Revised October 22, 2011

Group Size, Coordination, and the Effectiveness of the Punishment Mechanism in the VCM: An Experimental Investigation

Bin Xu (Public Administration College, Zhejiang Gongshang University and Experimental Social Science Laboratory, Zhejiang University)*

Bram Cadsby (Department of Economics, University of Guelph)

Liangcong Fan (College of Public Administration, Zhejiang University)

Fei Song (Ted Rogers School of Management, Ryerson University)

Abstract

In this study, we examine the effectiveness of the individual-punishment mechanism in larger groups, comparing groups of four to groups of 40 participants. We find that the individual punishment mechanism is remarkably robust when the MPCR is held constant despite the coordination problems inherent in an institution relying on decentralized individual punishment decisions in the context of a larger group. This reflects increased per-capita expenditures on punishment that offset the greater coordination difficulties in the larger group. However, if the marginal group return stays constant, resulting in an MPCR that shrinks with group size, no such offset occurs and punishment loses much but not all of its effectiveness at encouraging voluntary contributions to a public good.

*Bin Xu is the holder of the grant from the Social Science Experimental Center of Zhejiang University that funded this project. All authors contributed equally to the study. We thank Qiqi Cheng, Lu Liu, Chao Wang, Tongyu Wu, and Xinchao Zhang for their excellent research assistance, and Zhiwei Fang and Yanmin Qian for their support for the project.

Group Size, Coordination, and the Effectiveness of the Punishment Mechanism in the VCM: An Experimental Investigation

I. Introduction

The voluntary contribution mechanism (VCM) has been an important topic of research in experimental economics. Among the many issues addressed by laboratory experiments is the relationship between group size and the level of contributions. Isaac, Walker and Williams (1994) examined group sizes from four to 100, while simultaneously manipulating the marginal per-capita return (MPCR) between 0.04 and 0.75. Their main results show that with the MPCR held constant at 0.3, groups of 40 and 100 provide the public good at higher levels of efficiency than groups of 4 and 10 respectively. However, for an MPCR of 0.75, group size had no significant effect on public good provision.

In a separate line of research, Fehr and Gächter (2000; 2002) demonstrated that informing individual contributors of the contributions made by their peers, and then permitting those contributors to purchase punishments directed at individuals they specify is a remarkably effective means of motivating high contributions among groups of four participants. This is true under both partner and stranger designs. This result is especially noteworthy because the availability of these punishment opportunities does not alter the fact that complete free riding in contributions is still the unique stage-game Nash Equilibrium for the VCM with or without punishment opportunities.

A number of studies have examined the robustness of Fehr and Gächter's results with respect to punishment effectiveness and cost (Egas and Riedl, 2008; Nikiforakis and

Normann, 2008; Gardner and West, 2004), communication (Bochet, Page, and Putterman, 2006), self-selection of punishment versus non-punishment institution (Gürerk, Irlenbusch, and Rockenbach, 2006), monetary versus non-monetary punishment (Masclet et al., 2003), length of the game (Gächter, Renner, and Sefton, 2008), alternative punishment institutions (Casari and Luini, 2009), and country (Herrmann, Thöni, and Gächter, 2008).[1] Carpenter (2007) compares groups consisting of five versus ten participants. He also controls for the extent to which subjects can monitor each other. His results show that the availability of punishment promotes contributions for both groups of five and groups of ten, but that restrictions on monitoring can adversely affect contributions.

The effectiveness of the individual punishment mechanism in laboratory groups of four, five or ten provides a persuasive explanation of how free-riding behavior can be mitigated in relatively small groups that need to mobilize contributions of money or effort towards a common public good. However, it is uncertain whether such a mechanism would continue to be effective in the much larger groups that must often cooperate together for the common good. Carpenter (2007) finds that in ten-person groups there is some evidence that individuals punish less because of a bystander effect, i.e second-order free riding in bearing the cost of punishment. He finds however that this is largely offset by the presence of more potential punishers. Casari (2005) notes that Carpenter's design employs a punishment mechanism with a fine-to-fee ratio that increases with group size. As Casari points out, a higher fine-to-fee ratio has been associated with increased expenditures on punishment (Anderson and Putterman, 2006;

[1] Related literatures examine rewards versus punishments (e.g., Rand et al., 2009) and the evolutionary emergence of punishment (e.g., Boyd, Gintis and Bowles, 2010).

