punishment		*Democratic punishment (votes)*		*Democratic punishment (eventual)*	
No noise, P defect	0.019*	No noise, P defect	0.012**	No noise, P defect	0.002***
No noise, P contr	0.001***	No noise, P contr	0.000***	No noise, P contr	0.001***
Noise, P defect	0.063	Noise, P defect	0.002***	Noise, P defect	0.000***
Noise, P contr	0.001***	Noise, P contr	0.000***	Noise, P contr	0.000***

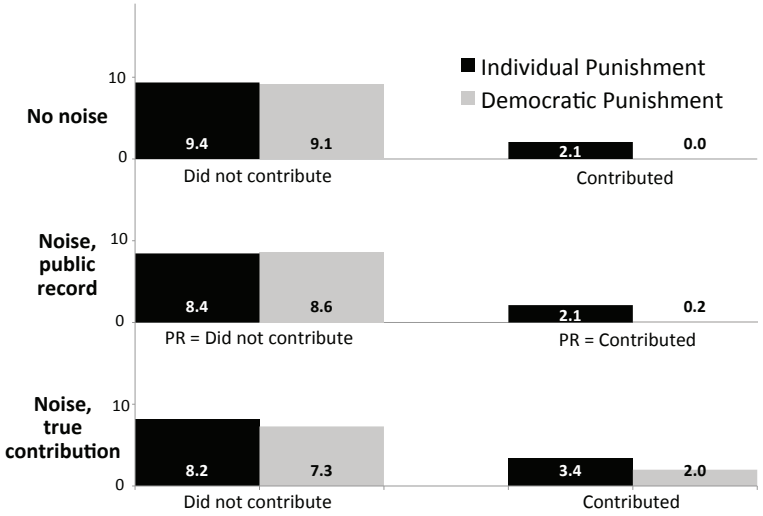
Note: All tests rely on averages at independent group level. For comparisons across treatments (within a treatment) we employ Wilcoxon Ranksum tests (Wilcoxon Matched Pairs Signed Ranks tests), respectively. *, **, and *** indicate significance at the 10%, 5%, and 1%-level, respectively, after applying an ex-post Bonferroni correction for repeated hypothesis tests, assuming each set of n=4 tests to be a test family (that is, dividing the required p-value for a level by 4).

Interestingly, under democratic punishment defectors are more likely to punish other defectors than other contributors, both when looking at votes as well as when looking at eventual outcomes. While the latter observation may be caused by majorities of cooperators dragging defectors along to punish another defector, the former result indicates that this seems not to be the case: defectors also intend to punish other defectors more than cooperators).

Across treatments, we observe a higher likelihood to vote for punishment in democratic decisions compared to the willingness to individually punish in the same situation (except for punishment of contributors towards contributors). These differences, however, are statistically only significant for the nominally largest differences, towards defectors under *Noise*, and are not significant at a reasonable level in the other conditions. In the *Democratic Punishment* conditions, we observe a drop in punishment frequency from votes to eventual punishments, indicating that often some group members wanted to punish but did not reach the required majority. This drop is significant across all types of punishment interactions for the *Noise* treatment, but only for punishment towards contributors (where it literally dropped down to zero) in the *No Noise* condition. As a result, as inspection of Figure 4 reveals, the *eventual* punishments in the different cases do not differ that much anymore between *Individual* and *Democratic Punishment* treatments.

The major statistically significant effect from introducing *Democratic Punishment* on punishment patterns is that cooperators are effectively not punished anymore (difference highly significant under *No noise*, and significant for punishment from defectors under *Noise*). Figure 5 visualizes these consequences. It displays the resulting average number of received punishment points conditional on whether (the public record indicated that) the participant has contributed in this round or not. Participants who did not contribute were deducted an average of 9.4 points (8.4 points) when there was *No Noise* (*Noise*), and this changed only slightly to 9.1 points (8.6 points) when employing *Democratic Punishment*. But for punishment towards contributors, introducing the voting procedure resulted in a drop from an average of 2.1 points (2.1 points) to literally 0 points (0.2 points)

FIGURE 5: Average punishment points deducted, conditional on punished subject's (true) contribution and public record



when there was *No Noise* (*Noise*). In the *Noise* treatment, this drop did not benefit all contributors, since some of them were burdened with a "no contribution" public record, such that the real expected punishment of a contributor decreased from 3.4 points on average in the noisy *Individual punishment* treatment to 2.0 points in the noisy *Democratic punishment* treatment.

