



# Endogenous choice of institutional punishment mechanisms to promote social cooperation

Anabela Botelho<sup>1</sup> · Glenn W. Harrison<sup>2,5</sup> · Lígia M. Costa Pinto<sup>3</sup> · Don Ross<sup>4,5,6</sup> · Elisabet E. Rutström<sup>6,7</sup>

Received: 10 November 2020 / Accepted: 7 December 2020 / Published online: 3 January 2021  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

## Abstract

Does the desirability of social institutions for public goods provision depend on the extent to which they include mechanisms for endogenous enforcement of cooperative behavior? We consider alternative institutions that vary the use of direct punishments to promote social cooperation. In one institution, subjects participate in a public goods experiment in which an initial stage of voluntary contribution is followed by a second stage of voluntary, costly sanctioning. Another institution consists of the voluntary contribution stage only, with no subsequent opportunity to sanction. In a third stage subjects vote for which institution they prefer for future interactions: do they prefer one that does allow sanctions or one that does not allow sanctions? Our results show that even though sanctions are frequently used when available, the clear majority of individuals vote for the institution that does not allow sanctions. Thus, a distinction is required between the principles that guide the *choice of institutions* and the principles that apply to actions *guided by institutions*. Our results indicate that it is the wealth generated by the institution that determines its desirability.

**Keywords** Public goods · Social institutions · Voting · Sanctions · Laboratory experiments

**JEL Classification** H41 · D72 · C92

## 1 Introduction

A growing consensus in the experimental literature suggests that institutions allowing the use of voluntary punishments can reduce the free-rider problem in public goods games. That behavioral phenomenon has been studied using variations of an experimental design introduced by Ostrom et al. (1992) and Fehr and Gächter (2000, 2002). The *inference* often drawn from those findings is that enforcement mechanisms tend to be welfare-enhancing to the extent that they align incentives.<sup>1</sup> We offer a different perspective on the value of institutions by constructing laboratory experiments wherein we *elicit institutional preferences*

<sup>1</sup> The efficacy of punishments is questioned in some studies. Gintis et al. (2005) collect many perspectives on the existence and behavioral role of reciprocity. Simonsohn (2006) provides a thoughtful critical review, noting in conclusion that “one of the challenges for social preference research is the abundance of theories

✉ Glenn W. Harrison  
gharrison@gsu.edu

Extended author information available on the last page of the article

*directly*<sup>2</sup> from individuals experienced in institutions with voluntary punishment and institutions without voluntary punishment. Our analysis of their expressed preferences explicitly recognizes that imposing punishments is costly, such that the value of an institution may not be correlated positively with its contribution to reducing free riding. We find that institutional preferences are not mysterious: subjects generally are motivated by self-interest when choosing institutions.

Gürerk et al. (2006) demonstrate that societies with rules that allow the voluntary imposition of costly sanctions, and that generate aggregate wealth *advantages* over societies without such rules, can grow over time at the expense of the “more anarchic” societies. On the other hand, in experiments with subjects who cannot switch between alternative institutional structures, we sometimes see voluntary sanctions leading to aggregate wealth *disadvantages* relative to those that do not allow sanctions. We review the data from Fehr and Gächter (FG, hereafter) (2000, 2002) and find that the same conclusion is the case in their experiments. Many other studies report detrimental effects on earnings from unrestricted sanctions. (Carpenter and Matthews 2004; Page et al. 2005; Sefton et al. 2007; Casari and Luini 2009; Anderson and Putterman 2006; Nikiforakis 2008; Nikiforakis and Normann 2008; Egas and Riedl 2008). We implement a new experimental design that allows us to identify the independent influences on institutional preferences of the institution’s incidence of free riding and its profitability.

We find that *none of our laboratory societies had a majority that voted to live in a world with sanctions* when we implemented an environment with no reputational effects. That result is robust to the order in which participants experienced the alternative institutions before voting and parametric variations in the opportunity cost of free riding. The vote is not even close, and in one case it is unanimous. When we allow for some reputation effects by using a random-strangers matching protocol, we find that *the majority votes for the world with sanctions in only one in nine of our laboratory societies*. Our findings suggest that the conditions under which a group or a society would choose a rule allowing for voluntary costly sanctions depend on its relative profitability and not on its ability to solve the free-rider problem. A strong negative correlation exists between the vote for sanctions and the loss in profits that the institution caused. Other motivations, such as fairness, likewise may have played a role, but effects on wealth are a key candidate for what motivated actions.<sup>3</sup>

---

Footnote 1 (continued)

that are often hard to tease apart empirically. Loosely applying the new term ‘strong reciprocity’ to phenomena that can be accounted for by preexisting theories is counterproductive”.