Andreoni, Harbaugh, and Vesterlund, 2003; Carpenter, 2002; Egas and Riedl, 2008; Nikiforakis and Normann, 2008; Gardner and West, 2004; Ostrom, Walker, and Gardner, 1992). This could have motivated more punishment expenditures in Carpenter's ten-person than in his five-person groups, mitigating the potential coordination problem in the ten-person groups.

As group size increases, potential coordination problems in the individual punishment mechanism multiply if each subject trying to decide whether or not to punish a low contributor is unable to observe which of those low contributors may be simultaneously receiving punishments from others. The primary objective of our study is thus to examine the robustness of the individual-punishment mechanism at a constant fine-to-fee ratio in the context of the potential coordination problems that may occur in larger groups. In particular, we follow Isaac, Walker and Williams (1994) in comparing groups of four versus 40 participants. In our four person-groups, the MPCR was set at 0.4. This implies a marginal group return (MGR) of $0.4 \cdot 4 = 1.6$, i.e. each contribution of one token results in 1.6 tokens divided equally among the four-person group. In half of our 40-person groups, we held the MPCR constant at 0.4, resulting in a MGR of $0.4 \cdot 40 = 16$, i.e. each contribution of one token creates 16 tokens divided equally among the 40-person group. In the other half of our 40-person groups, we held the MGR constant at 1.6, resulting in a reduced MPCR of just 0.04. Of course, we would expect the high-MPCR group to contribute more to the public good than the low-MPCR group with or without punishment. We also hypothesize that punishment will be more effective at raising contributions in the high- than in the low-MPCR group. This is because there is greater motivation to punish low contributors when their increased contributions would have a

greater effect on one's earnings.

It is less clear how an increase in group size with a constant MPCR would influence the effectiveness of the individual punishment mechanism. On the one hand, the increase in MGR might be expected to encourage the punishment of low contributors by those who care about the larger potential social surplus. On the other hand, the coordination problem described above may cause free-riding to take hold if some low contributors are not initially punished.

II. Experimental Design

Our specific experimental design adopted key elements from Fehr and Gächter's two important studies (2000; 2002). Like them, we employed a within-person design of punishment (P) versus non-punishment (N) conditions. In particular, each subject played ten rounds of N and ten rounds of P in a session. The order of P and N was reversed for half of the sessions. Following Fehr and Gächter (2000; 2002), we initially told the participants that they would be playing ten rounds in either the P or N condition. Afterwards, they were informed that they would be playing ten more rounds in a new experiment, and that the session would finish after this second set of ten rounds was played. We used a partner protocol both because of the practical difficulties of using a stranger design with 40-person groups and in order to focus on large groups that may have repeated opportunities for cooperation. We used scrambled IDs from round to round so that no reputation could be built over time. The fine-to-fee ratio was set at 3:1 as in Fehr and Gächter (2002). Thus, spending one token to punish another person resulted in a three-token loss for that person. This ratio did not vary with either group size or

MPCR.

Each subject was endowed with 20 tokens for each round and another 25 (500) tokens at the beginning of the P condition when group size was equal to 4 (40). The exchange rate was set at 21 Tokens = 1 RMB for group size = 4 and 39.23 (150) Tokens = 1 RMB for group size = 40 with MPCR = 0.04 (0.4). These exchange rates were calculated by holding the mean of the free-riding payoff and the full-contribution payoff plus the 25 (500) tokens for the four- (40-) person P condition equal in RMB between these treatments. Lastly, each subject was also given a 10 RMB show-up fee. As in Fehr and Gächter (2000; 2002), a subject who did not punish others could not lose money. Punishment points received could not reduce income from the contribution stage of the game to less than zero. However, spending money on punishing others created the possibility of losing money. For example, if one received enough punishment points to reduce one's earnings from the contribution stage to zero, any punishment points previously purchased would result in a loss. The purpose of the 25 (500) tokens received at the beginning of the P rounds for the four- (40-) person groups was to mitigate the possibility of somebody leaving the session owing the experimenter money.[2] This did not occur.

In summary, there are three independent variables: group size (four versus 40), MPCR (0.04 and 0.4), and decision order (NP and PN). MGR is the product of group size and MPCR. Since the MPCR of 0.04 could only be used for 40-person groups, there were six treatments in total:

[2] This design feature was borrowed from Fehr and Gächter (2000, 2002). See the experimental instructions associated with each paper for details.