III.C Reactions to received punishment

In Table 5 we report results from Probit regressions that explore how participants' contribution behavior responds to punishment received in the previous round. In Model 1 of Table 5, we regress the current contribution of a participant on the number of punishment points that were deducted from his income in the previous round ($RecPnmt_{PR}$). We control for whether the participant contributed in the previous round or not ($Contr_{PR}$), and interact previous punishment and previous contribution with each other as well as with treatment dummies that indicate whether noise was present (*Noise*), whether groups voted rather than assigned punishment individually (*Demo-*

TABLE 5: PROBIT ESTIMATIONS OF CURRENT CONTRIBUTION BASED ON PREVIOUS ROUND BEHAVIOR

	Model 1	Model 2
$RecPnmt_{PR}$	0.008***	0.006***
	[0.002]	[0.002]
$RecPnmt_{PR}$ × Noise	-0.001	
	[0.002]	
$RecPnmt_{PR}$ × DemocraticPun	0.002	0.001
	[0.002]	[0.002]
$RecPnmt_{PR}$ × Noise × DemocraticPun	-0.002	
	[0.003]	
$Contr_{PR}$	0.577***	0.458***
	[0.034]	[0.040]
$Contr_{PR}$ × $RecPnmt_{PR}$	-0.023***	-0.016***
	[0.006]	[0.005]
$Contr_{PR}$ × $RecPnmt_{PR}$ × Noise	0.008	
	[0.007]	
$Contr_{PR}$ × $RecPnmt_{PR}$ × DemocraticPun	0.012***	0.006
	[0.004]	[0.008]
$Contr_{PR}$ × $PRwrong_{PR}$		0.029
		[0.072]
$Contr_{PR}$ × $PRwrong_{PR}$ × $RecPnmt_{PR}$		0.014**
		[0.006]
$Contr_{PR}$ × $PRwrong_{PR}$ × $RecPnmt_{PR}$ x VotePun		-0.004
		[0.009]
N	6175	3230
Pseudo R-squared	0.223	0.148

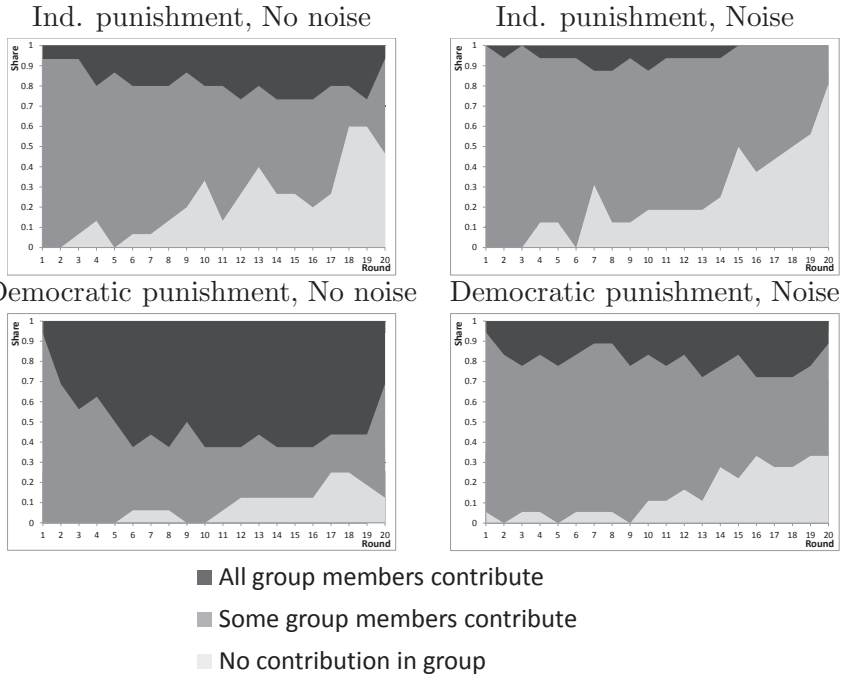
Note: We report marginal effects rather than coefficients. Robust standard errors, clustered at group level, are given in brackets. *, **, and *** indicate significance at the 10%, 5%, and 1%-level, respectively. $Contr_{PR}$ and $RecPnmt_{PR}$ refer to contribution and punishment received in the previous round, respectively, while $PRwrong_{PR}$ indicates whether the public record of a contributor in the previous round was wrong. *Noise* and *DemocraticPun* are dummies indicating whether noise was present or democratic punishment was employed, respectively.

