On the (In)effectiveness of Collective Punishment: An Experimental Investigation

Eric S. Dickson*

Abstract

The effectiveness of collective sanctioning strategies has been debated by scholars and policymakers; advocates argue that such measures can efficiently impel populations to undertake desired actions while encouraging enforcement of in-group norms. This paper describes a laboratory experiment in which "Group Members" may choose to make costly contributions to a public good while an "Outside Authority" – who observes only aggregate group outcomes – can choose to collectively punish Group Members in an attempt to influence their behavior. In the "Opposed Interests" treatment (where Outside Authority agents wish to *minimize* public goods production), collective punishment is found to be strictly counterproductive – it *provokes more* public goods provision instead of inhibiting it. In the "Aligned Interests" treatment (where Outside Authority agents wish to *maximize* public goods production), collective punishment provokes somewhat more provision in the short run, but appears to suppress it in the longer run, largely cancelling any short-run gains. Further, in neither treatment does the use of collective punishment have any significant effect on Group Members' propensities to punish one another *internally*.

*Direct correspondence to: Eric S. Dickson, Assistant Professor, Department of Politics and Center for Experimental Social Science, New York University. Email: eric.dickson@nyu.edu. I thank Julie Browne for excellent research assistance.

1. Introduction

In a variety of political settings, actors may find it desirable to influence the behavior of "external" populations. A head of government who desires the overthrow of a foreign regime may wish to increase the likelihood that the regime's own population will rise up *en masse*. Another leader, worried about the threat of terrorism that may emanate from a particular population, may wish to increase the likelihood that terrorist cells will be rooted out by in-group enforcement within that population. Such indirect means of pursuing these political ends can, if successful, offer actors profound advantages. From an instrumental perspective, the expenditure of vast quantities of blood and treasure could potentially be avoided; further, because local populations possess detailed local knowledge that is typically poorly grasped by invasion or counterterrorism forces, a process of change "from within" holds out the promise of potentially greater efficiency as well. From a normative perspective, particular outcomes may possess, or may be seen as possessing, greater legitimacy if they are arrived at via an internal process within a population rather than through direct intervention by outsiders.

While political leaders may attempt to influence external populations through a variety of strategies, one prominent approach involves the use of what has in different settings been termed *collective sanctions* or *collective punishment.* Under such measures, an authority who wishes to influence a group's "aggregate" behavior – overall outcomes such as the amount of terrorism or the nature of a country's leadership regime – indiscriminately penalizes group members, regardless of *which* specific individuals are causally responsible for the aggregate behavior. While collective sanctions can take many different forms, these different forms rely on a common underlying logic – that individuals within a group are liable to be sanctioned until any "unfavorable" (to the sanctioner) aggregate behavior is changed, and that such sanctions may be

effective by giving group members incentives either to change their *own* behavior, or to exert influence over responsible group members in an attempt to change *theirs* (Heckathorn 1988).

Collective punishment strategies have been used in law enforcement, within organizations, and in the context of international and intergroup conflicts, not only in the historical record but also in the present day.[1] Since the 1990s, the use (or threat) of economic sanctions has become an increasingly prominent tool of diplomacy (Cortright and Lopez 2000; Drezner 2003). Proponents of economic sanctions, in the phrasing of Weiss et al (1997), often believe "that the imposition of economic coercion will exercise sufficient 'bite' that citizens in the target country will exert political pressure to force either a change in the behavior of the authorities or their removal altogether." The implicit causal mechanism reflects the underlying logic of collective punishment – that the economic costs of sanctions, suffered by ordinary citizens who likely bear no causal responsibility for regime behavior, may give citizens incentives to try to influence, or replace, the regime in accordance with the sanctioner's wishes.

A similar logic underlies common arguments in favor of certain counterterrorism strategies. Tactics such as the imposition of curfews, the widespread use of checkpoints or border closures, and military raids that cause indiscriminate damage to private property or infrastructure can be regarded as kinds of collective punishment (e.g., Khawaja 1993). In June 2006, Gilad Shalit, a young Israeli corporal, was kidnapped by extremists in the Gaza Strip; as one component of its response, the Israeli government ordered a direct military strike on elements of the Gaza infrastructure, including a key power station and a number of bridges. Acknowledging that these actions would cause hardship to innocent Palestinians, Israeli Prime Minister Ehud Olmert defended the military response by stating that "our aim is not to mete out

[1] Levinson (2003) offers historical context for and many examples of the use of collective punishment.

punishment, but to apply pressure so the soldier will be freed. We want to create a new equation – freeing the abducted soldier in return for lessening the pressure on the Palestinians." (*The Guardian*, 29 June 2006) This justification clearly reflects the underlying logic above.

Academic views on the effectiveness of collective sanctions are mixed. At an empirical level, some scholars argue that economic sanctions are generally ineffective or have likely been ineffective in key cases (Galtung 1967; Pape 1997; Levy 1999), while others differ (e.g., Drezner 2003). At a theoretical level, advocates of collective sanctioning strategies claim that "collective sanctions mobilize groups to monitor and control the conduct of their members" (Levinson 2003, p. 373) and that, through this, "collective sanctions not only *leverage* but also *build* group solidarity" (Levinson 2003, p. 350; author's emphasis) – an outcome that would be especially valuable for civil societies that are initially weak, as they may often be under repressive regimes or in lawless areas where terrorism is endemic. Collective sanctions are also argued to be potentially *efficient* means of affecting change, in part because "group members…are in an advantageous position to identify, monitor, and control responsible individuals, and can be motivated by the threat of sanctions to do so" (Levinson 2003-4, p. 348). This claim resonates with a key challenge faced by many counterinsurgency or counterterror campaigns: poor information about the internal dynamics of relevant populations. It has been argued that sanctions are likely to be most effective when the "right" actors or interest groups are affected – and that in many cases the "optimal targets" may be bystanders rather than the perpetrators of objectionable deeds (Kaempfer and Lowenberg 1988; Major and McGann 2005).

Other scholars take an opposing view, arguing that collective sanctions are likely to be counterproductive in many settings. In the contexts of counterinsurgency and counterterrorism, for example, tactics involving indiscriminate coercion have been said to backfire among other

reasons because they induce moral outrage (DeNardo 1985); because the resulting economic damage lowers the opportunity cost of supporting extremists (Bueno de Mesquita and Dickson 2007); or because such tactics signal that the government is of a type that does not care about the population's welfare (Bueno de Mesquita and Dickson 2007). More generally, sanctions may in a wide variety of contexts increase the solidarity of the targeted group (Galtung 1967; Khawaja 1993) – a development which can, in some circumstances, cut against the sanctioner's objectives.

While the existing literature offers a variety of theoretical and empirical perspectives on collective punishment, a detailed micro-level understanding of how publics respond to such sanctions remains elusive. Under what circumstances are collective sanctions more likely to be effective – or more likely to backfire? Do the effects of collective punishment – and its effectiveness – vary depending on the perceived intentions of the collectively punishing agent? Do all actors within a group respond in a similar way to collective punishment, or is the nature of response heterogeneous in a way that is potentially predictable?

This paper explores these questions within the context of a laboratory experiment, a controlled research setting that can usefully complement observational studies. Lab experiments allow for the collection of a wealth of individual-level data, a key advantage of such methods for the purpose of drawing inferences about the underlying dynamics of responses to collective punishment. For this purpose, the stylized nature of the lab environment itself offers certain advantages, allowing for the clear emergence of basic behavioral tendencies that in complex, real-world cases could be obscured by – or mistakenly attributed to – idiosyncratic situational factors. And, of course, the laboratory setting also allows the researcher to directly manipulate situational parameters of interest. While the lessons of experimental results for specific empirical contexts will always be a matter for interpretation, the measurement of strong

behavioral trends in the lab may be useful both in helping to adjudicate between conflicting interpretations of observational data, as well as in generating new hypotheses and suggesting new empirical estimation strategies for more traditional studies.

The laboratory experiment set out below explores the effects of collective punishment within the context of a public goods game. Such an experimental paradigm has several key advantages. Prominent among these is that public goods games have been well-studied in the laboratory (for a review, see Ledyard 1995), including public goods games in which actors can choose to punish one another conditional on observations of each others' contribution decisions (Fehr and Gaechter 2000; Decker, Stiehler, and Strobel 2003; Carpenter 2007). However, this study appears to be the first lab experiment involving collective punishment (as it is defined here[2]) in public goods settings. In addition, a public goods setting also seems substantively quite appropriate. Many empirical instances of collective punishment seek either to *provoke* collective action (e.g., spurring on a mass population to overthrow a disfavored regime) or to *inhibit* it (e.g., harshly cracking down on disobedient behavior in an attempt to deter larger-scale rebellion); these kinds of collective action have been fruitfully explored in the literature within public goods frameworks (Opp 1994).

The experimental scenario adopted here takes as its starting point a setup resembling Fehr and Gaechter (2000); members of a group make public goods contribution decisions, observe each others' contributions, and choose whether or not to punish one another based on these observations. This setup, however, is embedded within a larger context, in which an "Outside Authority" who receives only *aggregate* information about the behavior of "Group Members"

[2] Decker, Stiehler, and Strobel (2003) describe a study of "collective punishment" – but in their context, this refers to individually-targeted punishments chosen by a collective decision rule.

has the opportunity to punish Group Members *collectively*. Because the dynamics of response to collective punishment may differ depending on the nature of the Outside Authority's interests, the experiments were carried out in the context of two treatments: under "Opposed Interests," the Outside Authority is best off when public goods production is *minimized*, but under "Aligned Interests," the Outside Authority is best off when public goods production is *maximized*. In the context of fixed matchings that interact over a number of periods, this design allows for inquiry into many of the key questions posed above, such as the effects of collective punishment on contributions behavior, and the effects of collective punishment on the propensity to employ *in-group* punishment.

The results of the experiments offer little encouragement for advocates of collective sanctioning strategies. In the Opposed Interests treatment, collective punishment was estimated to be strictly *counterproductive*, in the sense that it provoked *more* public goods contributions rather than inhibiting them in accordance with the Outside Authority's desires. In the Aligned Interests treatment, the estimated effects are more ambiguous; on average, collective punishment does appear to provoke more public goods contributions in the short run, but also to suppress them in the longer run, largely cancelling out short-term gains. Finally, there is no evidence in either treatment that collective punishment has any significant effect on Group Members' propensities to engage in internal punishment, in contrast to theoretical expectations that collective punishment may encourage the enforcement of in-group norms.

The remainder of the paper is organized as follows. Section 2 describes in detail the experimental games and the context in which the experimental sessions were carried out. Section 3 analyzes results from the Opposed Interests Treatment, while Section 4 does the same for the Aligned Interests Treatment. Section 5 concludes.