<sup>2</sup> That is, the set of rules, combined with their enforcement mechanisms, that constrain the choices of individuals.

<sup>3</sup> Consider a world of sanctions in which a majority of subjects gained more wealth on average than they would in a non-sanctions world, but a minority of subjects earned virtually nothing. Average gain is larger with sanctions, and for a majority, but one easily could imagine that some individuals in the majority might not want to live in such an inequitable world and would vote against it.

## 2 The value of a punishment rule

Relying on an established literature arguing that punishments can sustain cooperation beyond that achieved by other social norms, it is a natural extension to ask if the preference over such institutions depends on the earnings consequences that are generated.<sup>4</sup> More precisely, what is the net value of allowing the punishment technology that generates the cooperative behavior endogenously?

In the experiment of FG (2000) subjects play a voluntary contribution (VC) game over 20 periods, wherein one set of 10 periods they do not have the option to punish, but in another set they do. FG vary the ordering of the two within-subject treatments. Two between-subjects' treatments are implemented based on how subjects are matched into groups of four. In one set, FG employ a partners design in which the same subjects are matched throughout the full 20 periods; in another they use a random strangers design wherein subjects are rematched into new groups before the start of each round. In the VC game, all subjects are given an initial endowment of tokens, and they can choose to keep them or to invest them in a project. The private return on the tokens invested in the project is less than their value if kept, but all subjects are paid the return from the combined investment in the project, thus generating an efficient cooperative outcome that is not the Nash equilibrium of the game. In the punishment stage, each group member can send punishment points to any other group member. Punishment points reduce the receiving group members' earnings from the VC game by a proportional factor, but they also are costly to the sender. The unique Nash equilibrium prediction is for nobody to invest in the project or to send any punishment points.

The top panels of Fig. 1 display the relative gains from allowing sanctions, based on the data generated by FG (2000).<sup>5</sup> In the first series, subjects experienced ten periods with sanctioning after an initial ten periods without and in the second series that ordering was reversed. The earnings shown in the top panels of Fig. 1 correspond to a pattern of contributions in the public goods game that converges to the cooperative outcome in periods 9 or 10, as shown in FG (2000). But the accumulated cost of punishments on the convergence path in periods 1 through 8 more than offsets the incremental gains in periods 9 and 10. The aggregate loss is 12.5% in the first series, and 17% in the second series.<sup>6</sup> The experiment might trigger activation of pre-existing normative expectations in subjects that lead them to implement punishment. However, that possibility would depend on subjects integrating their model of the laboratory game into a larger “game of life,” in the sense of Binmore (1994, 1998), in which they model themselves as engaged.

The top panels of Fig. 2 report comparable calculations for the experiments in FG (2002). There, the design was similar to the strangers treatments in FG (2000), although the punishment cost schedule was linear rather than convex in punishment points. Each

<sup>4</sup> We take issue with the sense in which the previous literature in question actually identifies a social norm and discuss that issue in Sect. 4.

<sup>5</sup> We consider only the “strangers” design in FG (2000), since it controls for the possible role of strategic self-interest in imposing the sanctions. In their “partners” design, the same subjects played against each other for ten periods; in their “strangers” design, individuals were assigned randomly to groups after each period. FG (2000) considers only strangers designs. We are grateful to Simon Gächter for providing the data from their experiments.

<sup>6</sup> For example, average profits in the second series were 22.73 currency units and 18.85 currency units, respectively, without and with the institutional punishment mechanism: a difference of 3.88 currency units or 17.1%.

treatment consisted of only 6 periods. The results in Fig. 2 are similar to those in Fig. 1, but even more striking in terms of the persistent costliness of punishment. In each series the aggregate loss in value from implementing the punishment is roughly 15%. Moreover, perhaps owing to the shorter time horizon of the experiment, no strong indication exists that extrapolating beyond the horizon of the experiment into a “game of life” would generate a positive net value.