craticPun), or both (*Noise* × *VotePun*).[7] The second model in Table 5 only looks at choices in the two *Noise* treatments and analyzes whether having received a wrong public record in the previous round (dummy $PRwrong_{PR}$, indicating that the public record displayed that participant hasn't contributed even though he did) changes next-round reactions to received punishment.

[7]Since the only punishment of contributors with voting happens when there is *Noise* (and never when there is *No Noise*), we have a problem of perfect collinearity in that condition, and therefore do not include the variable "$Contr_{PR}$ × $RecPnmt_{PR}$ × Noise × VotePun" in our estimations.

Table 5 shows that participants who did not cooperate increase their next-round contribution for each punishment point they received. This effect is reversed when the participant cooperated and got punished (Model 1 post-estimation F-test of $RecPnmt_{PR} + Contr_{PR} \times RecPnmt_{PR} = 0$ rejected at $p = 0.006$), but this averse reaction of cooperators is not existent when punishment was the consequence a democratic vote (Model 1 post-estimation F-test of $RecPnmt_{PR} + Contr_{PR} \times RecPnmt_{PR} + Contr_{PR} \times RecPnmt_{PR} \times DemocraticPun = 0$ not rejected at $p = 0.652$), or when the punishment was received due to a wrong public record in the *Noise* treatments (Model 2 post-estimation F-test of $Contr_{PR} \times RecPnmt_{PR} + Contr_{PR} \times PRwrong_{PR} \times RecPnmt_{PR} = 0$ not rejected at $p = 0.570$).

FIGURE 6: Group cooperation over time and average punishment in different cooperation classes



III.D Evolution of cooperation in groups

We classify groups into whether all five group members cooperate, all five group members defect, or whether there is heterogenous behavior within a group. Figure 6 displays the emergence of such types of groups over the 20 rounds of the experiment, separately for each treatment. To statistically test for time effects on the frequency of particular group classifications, we use Probit regressions to predict the group type based on constant term and *Round* number, separately by treatment, with standard errors clustered at group level. We report the estimated marginal effect of *Round* and its significance level in the second panel of Table 6.[8] Further, to test for the dominance of a particular group classification at the beginning or the end of the experiment, we apply Wilcoxon Signed Ranks test to compare average frequency of the different group types in rounds 1 to 3 and in rounds 17 to 19, the results of which we also report in Table 6.

As the first panel in Table 6 statistically underlines, at the beginning of the experiment (in rounds 1 to 3) groups with heterogenous contribution behavior are significantly more frequent than groups where 'All members contribute' or where there is 'No contribution in group'. Generally, compared to results for three-person groups over 50 rounds in Ambrus and Greiner (2012), we observe less convergence to homogenous group types over time. That said, in all four treatments the estimated *Round* marginal effect on the likelihood of group type 'Some group members contribute' is negative. There is a statistically measurable trend in all treatments towards no-contribution groups, while a (weakly) significant increase in the likelihood of all-contribution groups is only observed with *Democratic Punishment* when there is *No Noise*.