2. Experimental Protocol

2.1. Basic Procedures

The experiments were carried out in a computerized social science lab at a large university. The results come from data collected in seven experimental sessions involving a total of 130 subjects. Subjects signed up for the experiment via a web-based recruitment system that draws from a broad pool of potential participants; individuals in the subject pool are mostly undergraduates from around the university, though a smaller number come from the broader community. Subjects were not recruited from the author's courses, and all subjects gave informed consent according to standard human subjects protocols. Subjects interacted anonymously via networked computers; the experiments were programmed and conducted with the software z-Tree (Fischbacher 1999). At the beginning of each lab session, a set of instructions describing the structure of the experimental game was not only distributed in hard copy form but also read aloud, in order to promote understanding of the lab scenarios and to induce common knowledge.[3]

In every experimental session, the number of subjects taking part was equal to a multiple of five. Initially, as described in the instructions, subjects were randomly assigned by the experimental software into quintets, which were to interact over the course of ten periods. At the same time, subjects were also randomly assigned into two roles, which were referred to using the neutral labels of "Role A" and "Role B" in the experimental instructions and protocol. Within each quintet, one subject, randomly assigned to Role A, acted as an *Outside Authority* with the ability to monitor and collectively punish four *Group Members*, who had been randomly

[3] EDITOR AND REFEREES: A copy of the instructions for subjects used in one of the treatments can be found in the "Online Appendix," which will be posted on the web prior to publication.

assigned to Role B. These four Group Members were also randomly assigned to the four "ID Numbers" 1, 2, 3, and 4. Both Role and ID Number assignments remained fixed throughout the ten periods of interaction for the quintet. At the end of these ten periods, subjects were then informed that they would be taking part in "another experiment," also lasting ten periods, which would follow the same rules. Subjects were then randomly reassigned into new quintets and new roles, subject to the constraint that any subject who had previously served in Role A would be assigned to Role B for the second ten periods. At the end of this process, subjects completed a debriefing questionnaire and received their payments for participation.

2.2. The Experimental Games

Each period of the experiment consisted of three distinct stages.

2.2.1. First Stage: Public Goods Provision

In the first stage of each period, the four Group Members were each given an endowment of 20 tokens. Each Group Member had to decide how many of these tokens to "allocate" to a "common pot" – that is, contribute to a public good – and how many to "keep" "for him or herself." As is standard in public goods games, individuals received a higher rate of return by keeping tokens than by allocating them, but tokens placed in the common pot had positive externalities for group members while kept tokens had no externalities. Specifically, the "first-stage payoffs" from the public goods game to a given Group Member i who allocated $(Contribution)_i$ tokens to the common pot in period t were:

$$(First\text{-}stage\ Group\ Member\ Payoffs)_{i,t} = (20 - (Contribution)_{i,t}) + 0.4 \sum_j (Contribution)_{j,t}$$

That is, the marginal return to a Group Member from keeping a token was 1, while the marginal return from contributing a token was 0.4; however, a decision to contribute also yielded identical

returns for other Group Members. The notation $(Others' Contributions)_{i,t}$ will be used to refer to the period-t sum of the contributions made by i's counterparts, that is, $\sum_{j \neq i} (Contribution)_{j,t}$, while $(Total\ Group\ Contributions)_t$ will refer to $\sum_j (Contribution)_{j,t}$. The Outside Authority agent made no public goods contribution decision; however, the Outside Authority did receive "first-stage payoffs" from the public goods game based on the contributions chosen by Group Members in the same quintet:

$$(First\text{-}stage\ Outside\ Authority\ Payoffs)_{k,t} = \alpha + \beta \sum_j (Contribution)_{j,t}$$

The values of α and β varied, depending on the particular treatment to which a given session was devoted. β took on a negative value during four sessions devoted to the "Opposed Interests Treatment" – so named because Group Members were collectively better off when maximizing public goods production, while the Outside Authority's payoffs *decreased* in public goods provision. In contrast, β took on a positive value during the other three sessions, devoted to the "Aligned Interests Treatment," in which the Outside Authority's payoffs instead *increased* as more public goods were provided. The values of α and β, along with the values of other parameters that varied across sessions, can be found in Table 1.

TABLE 1 ABOUT HERE

2.2.2. Second Stage: In-Group Punishment by Group Members

In the second stage of each period, Group Members had the opportunity to engage in in-group punishment. Group Members perfectly observed the first-stage choices of *(Contribution)* made by each of their counterparts, listed by their ID numbers. Group Members then had the opportunity to "reduce the first-period payoffs" of each of their counterpart Group Members, again referred to by these same ID numbers; potentially normatively-charged words such as

"punishment" or "enforcement" were avoided in the instructions and protocol. Decisions to reduce other Group Members' payoffs in this way were costly; Group Member *i* paid a marginal cost of one token for each token reduced from the payoffs of any other Group Member *j*. Such in-group punishment was subject to a budget constraint that a Group Member could spend no more than his or her first-stage payoffs on punishment. In the analyses below, $(In\text{-}Group\ Punishment)_{i,t}$ refers to the total amount by which Group Member *i* was punished by his or her three counterparts combined in period *t*, while $(Total\ In\text{-}Group\ Punishment\ Chosen)_{i,t}$ refers to the total amount by which Group Member *i* chose to reduce the payoffs of her three counterparts in that period.

2.2.3. Third Stage: Collective Punishment by the Outside Authority

In the third and final stage of each period, the Outside Authority had the opportunity to *collectively punish* all Group Members in his or her quintet. The Outside Authority agent observed Group Members' *total* first stage contributions ($\Sigma_j (Contribution)_{j,t}$), as well as the *total* amount by which Group Members chose to reduce each others' earnings in the second stage ($\Sigma_j (Total\ In\text{-}Group\ Punishment\ Chosen)_{j,t}$), but received no information about behavior at the *individual* level. The cost of collective punishment to the Outside Authority as well as the maximum feasible amount of collective punishment varied across sessions, as summarized in Table 1,[4] but in all cases the decision to collectively punish led to the subtraction of an equal number of tokens from every Group Member's payoffs. In the analyses below, *(Collective*

[4] The data from the Aligned Interests Treatment was not originally collected for the purpose of comparison with the Opposed Interests Treatment results; as a result, there were modest differences in the distributions of parameters employed across the treatments. Because the core subject of study is not direct measurement of the causal effect of Outside Authority interests on Group Member behavior, this fact does not pose any significant problem for the interpretations offered below.

Punishment)$_{i,t}$ refers to the total amount by which Group Member *i* was collectively punished by the Outside Authority in period *t*.

2.2.4 Overall Payoffs

Finally, subjects' overall payoffs for a period were constrained to be nonnegative. This constraint could never be binding for the Outside Authority, but potentially could be for Group Members who were heavily punished. Formally, the payoffs to Group Member *i* were given by:

$$(\text{Overall Period Payoffs})_{i,t} = \max(0, \{20 - (\text{Contribution})_{i,t} + 0.4\,(\text{Total Group Contributions})_t - (\text{Total In-Group Punishment Chosen})_{i,t} - (\text{In-Group Punishment})_{i,t} - (\text{Collective Punishment})_{i,t}\})$$

The conversation rates from experimental tokens into dollars for each of the sessions are depicted in Table 1. Subjects' overall payoffs for participation in a session were equal to the sum of their payoffs from each period, plus a show-up fee of US$7. In all, subjects earned US$21.66 on average.

3. Experimental Results: Opposed Interests Treatment

Section 3.1 offers descriptive statistics about overall patterns of behavior in the Opposed Interests treatment; Section 3.2 then analyzes the effects of collective punishment.

3.1. Descriptive Statistics

3.1.1. Opposed Interests, First Stage: Public Goods Provision

As in many existing experimental studies of public goods provision (see Ledyard 1995), an intermediate level of contributions was observed. On average across all periods of the Opposed Interests sessions, subjects contributed 7.09 tokens out of a possible 20 to the public good (standard deviation = 6.57). This average did not differ substantially between the first ten-period interaction (mean = 6.87, standard deviation = 6.33) and the second (mean = 7.31,

standard deviation = 6.81). Figure 1 depicts the overall time evolution of average contribution levels under Opposed Interests. After being maintained at higher levels in earlier periods, public goods contributions declined quite abruptly towards the end of each ten-period interaction, a clear indication of "end-game" effects. However, when subjects were randomly reassigned to new groups and new roles at the beginning of the second ten-period interaction (in period 11), contributions returned to approximately the same level as in period 1, indicating a "re-setting" effect. Both the "end-game" and "re-setting" effects have been observed in other experimental public goods studies (Ledyard 1995).

FIGURE 1 ABOUT HERE

The above large standard deviations in contributions are suggestive of the high level of heterogeneity in Group Member behavior. Across all periods, the four most-common contribution decisions were 0 (306 out of 1200 observations, or 25.5%); 5 (163/1200, 13.6%); 20 (121/1200, 10.1%); and 10 (120/1200, 10.0%). Considerable heterogeneity was evident within as well as between groups; across all periods, the highest-in-group contribution averaged 11.88 (standard deviation 6.50; minimum 0; maximum 20) while the lowest-in-group averaged 2.85 (standard deviation 4.25; minimum 0; maximum 20). Group-level outcomes also exhibited considerable heterogeneity, with a mean level of total group contributions equaling 28.36 (standard deviation 19.64; minimum possible value of 0 observed ten times [3.3%]; maximum possible value of 80 observed one time [0.3%]).