1. Small group (4), PN, MPCR = 0.4, MGR = 1.6, ten groups
2. Large group (40), PN, MPCR = 0.4, MGR = 16, three groups
3. Large group (40), PN, MPCR = 0.04, MGR = 1.6, three groups
4. Small group (4), NP, MPCR = 0.4, MGR = 1.6, ten groups
5. Large group (40), NP, MPCR = 0.4, MGR = 16, three groups
6. Large group (40), NP, MPCR = 0.04, MGR = 1.6, three groups

Subjects were randomly recruited via online advertisements at Zhejiang University in Hangzhou, China. All subjects were full-time undergraduate students in diverse majors across the Sciences, Social Sciences, and Humanities. A total of 560 subjects participated in the study. All sessions were run at the Zhejiang University Experimental Social Science Laboratory.

All sessions were computerized.[3] Upon arrival, each subject was seated at a private computer carrel. Each session lasted about 100 minutes. The average earnings for each subject were approximately 39.6 RMB including a 10 RMB show-up fee. At the time of the experiment, 39.6 RMB was equal to about $5.82 US. For comparison purposes, the wage rate for Zhejiang University undergraduates who had part-time jobs with the university administration was 12 RMB per hour.

III. Results

Table 1 presents a data summary by treatment of the sum of contributions per capita in the punishment rounds, in the non-punishment rounds and the difference between them. In all cases, the differences between contributions in the P condition and

[3] Zhijian Wang and Bin Xu jointly designed, tested and implemented the computer program used in this experiment.

contributions in the N condition are positive and significantly different from zero as indicated by the *p*-values from simple *t*-tests using group-level data, which are in parentheses below the calculated differences.

Table 2 presents regression results using individual data with random effects for each group. The dependent variable is the difference between contributions over all ten rounds of the P condition and contributions over all ten rounds of the N condition for each individual participant. The independent variables are all dummy variables representing the different treatments. Coding the dummy variables in a variety of ways permits us to run different hypothesis tests of interest. Thus, we report the results of this regression with six different permutations of dummy-variable codings. Each coding uses a different treatment as the reference group, represented by the constant term. The coding numbers at the top of each column correspond to the treatment numbers in Table 1, indicating the reference group for each coding. To correctly interpret the meaning of the independent dummy variables, recall that both high MGR and low MPCR occur only when the group size is large.

The first thing to notice is that all of the constant terms are significant with *p*-values less than 0.01. This confirms the results of the *t*-tests presented in Table 1. Punishment made a significant difference to contributions in all six treatments. Second, in the NP order, the effectiveness of punishment at increasing contributions is significantly lower in the low-MPCR than in the high-MPCR large-group treatment ($p = 0.044$). Third, in the NP order, the effectiveness of punishment at increasing contributions is also significantly lower in the low-MPCR large-group treatment than in the small-group treatment ($p = 0.009$).). Fourth, there is no significant difference in the

effectiveness of punishment related to group size for a constant high MPCR in the NP order. Fifth, there is a significant order effect in the small-group treatment with punishment being less effective in the PN order ($p = 0.007$). Sixth, nothing is significant in the PN order.

It may take time for participants to adjust to the change of condition. Thus, it is interesting to examine the analogous results for the last round under each condition. Table 3 reports these results. The constant terms are significant for both the small-group ($p = 0.000$ for both NP and PN orders) and high-MPCR large-group ($p = 0.000$ for NP order and $p = 0.001$ for PN order) treatments, indicating that punishment makes a significant difference in these cases. However, in contrast to the ten-round average data, the constant terms are not significant for the low-MPCR large-group treatment. Thus, we cannot reject the null hypothesis that punishment makes no difference to the level of contributions when the MPCR is low. In the NP order, the effectiveness of punishment at increasing contributions is significantly lower in the low-MPCR large-group treatment than in the small-group treatment ($p = 0.005$) and lower but with just marginal significance in comparison with the high-MPCR large-group treatment ($p = 0.070$). In the PN order, there is a significant difference in the effect of punishment only between the low-MPCR large-group and small-group treatments ($p = 0.052$). The effectiveness of punishment is not significantly influenced by group size for a constant high MPCR in either the NP or PN order.