In accordance with the detected trends, in *Individual Punishment* we observe a large share of no-cooperation groups towards the end of the experiment, both under *No Noise* and under *Noise*. However, the third panel of Table 6 shows that this apparent dominance is not statistically corroborated

[8]Non-parametric Wilcoxon Signed Ranks tests comparing frequency of group type in rounds 1-3 to rounds 17-19 yield literally the same statistical conclusions.

TABLE 6: Statistical evidence on effects of time on group classification

Group classification	No noise		Noise	
	Individual punishment	Democratic punishment	Individual punishment	Democratic punishment
P-values from Wilxcon Signed Ranks tests on frequency of group type in rounds 1-3				
No contribution vs. some contribute	0.000	0.000	0.000	0.000
Some contribute vs. all contribute	0.000	0.019	0.000	0.000
No contribution vs. all contribute	0.317	0.005	0.317	0.088
Marginal effect dy/dx of Round coefficient in predicting group classification				
No contribution in group	0.026***	0.013**	0.032***	0.019***
Some group members contribute	-0.032***	-0.022***	-0.029***	-0.023***
All group members contribute	0.006	0.029*	-0.003*	0.005
P-values from Wilxcon Signed Ranks tests on frequency of group type in rounds 17-19				
No contribution vs. some contribute	0.184	0.870	1.000	0.377
Some contribute vs. all contribute	0.352	0.058	0.001	0.183
No contribution vs. all contribute	0.260	0.095	0.001	0.843

Note: *, **, and *** indicate significance at the 10%, 5%, and 1%-level, respectively. Probit models take the form $Pr(grouptype = x) = \Phi(\alpha + \beta Round)$, and we report the average marginal effect of β in the second panel. Wilcoxon Signed Ranks tests rely on independent groups as observations.

in the *No Noise* environment, while under *Noise* no-contribution groups are not more frequent than heterogenous groups albeit full-contribution groups disappear completely. Under *Democratic Punishment* and *No Noise*, full-contribution groups end up to be (weakly significantly) more frequent than groups with some or no contributions at all. For *Democratic Punishment* with *Noise* no convergence to a particular group type can be statistically detected, at least over the course of 20 rounds.

IV Conclusion

In this paper we observed that democratic punishment, when punishment decisions in a group are decided by majority voting, facilitates more cooperation and higher payoffs than individual punishment. It achieves so by establishing a stronger connection between a member's contribution decision and whether the member gets punished, in particular by decreasing anti-social

punishment while keeping the same level of pro-social punishment. We also see some evidence that participation in democratic punishment makes punishment intentions themselves more pro-social. The findings suggest that social norms or institutions that help members of a group to coordinate punishment decisions, and make it contingent on majority approval, can be welfare enhancing, even without the ability to make future commitments for punishment. A direction for future research is investigating what voting rule for punishments is optimal for society's welfare, for different levels of noise in observations, although addressing this question would ideally require larger groups than in our study. Presumably the expected welfare in the group is non-monotonic in the strictness of the voting rule, since if the threshold for punishing is very low, outcomes might be similar to individual punishment, while if they are too high then it might become impossible for the group to agree upon punishing someone, resulting in a lot of free riding.

References

Aldeshev, G. and Zanarone, G. (2014), Endogenous enforcement institutions, Technical report, Working Paper, University of Namur.

Ambrus, A. and Greiner, B. (2012), 'Imperfect public monitoring with costly punishment: An experimental study', *American Economic Review* **102**(7), 3317–3332.

Andreoni, J. and Gee, L. (2012), 'Gun for hire: Does delegated enforcement crowd out peer punishment in giving to public goods?', *Journal of Public Economics* **96**, 1036–1046.