3.1.2. Opposed Interests, Second Stage: In-Group Punishment

As in Fehr and Gaechter (2000), Group Members sometimes exhibited a willingness to undertake costly punishment of others after observing their public goods contributions. During

each period, every Group Member had the opportunity separately to punish each of his or her in-group counterparts; of these individually-targeted punishment decisions, 8.7% (312/3600) involved a positive (non-zero) level of punishment. Overall, the average punishment choice was 0.26 tokens; conditioning on the choice being nonzero, the average punishment choice was 2.94 (minimum=1, maximum=23). The vast majority of these punishment choices were comparatively small; among the non-zero punishment choices, only 9.3% (29/312) were greater than five, and only 1.9% (6/312) were greater than ten. Individual Group Members chose to punish at least one of their three counterparts in 18.25% of periods (219/1200). On average, the total per-period punishments assigned by a Group Member came to 0.77 tokens (standard deviation = 2.50); conditioning on this value being nonzero, the average for a period was 4.19 (standard deviation = 4.45; minimum=1; maximum=23).[5]

It is important to consider *who* is being punished by *whom*; in the Opposed Interests treatment, one might imagine either that higher public goods contributors punish lower providers (because the latter did not contribute as much), or that lower public goods contributors punish higher providers (if the former fear that high contributions will induce collective punishment). It will prove useful to stratify individuals based on their ordinal in-group position with respect to public goods contributions; for all group-periods in which an identical choice was not made by all four Group Members, the following definitions will be used: (1) "max-in-group" contributors, who made the largest provision decision in their group; (2) "intermediate-in-group" contributors,

[5] The budget constraint for in-group punishment, equal to an individual Group Member's first-stage payoffs, was seldom binding. Only in 10 subject-periods (out of 1200) did a Group Member choose to spend the full available amount on in-group punishment.

who made neither the largest nor smallest provision decision in their group; and (3) "min-in-group" contributors, who made the smallest provision decision in their group.[6]

Punishment behavior is strongly correlated with these contribution-based strata. On average, individuals who were "max-in-group" contributors in the first stage of a period subtracted 1.26 tokens from other Group Members' payoffs in the second stage of that period (26.1% of punishments were non-zero), compared to 0.89 tokens by "intermediate-in-group" contributors (26.3% non-zero punishments) and only 0.23 tokens by "min-in-group" contributors (only 4.3% non-zero punishments). And, on average, "min-in-group" contributors had 1.49 tokens subtracted from their payoffs, as against only 0.43 tokens for "intermediate-in-group" contributors and 0.15 tokens for "max-in-group" contributors. Clearly, most though not all punishment is levied by *higher* public goods providers against *lower* public goods providers.

3.1.3. Opposed Interests, Third Stage: Collective Punishment

Subjects in the Outside Authority role did make substantial use of their power to collectively punish Group Members, choosing a non-zero level of collective punishment at 34.0% (102/300) of the available opportunities. While the most common collective punishment selected was 0 (66.0%, 198/300), the second-most common was the maximum possible value of 5 (12.3%, 37/300), followed by 2 (7.0%), 1 (6.3%), 3 (5.3%), and 4 (3.0%). On average, Group Members were collectively punished by 1.10 tokens (standard deviation = 1.79); conditioning on

[6] If all members of groups always made different provision decisions from one another, then there would be always be one designated "max-", two designated "intermediate-", and one designated "min-in-group" contributor(s) for each period. However, depending on the distribution of provision decisions, in a given period it is also possible that there might be, for example, two "max" contributors in the same group, if two individuals made an identical choice that exceeded the choices made by the two other group members. For the purposes of this stratified analysis, any group-periods involving identical contributions by all group members are dropped.

the collective punishment having been positive, the average was 3.24 (standard deviation = 1.56).

Each interacting quintet had four Group Members, but only one Outside Authority agent. This fact, along with variation in behavior across different Outside Authority subjects, meant that different groups were exposed to very different frequencies and intensities of collective punishment. 11 out of the 30 Outside Authority agents *never* used collective punishment in any of the ten periods they served in that role. The other 19 Outside Authority agents, who did employ collective punishment in at least one period, did so on average during 4.6 periods out of a possible 10 (ranging from three agents who did so only once, through two agents who did so in all ten periods).

While the subject of inquiry is the effects, and not the determinants, of collective punishment, it is worth noting two key facts about the behavior of Outside Authority agents. First, in the aggregate, straightforward regression analysis demonstrates that the extent of collective punishment is increasing in group-level public goods production, as is natural to expect given the structure of Outside Authority payoffs.[7] Second, though, as suggested by the small R^2 for this regression (R^2 = 0.2825), there was considerable heterogeneity in Outside Authority subjects' tastes for collective punishment. Indeed, of the 11 Outside Authority agents who never resorted to collective punishment, 5 were attached to groups in the *lower* half of the distribution of first-period *Total Group Contributions*, while 6 were attached to groups in the *upper* half of this distribution. Thus, while higher-providing groups were punished more on average, both low and high providing groups were exposed to both low and high levels of

[7] An OLS estimation (with period fixed effects) of $(Collective\ Punishment\ Chosen)_t = \beta_0 + \beta_1\ (Total\ Group\ Contributions)_t + \beta_2\ (Total\ In\text{-}Group\ Punishment)_t$ yields β_1 = 0.043*** (s.e. = 0.008) and β_2 = 0.018 (s.e. = 0.033) (robust standard errors clustered at the individual level). (N=300, 30 clusters, R^2 = 0.2825).

collective punishment during the experimental sessions. This naturally-occurring variation in Outside Authority behavior is highly useful in estimating the effects of collective punishment of group members' behavior. The following subsection turns its attention to this task.

3.2. Opposed Interests: The Effects of Collective Punishment on Public Goods Provision

TABLE 2 ABOUT HERE

Perhaps the simplest plausible method for estimating the effects of collective punishment on public goods provision is to carry out a group-level regression of *(Total Group Contributions)*$_t$ on lagged values of *(Collective Punishment)*.[8] Table 2 contains two such regression analyses, employing different lag structures. Both analyses employ ordinary least squares (OLS) regression with panel-corrected standard errors (PCSE). Because this method relies on an assumption that there is no serial correlation in the regression error terms, all OLS-with-PCSE specifications reported in all of the Tables were subjected to, and passed, a standard Lagrange multiplier test for temporal independence of the error process.[9] A lagged dependent variable was employed in all of these regressions; further lags were added when necessary to achieve serial independence. In addition, all of the reported regressions employ period fixed effects[10], and the final period of each ten-period interaction is dropped from the analysis because of the end-game effects evident in Figure 1 (and the corresponding Aligned Interests Figure 2).[11]

[8] From here on, the subscript *i* is dropped because the context is clear.

[9] That is, OLS regressions of the OLS residuals on their lags and all the independent variables returned insignificant coefficients for the lagged residuals (see Beck 2001: 279-280).

[10] Where for this purpose only the periods are treated as 1 through 20; that is, there are separate period dummies e.g. for period 1 and for period 11, even though both correspond to the first period of a ten-period interaction.

[11] It is natural to expect that dynamics will differ in the final period of a finite interaction. Because of this, unsurprisingly, including periods 10 and 20 generally slightly weakens

The specification in column (1) estimates the effect of collective punishment under the assumption that all relevant dynamics are encapsulated in a one-period lag model. This estimation suggests that, controlling for the prior level of group contributions, (*Collective Punishment*)$_{t-1}$ has a *positive* and statistically significant effect on group contributions – *an effect exactly contrary to the Outside Authority's best interests.* Specifically, group contributions are estimated to increase by about 1.19 tokens for every 1 token of collective punishment applied in the previous period.

However, there is no *ex ante* reason to expect that the relevant dynamics would necessarily be fully captured by a one-period lag model. Column (2) therefore presents results from a lag specification determined by the Akaike Information Criterion (AIC), a statistical method for choosing the most appropriate lag lengths in such analyses when theoretical expectations are weak (see e.g. Greene 1997: 786-787)[12]. In Column (2), the AIC indicates a specification with three lags of *(Collective Punishment)*. Notably, the coefficients for *(Collective Punishment)*$_{t-1}$, *(Collective Punishment)*$_{t-2}$, and *(Collective Punishment)*$_{t-3}$ are all positive and substantively large, although the last two fall somewhat short of individual statistical significance. Overall, the estimate of the direct effect of one token of *(Collective Punishment)* on *(Total Group Contributions)* over three periods – that is, the sum of these three coefficients – is to increase public goods provision by 2.269 tokens (standard error = 0.489), significant at the $p<0.01$ level. Thus, the analysis suggests that the use of collective punishment in the Opposed

---

the magnitude and occasionally the specific significance level of the estimated effects described here, but the basic picture remains the same.

[12] In this paper, because the time series are relatively short, the lag-selection process is carried out under the constraint that no more than four lags of any independent variable be selected.

Interests treatment may not only have counterproductive consequences for the Outside Authority, but that such counterproductive consequences may also be somewhat enduring.

TABLE 3 ABOUT HERE

This group-level analysis, while useful because it reflects the aggregated information actually available to the Outside Authority agents, nonetheless leaves the subtleties of individual-level behavior unexplored. Table 3 turns our attention to individual-level analyses. Column (1) of Table 3 employs a specification in which several factors that could potentially influence individual contribution decisions – *(Others' Contributions)*, *(In-Group Punishment)*, and *(Collective Punishment)* – are modeled according to a one-period lag.[13] Column (2), in contrast, presents an estimate from the specification indicated by the Akaike Information Criterion. Both Columns (1) and (2) indicate a positive effect of *(Collective Punishment)* on *(Contribution)*. The AIC specification implies an overall direct effect by which one token of *(Collective Punishment)* increases an individual's contributions by 0.51 tokens on average (s.e. = 0.13, $p<0.01$), roughly consistent with the group-level analysis given that each Outside Authority agent interacts with four Group Members.[14]

The strong estimated effects for collective punishment stand in stark contrast to the substantively small and statistically insignificant coefficient for *(In-Group Punishment)* in Column (1) – and the AIC recommended dropping this quantity altogether in generating the Column (2) specification. The regression in Column (3), which simply re-integrates *(In-Group Punishment)* into the Column (2) specification at the same number of lags as *(Collective*

[13] As noted above, additional lags of dependent variables are included when necessary, as here, to remove serial correlation from regression error terms.
[14] Note that the AIC barely indicated against a third lag for *(Collective Punishment)*, which would have yielded an overall direct effect of 0.60 tokens (s.e. = 0.15, $p<0.01$).

Punishment), illustrates that there is no significant estimated effect of *(In-Group Punishment)* on contribution behavior, and the inclusion (or omission) of this factor does not significantly influence other estimates.

Notably, the estimated coefficient for $(Others' Contributions)_{t-1}$ is positive while the coefficient for $(Others' Contributions)_{t-2}$ is *negative* in both columns (2) and (3). These findings are most naturally interpreted as indicating that, on average, individuals contribute more in a given period when their counterparts have contributed more in the *previous* period (when $(Others' Contributions)_{t-1}$ is higher), and that this is even *more* the case when their counterparts' previous-period contributions represented an *increase* over their prior contributions behavior (that is, when $(Others' Contributions)_{t-2}$ was *lower*, for fixed $(Others' Contributions)_{t-1}$.) These results fit well with natural intuitions about conditional cooperation and positive reciprocity in group dynamics.