In contrast to the ten-round average data, none of the order effects or interactions involving order effects is individually significant for the last-round data. Moreover, a joint test that the coefficients on the main order effect together with those on its

interactions with the two other treatment dummies all equal zero yields a Chi-Square statistic of 1.95 with three degrees of freedom ($p = 0.583$). This suggests that the observed differences between the effectiveness of punishment in the NP versus the PN order have to do with the transition from N to P relative to the transition from P to N, and vanish by the tenth repetition within the N or P condition. Dropping the order effects, we can aggregate the NP and PN data and re-estimate the regressions using the aggregated data. The results are reported in Table 4. The constant terms continue to be significant for the small-group and high-MPCR large-group cases ($p = 0.000$ in both cases). For the low-MPCR large-group treatment, the constant term now attains marginal significance ($p = 0.079$), yielding some weak evidence that punishment has an effect on contributions even in this case. However, the effectiveness of the punishment condition at increasing contributions is significantly lower in the low-MPCR large-group treatment than in either the small-group ($p = 0.001$) or the high-MPCR large-group ($p = 0.017$) treatments by the last round of each condition. Once again, group size has no significant effect for a constant high MPCR.

Is the punishment condition less effective in the low-MPCR large-group treatment simply because fewer punishments are purchased when the potential gains from further contributions are relatively small? The last column of Table 1 presents per-capita expenditures on punishment for each treatment. In both orders, such expenditures appear to be substantially higher in the high-MPCR large-group (high MGR) treatment than in the other two treatments. To investigate this issue further, we regress per-capita expenditures on punishment in each session aggregated over all ten punishment rounds on the same dummy variables representing the different treatments as used above. The

results are presented in Table 5 for the different dummy-variable codings. None of the order effects or their interactions with the treatment dummy variables is significant. While per-capita punishment expenditures in the high-MPCR large group treatment are significantly higher than in both the small group treatment ($p = 0.000$ and $p = 0.001$ for the NP and PN orders respectively) and the low-MPCR large group treatment ($p = 0.001$ and $p = 0.023$ for the NP and PN orders respectively), there is no significant difference in per-capita punishment expenditures between the small-group and the low-MPCR large group treatments for either order. A joint test that the coefficients on the main order effect together with those on its interactions with the two other treatment dummies all equal zero yields an $F(3, 26)$ statistic of 0.80 ($p = 0.5057$). Dropping these order effects leads to qualitatively identical inferences.[4]

While per-capita expenditures on punishment are significantly higher in the high-MPCR large-group treatment than in the other two treatments, the effectiveness of the punishment condition at increasing contributions is significantly higher in both the high-MPCR large-group treatment and the small-group treatment than in the low-MPCR large-group treatment. Thus, statistically indistinguishable levels of per-capita spending on punishment are significantly more effective at increasing contributions in the small-group treatment than in the low-MPCR large group treatment. Moreover, significantly higher levels of per-capita spending in the high-MPCR treatment relative to the small-group treatment produce increases in contributions that are statistically indistinguishable from each other. We hypothesize that this reflects a coordination problem that afflicts the decentralized punishment mechanism in large groups, making per-capita expenditures on

[4] To conserve space, these results are not reported in detail here. They are available from the authors upon request.

punishment less effective at increasing contributions in such groups.

Suppose for example that 25% of participants are low contributors. In a group of four, this implies that there is just one low contributor and three higher contributors who might decide to punish him or her. Suppose that each high contributor purchases one punishment point. The low contributor will receive three punishment points, perhaps an inducement to contribute more in the next round. In an analogous group of 40, there would be ten low contributors and thirty higher contributors who might decide to punish one or more of the ten low contributors. If each high contributor purchases one punishment point, the ten low contributors will together receive thirty punishment points, an average of three per person. It is possible that these thirty punishment points will be divided equally among the ten low contributors. In that case, each low contributor will receive three punishment points just as in the small four-person group. However, there is no mechanism to coordinate the distribution of punishment points among the low contributors. Therefore, it is unlikely that they will be distributed equally. Instead it is probable that some low contributors will receive more punishment points than necessary to motivate higher contributions, while others will receive fewer or none at all.