Aoyagi, M. and Fréchette, G. (2009), 'Collusion as public monitoring becomes noisy: Experimental evidence', *Journal of Economic Theory* **144**, 1135–1165.

Baldassarri, D. and Grossman, G. (2011), 'Centralized sanctioning and legitimate authority promote cooperation in humans', *Proceedings of the National Academy of Sciences of the USA* **108**(1-5).

Bó, P. D., Foster, A. and Putterman, L. (2010), 'Institutions and behavior; experimental evidence on the effects of democracy', *American Economic Review* **100**, 2205–2229.

Cinyabuguma, M., Page, T. and Putterman, L. (2006), 'Can second-order punishment deter perverse punishment?', *Experimental Economics* **9**, 265–279.

Ertan, A., Page, T. and Putterman, L. (2009), 'Who to punish? individual decisions and majority rule in mitigating the free rider problem', *European Economic Review* **53**, 495–511.

Fehr, E. and Fischbacher, U. (2004), 'Third-party punishment and social norms', *Evolution and human behavior* **25**, 63–87.

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

Fischbacher, U. (2007), 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics* **10**, 171–178.

Fischer, S., Grechenig, K. and Meier, N. (2013), Cooperation under punishment: Imperfect information destroys it and centralizing punishment does not help, Technical report, Working Paper, University of Bonn.

Frey, B. (1994), 'Direct democracy: Politico-economic lessons from swiss experience', *American Economic Review Papers & Proceedings* **84**, 338–342.

Frey, B., Benz, M. and Stutzer, A. (2004), 'Introducing procedural utility: Not only what, but also how matters', *Journal of Institutional and Theoretical Economics* **160**, 377–401.

Fudenberg, D., Rand, D. G. and Dreber, A. (2012), 'Slow to anger and fast to forget: Cooperation in an uncertain world', *American Economic Review* **102**, 720–749.

Gächter, S., Renner, E. and Sefton, M. (2008), 'The long-run benefits of punishment', *Science* **322**, 1510.

Grechenig, C., Nicklisch, A. and Thöni, C. (2010), 'Punishment despite reasonable doubt - a public goods experiment with sanctions under uncertainty', *Journal of Empirical Legal Studies* **7**, 847–867.

Greiner, B. (2004), An online recruitment system for economic experiments, *in* K. Kremer and V. Macho, eds, 'Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63', Göttingen: Ges. für Wiss. Datenverarbeitung, pp. 79–93.

Hauser, O., Nowak, M. and Rand, D. (2014), 'Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible', *Journal of Theoretical Biology* **360**, 163–171.

Herrmann, B., Thöni, C. and Gächter, S. (2008), 'Antisocial punishment across societies,', *Science* **319**, 1362–1367.

Kamei, K., Putterman, L. and Tyran, J. (forthcoming), 'State or nature? endogenous formal versus informal sanctions in the voluntary provision of public goods', *Experimental Economics* .

Leibbrandt, A. and López-Pérez, R. (2011), 'The dark side of altruistic third-party punishment', *Journal of Conflict Resolution* **55**, 761–784.

Leibbrandt, A. and López-Pérez, R. (2012), 'An exploration of third and second party punishment in ten simple games', *Journal of Economic Behavior and Organization* **84**, 753–766.

Markussen, T., Putterman, L. and Tyran, J. (2014), 'Self-organization for collective action: An experimental study on sanction regimes', *Review of Economic Studies* **81**, 301–324.

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

Pommerehne, W. and Weck-Hannemann, H. (1996), 'Tax rates, tax administration and income tax evasion in switzerland', *Public Choice* **88**, 161–170.

Sutter, M., Haigner, S. and Kocher, M. (2010), 'Choosing the carrot or the stick? - endogenous institutional choice in social dilemma situations', *Review of Economic Studies* **77**, 1540–1566.

Tyran, J. and Feld, L. (2006), 'Achieving compliance when legal sanctions are non-deterrent', *Scandinavian Journal of Economics* **108**, 135–156.