Columns (4) and (5) offer two further robustness checks to these basic findings. Column (4) offers an alternative specification for the time effects of the Column (2) quantities of interest – the last lag of each independent variable in replaced with a new quantity summarizing the "average past" values of the variable. For example, $(Avg.\ Past\ Coll.\ Punishment)_{t-1}$ takes the amount of collective punishment received in every period up through and including *t-2* and calculates a per-period average value. The intuitions conveyed by the other specifications carry over here; in particular, the estimated effect of $(Avg.\ Past\ Coll.\ Punishment)_{t-1}$ is positive and substantively large, although it just barely falls short of statistical significance at conventional levels ($p<0.102$). Column (5) carries out a tobit estimation of the AIC lag structure from Column (2). Individual contribution decisions cannot be less than 0 and cannot exceed 20; when the precise quantity of interest for estimation is the effect of some independent variable on

behavior when choice is not constrained, the presence of "censored" observations (here, 0 or 20) can downwardly-bias OLS estimations of effect sizes. In accordance with this intuition, the tobit estimation in Column (5), accounting for both left- and right-censored observations, suggests that the effect of collective punishment on unconstrained agents may be modestly larger than the effect estimated by OLS. As a final robustness check, it was noted that the estimated effect of collective punishment was found to be positive and of similar magnitude in each of two sub-samples defined when the data was split depending on whether $(Total\ Group\ Contributions)_{t-1}$ was above or below the median value of 24.

The analyses in Table 3 used data pooled from all subjects. Yet, it seems plausible that some of the factors affecting contribution behavior might affect different subjects in different ways. For example, in-group punishment might be expected to have a different effect on low-providers (who might be spurred to higher production upon receiving in-group sanctions) than on higher-providers (who might become annoyed at receiving in-group sanctions despite their higher initial contributions, and therefore decide to contribute less). This is especially true because, in the Opposed Interests Treatment, individuals might in principle choose to punish other Group Members for two more or less opposite reasons – to *encourage* contributions, or to *discourage* them out of fear of the Outside Authority's potential response. Similarly, the dynamics of response to collective punishment, the primary substantive interest here, could potentially be heterogeneous across individuals as well.

As a means of inquiring into such potential heterogeneity, Columns (2)-(4) of Table 4 present disaggregated analyses, stratified based on individuals' relative levels of public goods production in the previous period (*t-1*). This approach employs the stratification categories defined earlier: "max-in-group," "intermediate-in-group," and "min-in-group." Because the

dynamics of subject behavior are complex, these three strata do not conform precisely to ideal types of agents – for example, an intrinsically high provider of public goods, with a characteristic way of responding to different factors in the environment, may dramatically lower her contributions for one or a few periods in order to signal her annoyance with other group members' behavior – and end up being counted in the "min-in-group" category for that one (or those few) periods. Despite such behavioral "noise," however, Table 4 does offer useful insights.

TABLE 4 ABOUT HERE

Column (1) of Table 4 simply reproduces Column (1) of Table 3, the pooled analysis with one period of lag.[15] Columns (2), (3), and (4) contain estimations of the same specification, strictly disaggregated across the three contribution strata. Comparing the results across columns, it seems clear that the dynamics of individual contributions decisions do indeed vary for individuals with different contribution histories. $(\textit{Others' Contributions})_{t-1}$ has a strong effect on "max-in-group" contributors' period t decisions, and, to a modestly lesser extent, on "intermediate-in-group" contributors' decisions, but it has no significant effect on "min-in-group" contributors' decisions. In this sense, "min-in-group" contributors appear rather disengaged from the group interaction; on average, their contribution behavior is not much affected by others' prior contribution decisions. Results also differ markedly across strata with

[15] Because subjects move, not infrequently, from one stratum to another over periods, the amount of data available for disaggregated multi-lag specifications (in which membership in a given stratum remains fixed over all the lags) is quite limited. Because of this, and because in the Opposed Interests treatment the effect of *(Collective Punishment)* is of the same sign over all substantively significant lags anyway, attention is restricted here to the one-lag specification. Also, because individuals' entries into and exits from strata make traditional time series methods inapplicable, the estimations in Columns (2)-(4) of Table 4 are simply OLS with standard errors clustered at the individual subject level.

respect to $(\textit{In-Group Punishment})_{t-1}$. For both "intermediate-in-group" and "min-in-group" contributors, the estimated effects of $(\textit{In-Group Punishment})_{t-1}$ on $(\textit{Contributions})_t$ is positive and of intermediate magnitude – less than the coefficients seen earlier for collective punishment, but at the same time potentially meaningful substantively – although they do not achieve statistical significance. In contrast, the effect for "max-in-group" contributors is negative, very large in magnitude, and highly statistically significant. One natural interpretation of this result is that "max-in-group" contributors may become annoyed when punished by lesser contributors, and in consequence sharply reduce their willingness to contribute to a public good benefiting all Group Members. Finally, the estimated effect of $(\textit{Collective Punishment})_{t-1}$ on $(\textit{Contributions})_t$ is positive, substantively quite large, and highly statistically significant for "max-in-group" and "min-in-group" contributors, but is markedly smaller for "intermediate-in-group" contributors.

This finding sheds some useful light on the microfoundations of collective punishment's counterproductive aggregate effects, and invites potentially fruitful speculation as to its ultimate causes. High providers of public goods – those for whom group ideals are presumably initially more salient – react to collective punishment of the group by putting even *more* effort into group ends. Interestingly, the same is true of *low* providers of public goods – those for whom group ideals are presumably initially least salient. These low providers have, compared to others in their group, behaved in a way most congenial to the Outside Authority's interests – and were nonetheless punished by that Outside Authority. Low providers then respond by *increasing* their contributions to the group interest, contrary to the Outside Authority's desires. However, the estimated effect of collective punishment on "intermediate-in-group" contributors' behavior is much smaller and falls short of statistical significance; on average, such individuals may be less committed to the group good than "max-in-group" providers, while also having played a greater

role in bringing collective punishment on the group – and on themselves – than low-providers did. On average, agents in this position react least strongly to being collectively punished.

3.3. Opposed Interests: The Effects of Collective Punishment on In-Group Enforcement

TABLE 5 ABOUT HERE

While collective punishment is estimated to have a substantial effect on public goods contribution behavior in the Opposed Interests Treatment, there is no evidence of a significant effect on in-group punishment behavior. Table 5 describes several specifications in which *(Total In-Group Punishment Chosen)*$_t$ was regressed on the same right-hand-side variables previously employed. The Akaike criterion indicated that *all* lags of *(Collective Punishment)* should be dropped; as such, column (1) simply displays an estimation in which all independent variables are lagged to period *t-1*. The coefficient of (*Collective Punishment*)$_{t-1}$ is indeed tiny and statistically insignificant. Columns (2)-(4), which offer parallel analyses disaggregated by behavioral stratum, also show no clear or significant patterns indicating an effect of (*Collective Punishment*)$_{t-1}$ on *(Total In-Group Punishment Chosen)*$_t$. Meanwhile, the other coefficients in the table suggest that the extent of in-group punishment chosen is larger on average when one's own contribution is higher; when others' contributions are lower and decreasing; and when one has been punished by in-group members oneself. All of these effects appear to be driven jointly by max-in-group and intermediate-in-group contributors.

Several further estimations also failed to return statistically significant effects of *(Collective Punishment)* on *(Total In-Group Punishment Chosen)*$_t$. As noted above, the distribution of *(Total In-Group Punishment Chosen)*$_t$ contained a small number of unusually large values, the presence of which could imaginably cover up a systematic relationship between

(Collective Punishment) on *(Total In-Group Punishment Chosen)*$_t$. As such, parallel analyses were run with two variants of *(Total In-Group Punishment Chosen)*$_t$ – one which truncated this variable at 11 (i.e., all values higher than 11 were re-assigned to a value of 11), and another which truncated it at 6. No significant effects for *(Collective Punishment)* were found for either of these truncated forms either; nor did the use of tobit rather than OLS estimators make a difference.

4. Experimental Results: Aligned Interests Treatment

4.1. Descriptive Statistics

As above, this section begins with a description of overall behavior during the experimental sessions devoted to the Aligned Interests treatment.

4.1.1. Aligned Interests, First Stage: Public Goods Provision

Across all periods of the Aligned Interests Treatment, subjects contributed 10.47 tokens to the public good on average, out of a possible 20 tokens (standard deviation = 7.72 tokens). The average contribution levels were again similar between the first ten-period interaction (mean=9.86, standard deviation = 7.63) and the second ten-period interaction (mean=11.08, standard deviation = 7.77). Figure 2 depicts the time evolution of overall contribution levels in each of these ten-period interactions. As in the Opposed Interests Treatment, contributions drop off somewhat towards the end of each ten-period interaction, but "re-set" between the end of the first ten and the beginning of the second ten periods.

FIGURE 2 ABOUT HERE

Contributions behavior in the Aligned Interests Treatment also exhibits a considerable degree of cross-subject heterogeneity, within groups as well as across the sample as a whole. The four

most-common contribution decisions (across all periods) were: 20 (255 out of 880 observations, or 29.0%); 0 (147/880, 16.7%); 10 (90/880; 10.2%); and 5 (72/1200; 8.2%). The highest-in-group contribution averaged 15.2 (standard deviation 5.85; minimum 0; maximum 20), while the lowest-in-group contribution averaged only 5.71 (standard deviation 7.25; minimum 0; maximum 20). The mean level of contributions at the group level across all periods was 41.87 (standard deviation 24.39; minimum possible value of 0 occurred six times [2.7%]; maximum possible value of 80 occurred thirty times [13.6%]).

4.1.2. Aligned Interests, Second Stage: In-Group Punishment

In contrast to the Opposed Interests treatment, the Outside Authority's payoffs are *increasing* in public goods provision in the Aligned Interests treatment. This suggests a somewhat conceptually simpler dynamic for punishment under Aligned Interests – both in-group and collective punishment are likely to be employed in pursuit of the same goal, increasing public goods provision.

Overall, 6.3% (167/2640) of group members' individually-targeted in-group punishment choices involved a positive (non-zero) level of punishment, a level similar to that of the Opposed Interests treatment. The average punishment choice was 0.28; conditioning on the choice being nonzero, the average punishment choice was 4.41 tokens (minimum = 1, maximum = 27). The vast majority of these punishment choices were comparatively small; among the non-zero punishment choices, 19.2% (32/167) were greater than five, and only 8.4% (14/167) were greater than ten. Individual Group Members chose to punish at least one of their three counterparts in 14.3% of periods (126/880). On average, the total per-round punishments assigned by a Group

Member were 0.84 tokens (standard deviation = 3.10); conditioning on this value being nonzero, the average for a period was 5.84 (standard deviation = 6.17; minimum=1; maximum=28).[16]

As in the Opposed Interests Treatment, the bulk of in-group punishment was carried out by "max-in-group" contributors, who on average subtracted 1.47 tokens from other Group Members' payoffs (25.1% of punishments were non-zero). This compares with a figure of 0.61 tokens for "intermediate-in-group" contributors (18.1% non-zero punishments) and 0.74 tokens for "min-in-group" contributors (only 6.3% non-zero punishments), with virtually all of the figure for "min-in-group" contributors due to one aberrant subject.[17] The vast majority of these punishments were levied against "min-in-group" contributors, who on average had 2.52 tokens subtracted from their payoffs, versus 0.46 tokens for "intermediate-in-group" contributors and 0.03 tokens for "max-in-group" contributors.