Table 6 presents summary data on the proportion of "low" contributors that received at least one punishment point for each treatment. We use two definitions of a low contributor. The first is a relative definition. It defines a contributor to be low if his/her contribution is at or below the 25^{th} percentile in a round and s/he is not one of the highest contributors in that round. The second is primarily an absolute definition. It defines those contributing ten or fewer tokens as low contributors as long as they are not among the highest contributors in the round. According to both definitions, the proportion

of low contributors receiving at least one punishment point was substantially lower in the low-MPCR large group treatment than in either of the other two treatments in both the NP and PN orders.

To determine whether there is a significant difference in the likelihood of a low contributor being punished in the low-MPCR large group treatment than in the other two treatments, we employed a negative binomial regression for each definition of a low contributor. For each group of participants, we have one count of the number of times a low contributor received at least one punishment aggregated across all rounds. This is the dependent variable. In addition, we calculate the number of times a low contribution occurred aggregated across all rounds, the log of which is used as the exposure variable.[5] To facilitate interpretation, the coefficients are reported in the form of incidence rate ratios (IRRs).

Table 7 presents the results for the relative definition. Consider the reported IRR for mpcr_low in regression 1, which is 0.593. This means that the estimated rate at which low contributors received at least one punishment in the low-MPCR large-group treatment was 59.3% as high as the analogous rate in the small-group treatment for the NP order. Since the *p*-value is 0.007, this is a significant difference. Similarly, in regression 2 the IRR for mpcr_low indicates that the rate at which low contributors received at least one punishment in the low-MPCR large group treatment was 46.5% as high as the analogous rate in the high-MPCR large-group treatment ($p = 0.000$). Notice

[5] The exposure variable adjusts for the differing numbers of low contributions in each group. The proportions for each treatment presented in Table 6 are averages across such proportions calculated for each group in a treatment. The numerator of each such group proportion is the count of the number of times a low contributor received at least one punishment, while the denominator is the number of times a low contribution occurred aggregated across all rounds.

that the IRR in regression 3 for mpcr_high contains the same information, indicating that the incidence rate in the high-MPCR large-group treatment is 214.9% as high as the rate in the low-MPCR large group treatment, where 214.9% is the reciprocal of 46.5%. For the PN order, the incidence rate for the low-MPCR large-group treatment was 70.6% of the rate for the small-group treatment with marginal significance ($p = 0.066$), while the rate for the low-MPCR large-group treatment was 69.4% of the rate for the high-MPCR large-group treatment ($p = 0.053$). There is no significant difference between the incidence rates for the small-group versus the high-MPCR large-group treatment in either the NP or PN order. Moreover, there are no significant order effects. A joint test of the null hypothesis that the order effect and its interactions with the treatment variables all equal zero yields a chi-square statistic of 3.41 with three degrees of freedom ($p = 0.332$). Thus, the null hypothesis of no order effects or interactions involving order effects cannot be rejected. Dropping these order effects and reestimating this negative binomial regression leads to the likelihood of low contributors receiving at least one punishment being significantly lower in the low-MPCR large-group treatment than in either the small-group ($p = 0.003$) or the high-MPCR large-group ($p = 0.000$) treatment. As before, there is no significant difference between the incidence rates for the small-group versus the high-MPCR large-group treatment ($p = 0.370$).[6]

Table 8 presents the results for the primarily absolute definition of low contributor. There are marginally significant order effects for the high-MPCR large group treatments ($p = 0.081$) and a significant interaction between the effect of MPCR and order ($p = 0.048$). However, the treatment effects are robust to the altered definition of

[6] The detailed results are not reported in order to conserve space. They are available from the authors upon request.

low contributor. The incidence rate for the low-MPCR large group treatment is significantly lower than for the small-group treatment ($p = 0.000$ for both the NP and PN orders) and significantly lower than for the high-MPCR large group treatment ($p = 0.000$ for the NP and $p = 0.003$ for the PN order). There is no significant difference between the incidence rates for the small-group versus the high-MPCR large group treatment in either order.

These results together corroborate the coordination hypothesis, supporting the idea that a given per-capita expenditure on decentralized individual punishments is more efficient at increasing contributions for smaller than for larger groups. In small groups, for a given level of per-capita expenditures, a higher proportion of low contributors receive at least one punishment than in large groups. This is the reason that statistically indistinguishable amounts of expenditure on punishment are significantly more effective in the small-group treatment than in the low-MPCR large-group treatment at increasing contributions. It is also the reason that the significantly higher expenditures on punishment observed in the high-MPCR large-group treatment relative to the small-group treatment are necessary to produce similar increases in contributions that are statistically indistinguishable from each other.