The foregoing experiments show that sanctions can have strong effects on cooperation. What is less clear, however, is the extent to which the earnings effect of the sanctions is perceived as favorable by the participants. More to the point, would the participants in any of these experiments want to have the punishment mechanism available if they were to participate in a social choice after their experience? Since the FG experiments were not designed to answer that question, we can only speculate about such choices based on their data. The results in the bottom panels of Figs. 1 and 2 consider this question, using two possible voting rules<sup>7</sup> for social choice:

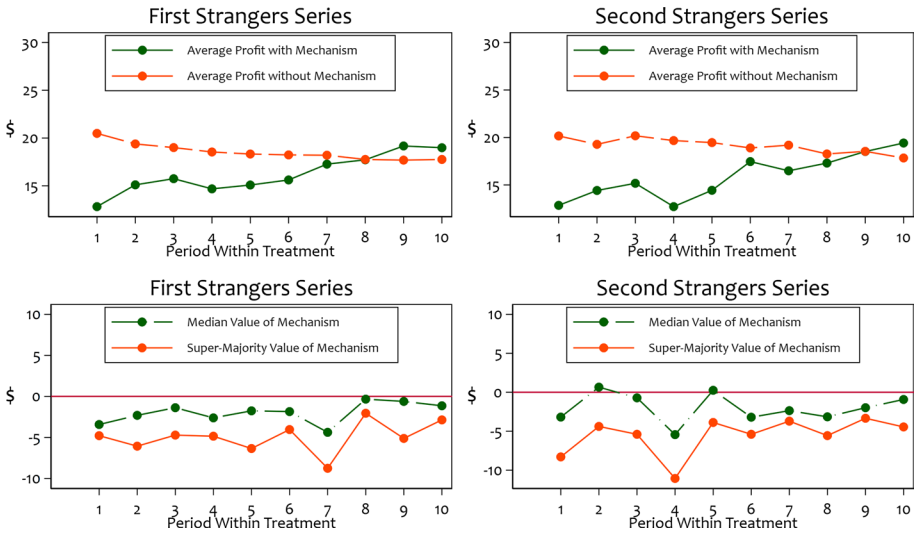
- Majority rule referendums—would the median voter opt for the social technology?
- Super-majority rule referendums—would 67% of the population vote for the punishment mechanism as an institutional rule?

One particularly nice feature of the FG (2000, 2002) design is that it allows in-sample comparisons of the value of the mechanism to each individual subject. Each subject participated in both conditions, so it is a simple matter to calculate the earnings for each subject with and without the mechanism. From the distribution of within-subject net profits, so calculated, one can determine the period-wise median and 33<sup>rd</sup> percentile. The results are shown in the bottom panels.

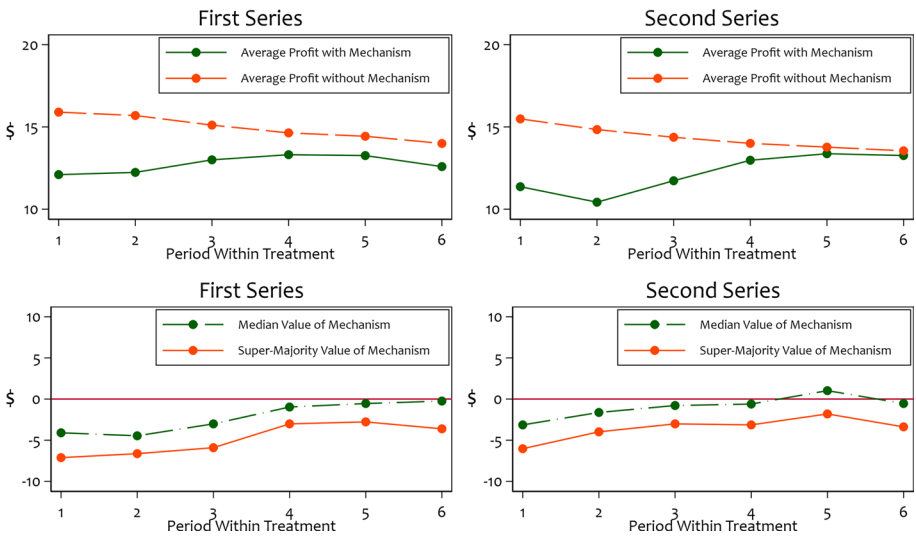
The implication from Figs. 1 and 2 is that, with two exceptions, *the institutional punishment mechanism would not be adopted under either of those social choice criteria*.<sup>8</sup> Furthermore, it is much harder to argue a priori that simple extrapolation into a “game of life” beyond the laboratory provides any basis for predicting that the institution would be chosen socially under either criterion. Our experiment was designed to investigate the question of social choice directly, by allowing participants to vote over institutional mechanisms after they have experienced their effects. The design allows us simply to observe the choices made by the subjects and then infer whether they seem to be extrapolating into “games of life” or not.