4.1.3. Aligned Interests, Third Stage: Collective Punishment

Subjects in the Outside Authority role did make substantial use of their power to collectively punish Group Members, choosing a non-zero level of collective punishment at 40.5% (89/220) of the available opportunities (36.3% in the Aligned-1 and Aligned-2 sessions combined, 51.7% in the Aligned-3 session). On average, Outside Authority agents collectively punished Group Members by 1.96 tokens in the Aligned-1/Aligned-2 sessions (standard deviation = 3.32) and by 1.63 tokens (standard deviation = 1.87) in the Aligned-3 session.

[16] Only in 14 subject-periods (out of 880) did a Group Member choose to spend the maximum possible amount on in-group punishment.

[17] Over the course of seven periods as a min-in-group contributor, this subject spent 160 tokens to remove 160 tokens from low-performing counterparts (even though the subject contributed nothing during these periods). Without this one subject, "min-in-group" contributors subtracted on average 0.08 tokens from others' payoffs, with only 3.5% of punishments being non-zero. Overall, without this subject, the average punishment choice (conditioning on the choice being nonzero) was 3.52; 14.2% (4.5%) of non-zero punishment choices were greater than five (ten); and the total per-round punishment assigned by Group Members (conditioning on this value being nonzero) averaged 4.67.

Conditional on the collective punishment being nonzero, these values were 5.40 in Aligned-1/Aligned-2 (standard deviation = 3.44, minimum = 1, maximum = 10) and 3.16 in Aligned-3 (standard deviation = 1.37, minimum = 1, maximum = 5).[18] In Aligned-1/Aligned-2, 6 out of the 16 Outside Authority agents *never* used collective punishment; the other 10 did so on average 5.8 times. In Aligned-3, all Outside Authority agents used collective punishment at least once, doing so on average 5.2 times.

As one might expect given Outside Authority incentives in the Aligned Interests Treatment, the extent of collective punishment was decreasing in group-level public goods production.[19] However, as in the parallel Opposed Interests analysis, the R^2 associated with this result was low (R^2 = 0.2140), again indicating considerable heterogeneity in Outside Authority subjects' collective punishment responses to group public goods provision behavior.

4.2. Aligned Interests: The Effects of Collective Punishment on Public Goods Provision

TABLES 6 & 7 ABOUT HERE

Tables 6-8, which contain analyses of the effects of collective punishment under Aligned Interests, closely parallel Tables 2, 3, and 5 from the Opposed Interests Treatment. These analyses use data from all three Aligned Interests sessions. The first glimpse at the data offered by Column (1) in both Tables 6 and 7 suggests that collective punishment may again provoke increases in public goods contributions – which in the Aligned Interests Treatment would work in the Outside Authority's favor. These 1-period lag models at the group and individual level both show a positive coefficient for $(Collective\ Punishment)_{t-1}$, although the estimates are less

[18] In Aligned-1/Aligned-2, 13.8% (22/160) of collective punishment choices exceeded 5.

[19] An OLS estimation (with period fixed effects) of $(Collective\ Punishment)_t = \beta_0 + \beta_1 (Total\ Group\ Contributions)_t + \beta_2 (Total\ In\text{-}Group\ Punishment)_t$ yields β_1 = -0.059*** (s.e. = 0.017) and β_2 = -0.048 (s.e. = 0.039) (robust standard errors clustered at the individual level). (N=220, 22 clusters, R^2 = 0.2140).

than half as large as the corresponding estimates in the Opposed Interests Treatment, and fall somewhat short of statistical significance. However, Column (2) in both Tables 6 and 7 tells a more complicated story. In both tables, the Akaike Information Criterion indicates that a lengthy lag structure is most appropriate, given the data. Strikingly, the coefficients for *(Collective Punishment)*$_{t-1}$ and *(Collective Punishment)*$_{t-2}$ are positive, while the coefficients for *(Collective Punishment)*$_{t-3}$ and *(Collective Punishment)*$_{t-4}$ are *negative* in both tables. While the coefficient for *(Collective Punishment)*$_{t-1}$ is large and strongly significant – indicating an additional 1.35 tokens of public goods provision for each token of collective punishment – the first two coefficients are nearly, but not quite, balanced by the second two. Over four periods, the estimated direct effect of one token of collective punishment in Table 6 (at the group level) is 0.23 (s.e.=0.43, not significant), and in Table 7 (at the individual level) is 0.06 (s.e.=0.11, n.s.).[20]

This pattern is quite striking. It suggests that, in the Aligned Interests treatment, collective punishment by the Outside Authority does, in accordance with the Authority's interests, provoke further public goods provision – in the *short* run. Yet, in the *medium* run, past collective punishment is estimated to *suppress* contributions instead. The short-run effect is easy to understand, and congruent with the theoretical expectations of collective sanctions supporters. The Outside Authority imposes a cost, implicitly linked to a behavior the Authority wishes to influence – and Group Members, in response to that cost, adjust their behavior in the desired direction. Interestingly, though, the direction of the medium-run effect is entirely the opposite. One potential interpretation is that Group Members at first increase their contributions upon

[20] These overall effect estimates are quite robust to changes in the specification. For example, perturbing the AIC-selected lag structure in Table 7, the overall effect of *(Collective Punishment)* is 0.04 (s.e.=0.10) with three lags of this variable, and 0.025 (s.e.=0.097) with five lags.

being collectively punished, as a concession to the Outside Authority, but then modestly lower their contributions later once it seems to them that they may be able to get away with it.

These results were obtained using the data from all three Aligned Interests sessions. It is worth noting that the estimated overall direct effects obtained when separately estimating specification (2) in Table 7 for the Aligned-1 and Aligned-2 sessions were essentially zero and for the Aligned-3 session was -0.265 tokens. Thus, there is no evidence for an overall positive direct effect either in Aligned-1/Aligned-2 or in Aligned-3, across which the parameters varied modestly.

The above results have been posed in terms of the "direct effects" of collective punishment on $(Contribution)_t$. In principle, there may be indirect effects as well. For example, $(Collective\ Punishment)_{t-2}$ may have a direct effect on $(Contribution)_{t-1}$; the value of $(Contribution)_{t-1}$, higher or lower than it would have been in the absence of collective punishment, may then itself directly influence $(Contribution)_t$. Such indirect effects cannot be clearly discerned from Table 7, for example because the relationship between $(Contribution)_t$ and $(Contribution)_{t-1}$ may differ depending on the past collective punishment history. One potential strategy to get a sense of the sign and likely magnitude of any indirect effects is to estimate $(Contribution)_t = \beta_0 + \beta_1 (Contribution)_{t-k} + \beta_2 (Others' Contributions)_{t-k} + \beta_3 (In\text{-}Group\ Punishment)_{t-k} + \beta_4 (Collective\ Punishment)_{t-k}$ for k = {2,3,4}. The coefficients β_4 can be thought of as containing information about both any direct and any indirect effects of $(Collective\ Punishment)_{t-k}$ on $(Contribution)_t$ in these regressions because values of the independent variables subsequent to $(Collective\ Punishment)_{t-k}$ are omitted. Such a specification is obviously an incomplete model of the overall system, but has the advantage that any indirect effects of $(Collective\ Punishment)_{t-k}$ would load onto β_4. This coefficient takes on the values 0.05 for k=2,

-0.01 for k=3, and -0.19 for k=4, suggesting that there are no substantial positive indirect effects of collective punishment in the Aligned Interests treatment.[21]

Columns (3) and (4) carry out robustness checks analogous to those in columns (3) and (5) of Table 3. Neither the re-introduction of *(In-Group Punishment)*, again dropped by the AIC, or use of a tobit estimator makes any qualitative difference to the results. In addition, the same overall qualitative patterns shown in Table 7 were visible in split-sample analyses carried out separately on data for which $(Total\ Group\ Contributions)_{t-1}$ was above vs below the median value of 38. No analysis analogous to Column (4) of Table 3 is carried out, because the signs of the direct effects of collective punishment vary here across different lags.

For the Opposed Interests analysis, Table 4 offered a glimpse into heterogeneity in behavior by disaggregating the one-period-lag analysis into the three behavioral strata described earlier. Clearly, Tables 6 and 7 suggest that a one-period-lag analysis would not cut to the heart of the Aligned Interests dynamics; and, because of individual subject mobility across strata over different periods, only very limited data exists allowing for a multi-period-lag analysis of this kind for Aligned Interests. The data that does exist is however consistent with a pattern of initially positive, and then negative response both for "max-in-group" and "min-in-group" contributors, though not at conventional levels of statistical significance; a meaningful amount of data does not exist for "intermediate-in-group" contributors. Thus, there is no clear evidence of differential response to collective punishment across strata, though limitations in the data allow no firm conclusions.

4.3. Aligned Interests: The Effects of Collective Punishment on In-Group Enforcement

[21] The corresponding figures for the Opposed Interests treatment were 0.66 for k=2, 0.89 for k=3, and 1.01 for k=4; comparison with the direct effect coefficients suggests the possibility of modestly positive indirect effects.

TABLE 8 ABOUT HERE

As with Table 5 for Opposed Interests, Table 8 does not indicate any pattern of significant effects of collective punishment on the propensity to carry out in-group punishment. The AIC specification, which as in the Opposed Interests case recommended dropping all *(Collective Punishment)* terms, is not shown for this reason; instead, the lack of effects is depicted in the table, for pooled data as well as for analyses disaggregated by contribution stratum, for independent variables lagged back to *t-1*. The same robustness checks carried out in Section 3.3 were again put to work here – employing truncated forms of the dependent variable to reduce the impact of outliers, and using tobit estimators instead of OLS – and in addition, parallel analyses were run in which groups containing the aberrant subject described earlier were dropped. All of these estimations failed, as before, to uncover any significant effect of collective punishment on group members' in-group punishment choices.[22]

For theoretical perspectives arguing that collective punishment can encourage the enforcement of in-group norms, the Aligned Interests treatment would seem to constitute an ideal scenario; the collective punisher's interests are in line with "group interests," and in-group punishment of low contributions is a clear mechanism by which group norms can be enforced. Therefore, it is especially striking that the use of collective punishment had no apparent impact on group members' propensities to punish each others' transgressions internally.