IV. Conclusion

The effectiveness of the individual punishment mechanism at increasing contributions to a public good depends critically on what happens to the MPCR of a public good as the potential community of contributors grows. For a pure public good with non-rivalry in consumption, MPCR stays constant and MGR increases

proportionally with the size of the community. In this paper, we have demonstrated that the higher MGR produces a significant increase in per-capita expenditures on punishment in 40-person relative to four-person groups. At the same time, the larger group creates a coordination problem for the decentralized punishment mechanism, making each dollar spent on punishment less effective at increasing contributions. This occurs because some punishment dollars are inevitably wasted on low contributors who are simultaneously punished sufficiently to increase their contributions by other purchasers of punishment points, while other low contributors escape punishment. In this experimental study, the increase in punishment expenditures was sufficient to offset the reduction in the efficiency of each punishment dollar. Thus, for a constant MPCR, the individual punishment mechanism proved remarkably robust despite the coordination problems inherent in an institution relying on decentralized individual punishment decisions in the context of a larger group.

However, if the MGR stays constant, resulting in an MPCR that shrinks with group size, per-capita expenditures on punishment do not increase. In this case, the coordination problem associated with the 40-person group is not offset by increases in punishment expenditures. This results in the individual punishment mechanism being significantly less effective at increasing contributions for a 40-person than for a four-person community with the same MGR. Examining institutional modifications to mitigate the coordination problem associated with the decentralized individual punishment mechanism is an important issue deserving further study.

References

Anderson, C., and Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Journal of Economic Behavior and Organization*, 54: 1-24.

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, 93: 893-903.

Bochet, O., Page, T., and Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization*, 60: 11-26.

Boyd, R., Gintis, H., and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328: 617-620.

Carpenter, J. (2002). The demand for punishment. Working Paper 0243, Middlebury College, Department of Economics.

Carpenter, J. P. (2007). Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, 60: 31-51.

Casari, M. (2005). On the design of peer punishment experiments. *Experimental Economics*, 8: 107-115.

Casari, M., and Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior and Organization*, 71: 273-282.

Egas, M., and Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 275: 871-878.

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

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature,* 415: 137-140.

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

Gardner, A., and West, S. A. (2004). Cooperation and punishment, especially in humans. *American Naturalist*, 164: 753-764.

Gürerk, O., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312: 108-111.

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

Isaac, R. M., Walker, J. M., and Williams, A. W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of Public Economics,* 54: 1-36.

Masclet, D., Noussair, C., Tucker, S., and Villeval, M. (2003). Monetary and non-monetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93: 366-380.

Nikiforakis, N., Normann, H. (2008). A comparative analysis of punishment in public-good experiments. *Experimental Economics*, 11: 358-369.

Ostrom, E., Walker, J. and Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Economic Review*, 86: 404-417.

Rand, D., Dreber, A., Ellingsen, T., Fudenburg, D., and Nowak, M. (2009). Positive interactions promote public cooperation. *Science*, 325: 1272-1275.

Table 1: Data Summary (*t*-test *p*-values in parentheses)

Treatment	Sample Size	P_N difference per capita	P contribution per capita	N contribution per capita	Punishments per capita
1: Small, NP, MPCR=0.4	10	69.825 (0.000)	124.40	54.58	10.6
2: Large, NP, MPCR=0.4	3	65.15 (0.016)	175.83	110.68	32.892
3: Large, NP, MPCR=0.04	3	32.608 (0.016)	63.60	30.99	8.467
4: Small, PN, MPCR=0.4	10	38.075 (0.001)	125.18	87.10	6.075
5: Large, PN, MPCR=0.4	3	43.767 (0.044)	126.54	82.78	26.3
6: Large, PN, MPCR=0.04	3	36.408 (0.013)	54.10	17.70	9.675

Table 2: Regression Results on Ten-Round Per-Capita Differences in Contributions between the Punishment and No-Punishment Conditions (*p*-values in parentheses)