5. Conclusion

The experiments described in this paper explored individual responses to collective punishment in a public goods provision setting. The results suggest, at least in the context of the

[22] Dropping groups containing the aberrant subject also had no meaningful effect on the estimations in Table 7.

specific experimental design, that collective punishment is fairly ineffective at best and strongly counterproductive at worst in shaping group behavior according to the desires of an outside authority. When the outside authority's interests were best served by minimum provision of public goods, collective punishment actually led to a marked *increase* in individual contributions, leaving the outside authority strictly worse off. When instead the outside authority's interests were best served by maximum public goods provision, collective punishment on average modestly stimulated contributions in the short run, but on average also modestly suppressed them in the medium run, largely cancelling out the short-run gains. Finally, in neither treatment was there any evidence that collective punishment affected the propensity of group members to enforce in-group norms, a hypothesized mechanism that lies at the heart of many arguments in favor of collective sanctioning strategies.

Of course, the results presented here were obtained in the context of specific experimental scenarios, and these results cannot logically exclude the possibility that collective punishment might be found to be more effective in other settings. At the same time, it is striking that collective punishment in the Opposed Interests treatment provoked a backlash despite the instantiation of the experiment in a stylized laboratory environment, in which individual subjects were randomly assigned to groups and to strategic roles. Such an environment lacks myriad factors which might naturally be supposed to facilitate group response to collective punishment in real-world settings – for example, preexisting group loyalties, long-standing grievances against the punishing agent, the ability to exchange communications, or leaders who could serve as focal points for coordinating dissent. It is equally striking that collective punishment did not appear to affect the extent to which Group Members punished one another, in spite of the facts that information about individual contribution decisions was perfect and that an anonymous

mechanism for enforcing group norms was readily at hand.

An interesting extension to this research would be to consider the dynamics of collective punishment when the intentions of the outside authority are uncertain. Such a context may arise, for instance, in the aftermath of a military victory by an external force or a relatively unknown faction. In such contexts, exactly what constitutes the "public good" may be unclear or seem ambiguous to members of the population potentially subject to collective punishment; assisting the outside authority in, for example, disarming an insurgency might be in group members' interests if the authority is benign, whereas group members might be better off *supporting* the insurgency if the authority has darker motives. With an appropriate experimental design, it should be possible to explore in more depth the way in which individuals form assessments of actors' intentions based on the punishment strategies those actors do or do not employ. Given the responses to collective punishment observed in the current experiments, it is easy to imagine that collective punishment may prove counterproductive in this broader sense as well.

References

Beck, Nathaniel. 2001. "Time-Series—Cross-Section Data: What Have We Learned in the Past Few Years?" *Annual Review of Political Science* 4: 271-293.

Bueno de Mesquita, Ethan and Eric S. Dickson. 2007. "The Propaganda of the Deed: Terrorism, Counterterrorism, and Mobilization." *American Journal of Political Science* 51(2): 364-381.

Carpenter, Jeffrey P. 2007. "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods." *Games and Economic Behavior* 60(1): 31-51.

Cortright, David and George A. Lopez. 2000. *The Sanctions Decade: Assessing UN Strategies in the 1990s.* Boulder, CO: Lynne Reiner.

Decker, Torsten, Andreas Stiehler, and Martin Strobel. 2003. "A Comparison of Punishment Rules in Repeated Public Good Games: An Experimental Study." *Journal of Conflict Resolution* 47(6): 751-772.

DeNardo, James. 1985. *Power in Numbers: The Political Strategy of Protest and Rebellion.* Princeton NJ: Princeton University Press.

Drezner, Daniel W. 2003. "The Hidden Hand of Economic Coercion." *International Organization* 57: 643-659.

Fehr, Ernst and Simon Gaechter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90(4): 980-994.

Fischbacher, Urs. 1999. "z-Tree – Zurich Toolbox for Readymade Economic Experiments – Experimenter's Manual." Working Paper Nr. 21, Institute for Empirical Research in Economics, University of Zurich.

Galtung, Johan. 1967. "On The Effects of International Economic Sanctions: With Examples from the Case of Rhodesia." *World Politics* 19(3): 378-416.

Greene, William H. 1997. *Econometric Analysis* (Third Edition). Upper Saddle River, NJ: Prentice Hall.

The Guardian (newspaper). 29 June 2006. "Israeli Tanks Turn Screw on Besieged Gaza Strip."

Heckathorn, Douglas D. 1988. "Collective Sanctions and the Creation of Prisoner's Dilemma Norms." *American Journal of Sociology* 94(3): 535-562.

Jervis, Robert. 1976. *Perception and Misperception in International Politics.* Princeton NJ: Princeton University Press.

Kaempfer, William H. and Anton D. Lowenberg. 1988. "The Theory of International Economic Sanctions: A Public Choice Approach." *American Economic Review* 78(4): 786-793.

Khawaja, Marwan. 1993. "Repression and Popular Collective Action: Evidence from the West Bank." *Sociological Forum* 8(1): 47-71.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research," in *Handbook of Experimental Economics*, John H. Kagel and Alvin E. Roth, eds. Princeton: Princeton University

Levinson, Daryl J. 2003. "Collective Sanctions." *Stanford Law Review* 56: 345-428.

Levy, Philip I. 1999. "Sanctions on South Africa: What Did They Do?" *American Economic Review* 89(2): 415-420.

Major, Solomon and Anthony J. McGann. 2005. " 'Innocent Bystanders' as Optimal Targets of Economic Sanctions." *Journal of Conflict Resolution* 49(3): 337-359.

McDermott, Rose. 2004. *Political Psychology in International Relations.* Ann Arbor: University of Michigan Press.

Mercer, Jonathan. 1996. *Reputation and International Politics.* Ithaca: Cornell University Press.

Pape, Robert A. 1997. "Why Economic Sanctions Do Not Work." *International Security* 22(2): 90-136.

Weiss, Thomas, David Cortright, George Lopez, and Larry Minear, eds. 1997. *Economic Gain and Civilian Pain: Humanitarian Impacts of Economic Sanctions.* New York: Rowman & Littlefield.

Table 1. Details of Experimental Sessions

Session	Subjects	Periods	α	β	Unit cost of collective punishment	Maximum amount of collective punishment	Token Conversion Rate
Opposed-1	20	20 (=10+10)	60	–0.4	0	5	30
Opposed-2	20	20 (=10+10)	60	–0.4	0	5	40
Opposed-3	15	20 (=10+10)	60	–0.4	0	5	40
Opposed-4	20	20 (=10+10)	60	–0.4	0	5	40
Aligned-1	20	20 (=10+10)	10	+0.4	0.5	10	30
Aligned-2	20	20 (=10+10)	10	+0.4	0.5	10	30
Aligned-3	15	20 (=10+10)	10	+0.4	0	5	40

Table 2. Opposed Interests Treatment: Collective Punishment and Group Public Goods Contributions

Dependent Variable: *(Total Group Contributions)*$_t$

Ordinary Least Squares (OLS) with Panel-Corrected Standard Errors (PCSE)

	(1) 1-period lag	(2) lag structure selected by Akaike Information Criterion (AIC)
(Total Group Contributions)$_{t-1}$	0.886 (0.069)***	0.720 (0.160)***
(Total Group Contributions)$_{t-2}$		0.111 (0.143)
(*Collective Punishment*)$_{t-1}$	1.194 (0.388)***	0.919 (0.425)**
(*Collective Punishment*)$_{t-2}$		0.681 (0.482)
(*Collective Punishment*)$_{t-3}$		0.669 (0.513)
Period fixed effects?	Yes	Yes
N	240	180
Groups	30	30
R^2	0.8114	0.8445

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

Table 3. Opposed Interests Treatment : Effect of Collective Punishment on Individual Public Goods Contributions

Dependent Variable: $(Contribution)_t$

	(1) OLS w/ PCSE, 1 lag in indep vars	(2) OLS w/ PCSE, AIC lag structure	(3) OLS w/ PCSE	(4) OLS w/ PCSE	(5) Tobit w/ uncorrected SE
$(Contribution)_{t-1}$	0.353*** (0.129)	0.322** (0.160)	0.329** (0.161)	0.328** (0.154)	0.446*** (0.055)
$(Contribution)_{t-2}$	0.207* (0.113)	0.196 (0.147)	0.187 (0.149)	0.213 (0.149)	0.283*** (0.055)
$(Contribution)_{t-3}$		0.092 (0.114)	0.092 (0.114)		0.129** (0.052)
$(Avg\ Past\ Contribution)_{t-1}$				0.073 (0.079)	
$(Others'\ Contributions)_{t-1}$	0.112*** (0.026)	0.133*** (0.026)	0.132*** (0.026)	0.140*** (0.027)	0.178*** (0.033)
$(Others'\ Contributions)_{t-2}$		-0.050* (0.027)	-0.048* (0.028)		-0.086*** (0.032)
$(Others'\ Avg\ Past\ Contributions)_{t-1}$				-0.073*** (0.022)	
$(In\text{-}Group\ Punishment)_{t-1}$	0.025 (0.078)		0.042 (0.073)		
$(In\text{-}Group\ Punishment)_{t-2}$			-0.068 (0.082)		
$(In\text{-}Group\ Punishment)_{t-3}$					
$(Collective\ Punishment)_{t-1}$	0.374*** (0.109)	0.277** (0.110)	0.266** (0.109)	0.185* (0.110)	0.340** (0.161)
$(Collective\ Punishment)_{t-2}$		0.235* (0.123)	0.243** (0.121)		0.308* (0.163)
$(Avg\ Past\ Coll.\ Punishment)_{t-1}$				0.307 (0.188)	
Period fixed effects?	Yes	Yes	Yes	Yes	Yes
N	840	720	720	720	720
Individuals	75	75	75	75	75
R^2	0.5641	0.5823	0.5829	0.5841	-

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

Table 4. Opposed Interests Treatment: Effect of Collective Punishment by Contribution Stratum

Dependent Variable: $(Contribution)_t$

	(1) OLS w/ PCSE, all subjects	(2) OLS w/ clustered SE, only period *t*-1 max-in-group contributors	(3) OLS w/ clustered SE, only period *t*-1 intermediate-in-group contributors	(4) OLS w/ clustered SE, only period *t-1* min-in-group contributors
$(Contribution)_{t-1}$	0.353(0.129)***	0.254(0.089)***	0.413(0.121)***	0.667(0.119)***
$(Contribution)_{t-2}$	0.207(0.113)*	0.175(0.098)*	0.129(0.073)*	0.237(0.048)***
$(Others' Contributions)_{t-1}$	0.112(0.026)***	0.190(0.040)***	0.127(0.040)***	0.005(0.029)
$(In\text{-}Group\ Punishment)_{t-1}$	0.025(0.078)	-0.763(0.329)**	0.114(0.118)	0.160(0.118)
$(Collective\ Punishment)_{t-1}$	0.374(0.109)***	0.495(0.187)***	0.176(0.171)	0.401(0.152)***
Period fixed effects?	Yes	Yes	Yes	Yes
N	840	265	249	318
Indivs./Clusters	75	67	64	70
R^2	0.5641	0.6508	0.4976	0.5182

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

Table 5. Opposed Interests Treatment: Effect of Collective Punishment on In-Group Punishment Choices

Dependent Variable: *(Total In-Group Punishment Chosen)$_t$*

	(1) OLS w/PCSE, all subjects	(2) OLS w/ clustered SE, only period *t* max-in-group contributors	(3) OLS w/ clustered SE, only period *t* intermediate-in-group contributors	(4) OLS w/ clustered SE, only period *t* min-in-group contributors
(Total In-Group Punishment Chosen)$_{t-1}$	0.230(0.142)	0.468(0.187)**	0.199(0.047)***	0.020(0.022)
(Contribution)$_t$	0.105(0.015)***	0.136(0.078)*	0.151(0.092)	0.018(0.016)
(Contribution)$_{t-1}$	-0.029(0.021)	-0.044(0.059)	-0.025(0.036)	-0.017(0.011)
(Others' Contributions)$_t$	-0.032(0.010)***	-0.060(0.031)*	-0.071(0.031)**	0.003(0.008)
(Others' Contributions)$_{t-1}$	0.018(0.010)*	0.043(0.027)	0.027(0.023)	-0.007(0.008)
(In-Group Punishment)$_{t-1}$	0.046(0.028)*	0.073(0.076)	0.090(0.109)	-0.014(0.013)
(Collective Punishment)$_{t-1}$	-0.027(0.045)	0.078(0.090)	-0.106(0.083)	-0.021(0.023)
Period fixed effects?	Yes	Yes	Yes	Yes
N	960	294	282	360
Indivs./Clusters	75	67	62	70
R^2	0.1310	0.2749	0.1127	0.0373

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

Table 6. Aligned Interests Treatment: Collective Punishment and Group Public Goods Contributions

Dependent Variable: *(Total Group Contributions)*$_t$

Ordinary Least Squares (OLS) with Panel-Corrected Standard Errors (PCSE)

	(1) 1-period lag	(2) lag structure selected by Akaike Information Criterion (AIC)
(Total Group Contributions)$_{t-1}$	0.998 (0.054)***	0.825 (0.200)***
(Total Group Contributions)$_{t-2}$		0.362 (0.257)
(Total Group Contributions)$_{t-3}$		-0.105 (0.229)
(Total Group Contributions)$_{t-4}$		-0.129 (0.140)
(*Collective Punishment*)$_{t-1}$	0.506 (0.311)	1.351 (0.427)***
(*Collective Punishment*)$_{t-2}$		0.473 (0.501)
(*Collective Punishment*)$_{t-3}$		-0.753 (0.400)*
(*Collective Punishment*)$_{t-4}$		-0.842 (0.522)
Period fixed effects?	Yes	Yes
N	176	110
Groups	22	22
R^2	0.8496	0.8850

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

Table 7. Aligned Interests Treatment: Effect of Collective Punishment on Individual Public Goods Contributions

Dependent Variable: $(Contribution)_t$

	(1) OLS w/ PCSE, 1 lag in indep vars	(2) OLS w/ PCSE, AIC lag structure (up to 4 lags)	(3) OLS w/ PCSE	(4) Tobit w/ uncorrected SE
$(Contribution)_{t-1}$	0.448*** (0.133)	0.355* (0.198)	0.364* (0.197)	0.548*** (0.089)
$(Contribution)_{t-2}$	0.200* (0.116)	0.227 (0.186)	0.237 (0.189)	0.289*** (0.102)
$(Contribution)_{t-3}$		0.168 (0.159)	0.165 (0.158)	0.292*** (0.095)
$(Others' Contributions)_{t-1}$	0.123*** (0.029)	0.155*** (0.029)	0.155*** (0.029)	0.262*** (0.057)
$(Others' Contributions)_{t-2}$		0.046 (0.046)	0.044 (0.047)	0.031 (0.074)
$(Others' Contributions)_{t-3}$		-0.085* (0.046)	-0.085* (0.047)	-0.114 (0.076)
$(Others' Contributions)_{t-4}$		-0.049 (0.041)	-0.049 (0.041)	-0.020 (0.057)
$(In\text{-}Group Punishment)_{t-1}$	0.042 (0.066)		0.052 (0.098)	
$(In\text{-}Group Punishment)_{t-2}$			0.052 (0.080)	
$(Collective Punishment)_{t-1}$	0.100 (0.084)	0.336*** (0.109)	0.351*** (0.113)	0.421** (0.177)
$(Collective Punishment)_{t-2}$		0.120 (0.125)	0.130 (0.126)	0.145 (0.197)
$(Collective Punishment)_{t-3}$		-0.182 (0.126)	-0.187 (0.127)	-0.247 (0.191)
$(Collective Punishment)_{t-4}$		-0.216 (0.149)	-0.229 (0.149)	-0.217 (0.207)
Period fixed effects?	Yes	Yes	Yes	Yes
N	616	440	440	440
Individuals	55	55	55	55
R^2	0.6397	0.6559	0.6568	-

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

Table 8. Aligned Interests Treatment: Effect of Collective Punishment on In-Group Punishment Choices

Dependent Variable: *(Total In-Group Punishment Chosen)$_t$*

	(1) OLS w/PCSE, all subjects	(2) OLS w/ clustered SE, only period *t* max-in-group contributors	(3) OLS w/ clustered SE, only period *t* intermediate-in-group contributors	(4) OLS w/ clustered SE, only period *t* min-in-group contributors
(Total In-Group Punishment Chosen)$_{t-1}$	0.280(0.149)*	0.257(0.142)*	-0.027(0.103)	0.332(0.175)*
(Total In-Group Punishment Chosen)$_{t-2}$	0.253(0.146)*	0.172(0.066)**	0.355(0.162)**	0.403(0.181)**
(Total In-Group Punishment Chosen)$_{t-3}$	0.270(0.126)*	0.284(0.195)	-0.040(0.045)	0.241(0.178)
(Contribution)$_t$	0.012(0.015)	-0.045(0.070)	0.057(0.031)*	0.012(0.032)
(Contribution)$_{t-1}$	0.010(0.020)	0.067(0.045)	0.010(0.029)	-0.031(0.015)**
(Others' Contributions)$_t$	-0.029(0.010)***	-0.051(0.022)**	-0.030(0.014)**	0.003(0.023)
(Others' Contributions)$_{t-1}$	0.026(0.011)**	0.042(0.025)*	0.016(0.014)	0.001(0.020)
(In-Group Punishment)$_{t-1}$	0.018(0.024)	-0.011(0.036)	-0.003(0.012)	0.058(0.014)***
(Collective Punishment)$_{t-1}$	0.024(0.029)	0.049(0.065)	-0.050(0.037)	0.020(0.038)
Period fixed effects?	Yes	Yes	Yes	Yes
N	528	160	131	141
Indivs./Clusters	55	47	43	41
R^2	0.6464	0.4875	0.4196	0.8671

(***) corresponds to p<0.01; (**) corresponds to p<0.05; (*) corresponds to p<0.10

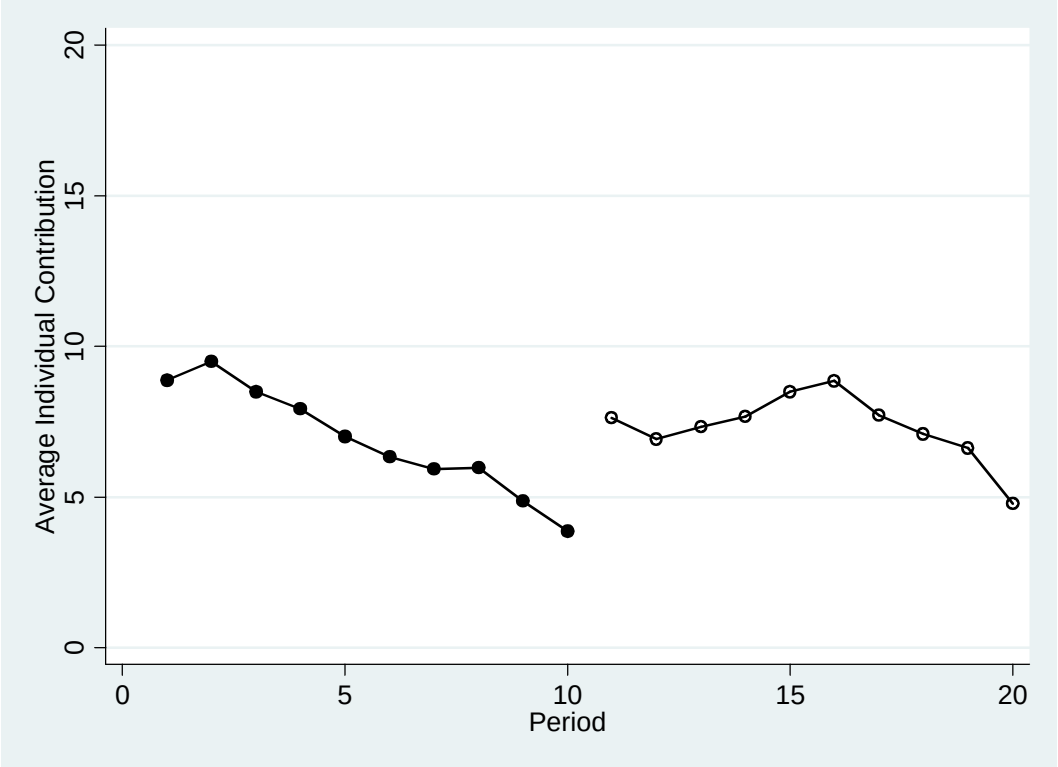

Figure 1. Average Individual Public Goods Contributions (Opposed Interests Treatment)