	Reg. 1	Reg. 2	Reg. 3	Reg. 4	Reg. 5	Reg. 6
Constant	**69.825**	**65.15**	**32.61**	**38.075**	**43.77**	**36.41**
	(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
MPCR_Low	**-37.22**	**-32.54**		-1.67	-7.36	
	(0.009)	**(0.044)**		(0.906)	(0.648)	
MPCR_High			**32.54**			7.36
			(0.044)			(0.648)
MGR_High	-4.68			5.69		
	(0.741)			(0.688)		
Small		4.68	4.68		-5.69	-5.69
		(0.741)	(0.741)		(0.688)	(0.688)
Order_PN	**-31.75**	-21.38	3.80			
	(0.007)	(0.185)	(0.814)			
Order_NP				**31.75**	21.38	-3.80
				(0.007)	(0.185)	(0.814)
MPCRL x Order	35.55	25.18		-35.55	-25.18	
	(0.076)	(0.269)		(0.076)	(0.269)	
MPCRH x Order			-25.18			25.18
			(0.269)			(0.269)
MGRH x Order	10.37			-10.37		
	(0.604)			(0.604)		
Small x Order		-10.37	10.37		10.37	10.37
		(0.604)	(0.604)		(0.604)	(0.604)
Observations	560	560	560	560	560	560
Number of groups	32	32	32	32	32	32
R squared	0.109	0.109	0.109	0.109	0.109	0.109

Table 3: Regression Results on Last-Round Per-Capita Differences in Contributions between the Punishment and No-Punishment Conditions (*p*-values in parentheses)

	Reg. 1	Reg. 2	Reg. 3	Reg. 4	Reg. 5	Reg. 6
Constant	**12.275**	**10.100**	3.442	**9.000**	**8.375**	2.892
	(0.000)	**(0.000)**	(0.186)	**(0.000)**	**(0.001)**	(0.266)
MPCR_Low	**-8.833**	-6.658		**-6.108**	-5.483	
	(0.005)	(0.070)		**(0.052)**	(0.136)	
MPCR_High			6.658			5.483
			(0.070)			(0.136)
MGR_High	-2.175			-0.625		
	(0.490)			(0.843)		
Small		2.175	2.175		0.625	0.625
		(0.490)	(0.490)		(0.843)	(0.843)
Order_PN	-3.275	-1.725	-0.550			
	(0.191)	(0.639)	(0.881)			
Order_NP				3.275	1.725	0.550
				(0.191)	(0.639)	(0.881)
MPCRL X Order	2.725	1.175		-2.725	-1.175	
	(0.540)	(0.821)		(0.540)	(0.821)	
MPCRH X Order			-1.175			1.175
			(0.821)			(0.821)
MGR X Order	1.550			-1.550		
	(0.728)			(0.728)		
Small X Order		-1.550	-1.550		1.550	1.550
		(0.728)	(0.728)		(0.728)	(0.728)
Observations	560	560	560	560	560	560
Number of groups	32	32	32	32	32	32
R squared	0.164	0.164	0.164	0.164	0.164	0.164

Table 4: Regression Results on Last-Round Per-Capita Differences in Contributions between the Punishment and No-Punishment Conditions Dropping Insignificant Order Effects (*p*-values in parentheses)

	Reg. 1	Reg. 2	Reg. 3
Constant	**10.638**	**-9.238**	3.167
	(0.000)	**(0.000)**	(0.079)
MPCR_Low	**-7.47**	**-6.07**	
	(0.001)	**(0.017)**	
MPCR_High			**6.07**
			(0.017)
MGR_High	-1.40		
	(0.522)		
Small		1.40	1.40
		(0.522)	(0.522)
Observations	560	560	560
Number of groups	32	32	32
R squared	0.154	0.154	0.154

Table 5: Regression Results on Ten-Round Per-Capita Expenditures on Punishment (*p*-values in parentheses)

	Reg. 1	Reg. 2	Reg. 3	Reg. 4	Reg. 5	Reg. 6
Constant	**10.600**	**32.892**	8.467	**6.075**	**26.300**	9.675
	(0.000)	**(0.000)**	(0.093)	**(0.031)**	**(0.000)**	(0.057)
MPCR_Low	-2.133	**-24.425**		3.600	**-16.625**	
	(0.703)	**(0.001)**		(0.521)	**(0.023)**	
MPCR_High			**24.425**			**16.625**
			(0.001)			**(0.023)**
MGR_High	**22.292**			**20.225**		
	(0.000)			**(0.001)**		
Small		**-22.292**	**-22.292**		**-20.225**	**-20.225**
		(0.000)	**(0.000)**		**(0.001)**	**(0.001)**
Order_NP				4.525	6.592	-1.208
				(0.240)	(0.346)	(0.862)
Order_PN	-4.525	-6.592	1.208			
	(0.240)	(0.346)	(0.862)			
MPCRL X Order	5.733	7.800		-5.733	-7.800	
	(0.471)	(0.429)		(0.471)	(0.429)	
MPCRH X Order			-7.800			7.800
			(0.429)			(0.429)
MGR X Order	-2.067			2.067		
	(0.794)			(0.794)		
Small X Order		2.067	2.067		-2.067	-2.067
		(0.794)	(0.794)		(0.794)	(0.794)
Observations	32	32	32	32	32	32
Adjusted R squared	0.475	0.475	0.475	0.475	0.475	0.475

Table 6: Proportion of Low Contributors Punished averaged across Sessions by Treatment for Two Definitions of Low Contributor

Treatment	Sample Size	25th Percentile or Lower and not Highest in Round	Ten or Lower and not Highest in Round
1: Small, NP, MPCR=0.4	10	0.609	0.606
2: Large, NP, MPCR=0.4	3	0.761	0.741
3: Large, NP, MPCR=0.04	3	0.347	0.188
4: Small, PN, MPCR=0.4	10	0.674	0.538
5: Large, PN, MPCR=0.4	3	0.689	0.486
6: Large, PN, MPCR=0.04	3	0.476	0.238

Table 7: Negative Binomial Regression Results for the Proportion of Times People in the Lowest Contribution Quartile and were not among the Highest Contributors in a Round were Punished
(*p*-values in parentheses)

	Reg. 1	Reg. 2	Reg. 3	Reg. 4	Reg. 5	Reg. 6
MPCR_Low	**0.593**	**0.465**		0.706	**0.694**	
	(0.007)	**(0.000)**		(0.066)	**(0.053)**	
MPCR_High			**2.149**			**1.441**
			(0.000)			**(0.053)**
MGR_High	1.276			1.017		
	(0.202)			(0.929)		
Small		0.784	0.784		0.983	0.983
		(0.202)	(0.202)		(0.929)	(0.929)
Order_NP				0.869	1.090	0.731
				(0.462)	(0.646)	(0.103)
Order_PN	1.151	0.917	1.369			
	(0.462)	(0.646)	(0.103)			
MPCRL X Order	1.189	1.492		0.841	0.670	
	(0.522)	(0.137)		(0.522)	(0.137)	
MPCRH X Order			0.670			1.492
			(0.137)			(0.137)
MGR X Order	0.797			1.254		
	(0.397)			(0.397)		
Small X Order		1.254	1.254		0.797	0.797
		(0.397)	(0.397)		(0.397)	(0.397)
Observations	32	32	32	32	32	32
Pseudo R squared	0.061	0.061	0.061	0.061	0.061	0.061

Table 8: Negative Binomial Regression Results for the Proportion of Times People who Contributed Ten or Less and were not among the Highest Contributors in a Round were Punished (*p*-values in parentheses)

	Reg. 1	Reg. 2	Reg. 3	Reg. 4	Reg. 5	Reg. 6
MPCR_Low	**0.347**	**0.260**		**0.407**	**0.505**	
	(0.000)	**(0.000)**		**(0.000)**	**(0.003)**	
MPCR_High			**3.844**			**1.980**
			(0.000)			**(0.003)**
MGR_High	1.335			0.805		
	(0.228)			(0.319)		
Small		0.749	0.749		1.242	1.242
		(0.228)	(0.228)		(0.319)	(0.319)
Size_large						
Order_NP				0.928	1.538	0.792
				(0.720)	(0.081)	(0.304)
Order_PN	1.078	0.650	1.263			
	(0.720)	(0.081)	(0.304)			
MPCRL X Order	1.171	**1.942**		0.854	**0.515**	
	(0.608)	**(0.048)**		(0.608)	**(0.048)**	
MPCRH X Order			**0.515**			**1.942**
			(0.048)			**(0.048)**
MGR X Order	0.603			1.658		
	(0.119)			(0.119)		
Small X Order		1.658	1.658		0.603	0.603
		(0.119)	(0.119)		(0.119)	(0.119)
Observations	32	32	32	32	32	32
Pseudo R squared	0.104	0.104	0.104	0.104	0.104	0.104