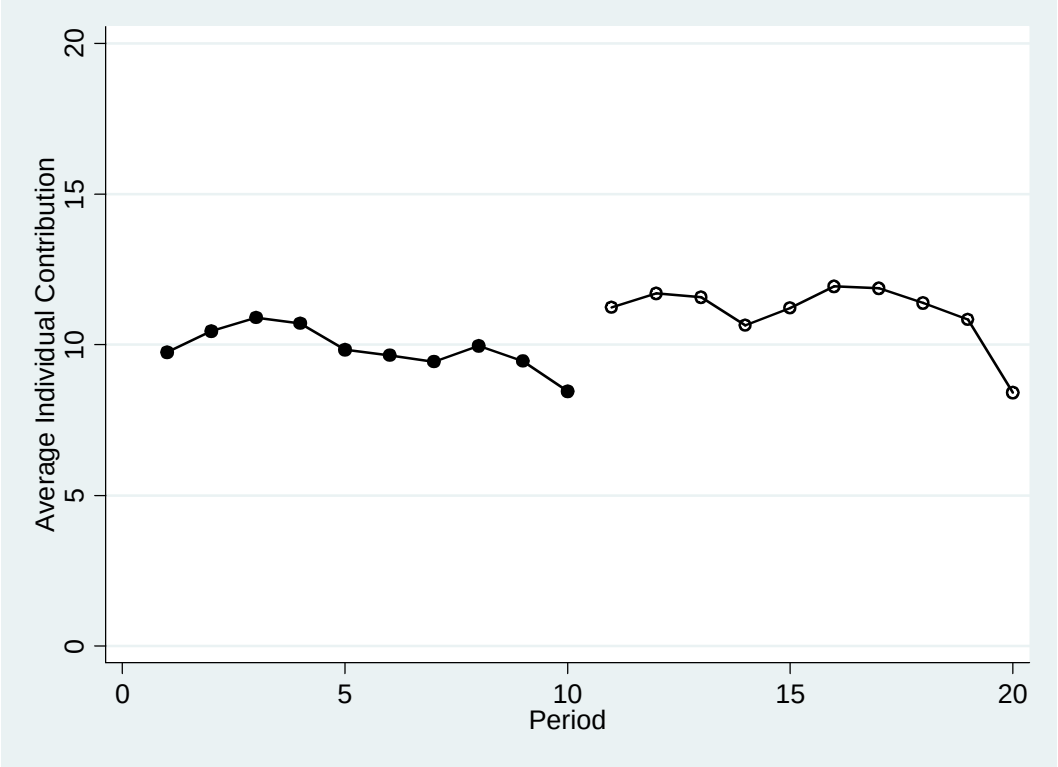

Figure 2. Average Individual Public Goods Contributions (Aligned Interests Treatment)

Online Appendix: Sample Instructions to Subjects

This is an experiment on decision making. In the following experiment you will make a series of choices. At the end of the experiment, you will be paid depending on the specific choices that you made during the experiment and the choices made by other people. If you follow the instructions and make appropriate decisions, you may make an appreciable amount of money. Please listen carefully to the instructions.

During the course of the experiment, you will have the opportunity to earn "tokens" that will be converted into dollars at the end of the experiment. The conversion rate is:

40 tokens = 1 dollar

You will be assigned, at random, into a group of five people. Within that group, you will also be randomly assigned to one of two roles in the experiment: Role A, or Role B. Within a given group, one person is randomly assigned to Role A; the other four people are assigned to Role B. The assignments will remain fixed for the duration of the experiment: that is, you will interact with the same group of other people, and remain in the same role, for the duration of the experiment. In addition, every person assigned to Role B will also receive an ID number – 1, 2, 3, or 4 – that will also remain fixed throughout the experiment. All of your interactions will be through the computer terminals at which you are sitting, and your true identity will never be revealed to any other person in the laboratory.

The experiment consists of **10** *periods*, all of which have the same structure. In each period, there are three separate *stages*. In the first stage of each period, each person in Role B will be given a supply of tokens, and must choose how many of these to allocate to a common pot and how many of these to keep for him- or herself. In the second stage of each period, each person in Role B will learn about the individual choices (listed by ID number) made in the first stage by the other people in Role B, and will then have an opportunity to decide whether, and by how much, to reduce the earnings of other individual group members in Role B. Finally, in the third stage of each period, the person in Role A will observe both the *total* number of tokens allocated to the common pot in the first stage by people in Role B, and also the *total* amount by which people in Role B reduced each others' earnings in the second stage. The person in Role A will then choose whether, and by how much, to reduce the earnings of all of the people in Role B.

This same process will be repeated in all **10** periods. A more complete description of this process now follows.

First Stage

At the beginning of each period, each person in Role B receives **20** tokens. Each person in Role B must then decide how many of the **20** tokens to allocate to a common pot, and how many to keep for him- or herself. Only a whole number of tokens (no decimals) can be allocated or kept.

The payoffs of a person in Role B in the first stage are composed of two parts:

- The number of tokens that person keeps for him- or herself.
- 0.4 times the number of tokens that *all* people in Role B allocate to the common pot (including tokens that person allocates him- or herself).

That is, the payoffs of a person in Role B in the first stage can be written as

first-stage payoffs for person in Role B = (tokens kept) + 0.4 × (total tokens allocated to common pot by people in Role B).

Although the person in Role A does not make a choice in the first stage, he or she also receives first-stage payoffs that depend on the choices made by people in Role B. These payoffs are composed of two parts:

- An automatic supply of **60** tokens
- -0.4 times the number of tokens that *all* people in Role B allocate to the common pot.

first stage payoffs for person in Role A = **60** - 0.4 *× (total tokens allocated to common pot by people in Role B).*

Therefore, every token kept by a given person in Role B increases that person's first-stage payoffs by one token (and does not contribute to the earnings of other group members). Every token allocated to the common pot increases the first-stage earnings of *every* person in Role B by 0.4 tokens, but also decreases the first-stage earnings of the person in Role A by 0.4 tokens.

Consider the following examples:

(1) Suppose that every person in Role B keeps all of his or her 20 tokens for him- or herself. Then the first-stage payoffs of each person in Role B will be equal to 20 + (0.4 x 0) = 20 tokens. The first-stage payoffs of the person in Role A will be equal to 60 - 0.4 *× (0)* = 60 tokens.

(2) Suppose that every person in Role B allocates all of his or her 20 tokens to the common pot. Then the first-stage payoffs of each person in Role B will be equal to 0 + (0.4 x 80) = 32 tokens. The first stage payoffs of the person in Role A will be equal to 60 - 0.4 *× (80)* = 60 – 32 = 28 tokens.

(3) Suppose that, in total, 30 tokens are allocated to the common pot by the people in Role B as a whole. Then each person in Role B receives 0.4 x 30 = 12 tokens from the common pot. A specific person in Role B who allocated 15 tokens to the common pot while keeping 5 for him- or herself would therefore have first-stage payoffs equal to 5 + (0.4 x 30) = 17 tokens. And another specific person in Role B who allocated 5 tokens to the common pot while keeping 15 for him- or herself would therefore have first stage payoffs equal to 15 + (0.4 x 30) = 27

tokens. The person in Role A would receive first stage payoffs equal to 60 - 0.4 $\times$ *(30)* = 48 tokens because a total of 30 tokens were placed in the common pot.

Second Stage

At the beginning of the second stage, everyone receives some information about what happened in the first stage.

- The person in Role A is told what his or her first-stage payoffs were, and how many tokens were allocated in total to the common pot by the players in Role B. The person in Role A learns *only* about the *total* number of tokens allocated to the common pot; the person in Role A never learns about the specific choices made by specific group members.
- A person in Role B is told:
  - His or her total first-stage earnings
  - How many tokens in total were allocated to the common pot by all the people in Role B
  - How many tokens each of the other people in Role B specifically allocated to the common pot, listed by ID number (1, 2, 3, or 4).

Once this information has been received, each person in Role B must decide if, and by how much, he or she wishes to reduce the first-period payoffs of each of the other people in Role B, as identified by their ID numbers. If a person in Role B wishes to do this, the cost is **1** token for each token reduced from the payoffs of any other person. A person in Role B can spend as much as he or she likes, up to his or her payoffs in tokens from the first stage, towards reducing the payoffs of others in Role B. Only a whole number of tokens (no decimals) can be reduced from the payoffs of another person.

The payoffs of a person in Role B after the second stage can therefore be written as

second-stage payoffs for person in Role B = first stage payoffs – (tokens spent reducing others' payoffs) – (number of own tokens reduced by others)

The person in Role A does not make a choice in the second stage, and his or her payoffs are not directly affected by the choices made in the second stage by people in Role B.

Third Stage

At the beginning of the third stage, everyone receives some information about what happened in the second stage.

- A person in Role B:
  - Is told the *total* number of tokens by which other people in Role B reduced his or her earnings. He or she is never specifically informed of *which* other people in Role B (if any) reduced their payoffs, or the specific amounts by which individuals in Role B reduced their payoffs.
  - Is reminded of how much he or she spent reducing others' payoffs.
  - Is told her second-stage payoff.
- The person in Role A:

- Is told how many tokens in *total* people in Role B reduced from each others' first-round earnings. As before, the person in Role A never learns about the specific choices made by individuals in Role B.
- Is reminded how many tokens in *total* were contributed to the common pot by people in Role B in the first stage.
- Is reminded of his or her payoffs from the first stage.

Once this information has been received, the person in Role A must decide if, and by how much, he or she wishes to reduce the payoffs of people in Role B. The person in Role A cannot reduce the payoffs of different individuals in Role B by different amounts; if the person in Role A wishes to reduce the earnings of people in Role B, an *identical* amount is subtracted from the payoffs of *every* person in Role B. The person in Role A simply chooses what this amount will be. There is no cost to the person in Role A for reducing the payoffs of the people in Role B. The person in Role A can reduce the payoffs of the people in Role B by as much as he or she likes, up to a total **5** tokens from each person in Role B. Only a whole number of tokens (no decimals) can be reduced from the payoffs of all the people in Role B.

The people in Role B will learn what the person in Role A has chosen to do in the third stage as soon as a decision has been made.

Summary of Net Earnings for a Period

For a person in Role B, the following is calculated:

- First-stage payoffs (from common pot and from tokens kept)….
- ….MINUS tokens spent reducing payoffs of others in Role B (second stage)
- ….MINUS own tokens reduced by others in Role B (second stage)
- ….MINUS tokens reduced by the person in Role A (third stage)

If the result of this calculation is greater than zero, the amount will be the Role B person's net payoffs for the period. If the result is less than or equal to zero, the Role B person's net payoffs for the period will simply be set equal to zero.

For the person in Role A, the following is calculated:

- First-stage payoffs (from common pot and from automatic token supply)

The result is the net payoffs for the period for the person in Role A.

Conclusion

This concludes the description of the choices that are made and the payoffs that are earned in one period. This process will be repeated until all of the **10** identical periods are completed. Your final earnings will be equal to the sum of your payoffs in each of the **10** periods. Remember that you will interact with the same group of other people, and remain in the same role, throughout this process.

We ask everyone to remain silent until the end of the last period and then to await further instructions. If you have any questions, please ask them at this time.