<sup>7</sup> The results in the top panels imply what would happen if a classical utilitarian social choice rule was employed, in which aggregate benefits were compared to aggregate costs. Over the life of the experiment, the institutional punishment mechanism would not be approved. However, it would be approved if one were just to use the results of the last period or two to calculate benefits or costs in the experiments of FG (2000) (Fig. 1).

<sup>8</sup> The exceptions are period 9 of the first series in Fig. 1 and period 5 of the second series of experiments in Fig. 2, where the median voter would just vote *for* the institutional punishment mechanism. The mechanism would not survive a constitutional referendum using a super-majority rule in those periods.



**Fig. 1** Value of the institutional punishment mechanism in *AER* experiments. Data from Fehr and Gächter (2000)



**Fig. 2** Value of the institutional punishment mechanism in *Nature* experiments. Data from Fehr and Gächter (2002)

### 3 Voting for a punishment institution

#### 3.1 Basic experimental design

We design a simple experiment to test whether subjects would choose to live in a world with mechanisms for costly sanctions. In the first part of the experiment, we replicate the design of FG (2000) by providing subjects with experience in public goods contribution games in which a punishment mechanism is in place and also in games in which no such mechanism is available. We examine order effects as FG do by running one set of subjects through experiments in which the punishment option comes first and then the no-punishment option is experienced. A separate set of subjects plays the same game, but in the reverse order. We allow 10 periods in each setting, so that each subject plays 20 periods prior to the vote.

To ensure that no confounding reputation effects arise, and to provide the cleanest possible test, we adopt a perfect strangers design in which no subject ever meets the same subject more than once. Virtually all previous public goods experiments rely on a random strangers design in which subjects are re-paired randomly every period.<sup>9</sup> Although that design reduces the chance that a subject will meet the same person dramatically, and is coupled with anonymity, the critical behavioral issue is whether the subjects believe that no reputation effect is associated with their choices in a given round. In a perfect strangers design, it is clear that subjects should hold a belief that the probability of ever meeting any other player again is a zero. In our experiments, subjects participate in groups of 2 in each round.<sup>10</sup> We explain carefully to them how we ensure that they have no chance of meeting the same person in any other round. Since ours is a departure from previous experimental practice, we also implement between-subject controls to see the effect of using a random strangers design instead of a perfect strangers design. We vary the cohort size in the random strangers design in an effort to vary the reputational conditions.<sup>11</sup> Subjects are matched randomly into pairs from within the same cohort.

---

<sup>9</sup> Andreoni and Croson (2005) review the literature on public goods contributions with partners and strangers. FG (2000, fn.3) report that the results of a perfect strangers replication of their design generated essentially the same results as their ordinary strangers experiments. However, they considered only one sequence of regimes (punishment followed by non-punishment), and did not maintain the perfect strangers treatment after the first regime of 6 periods. Compared with a random strangers design, Botelho et al. (2009) found a statistically significant negative effect of the perfect strangers design on subjects' propensity to contribute to the public good in experiments using both four subjects per group and two subjects per group. Rather than debate whether any of such comparisons are conclusive, we prefer to ensure the control against any reputational effects afforded by a perfect strangers design. What we call a perfect strangers design is the same as "turnpike matching" introduced by Cooper et al. (1996) and discussed by Fréchette (2012).

<sup>10</sup> Most public goods experiments use four subjects per group, although the effect of larger group sizes has been studied by Isaac and Walker (1988) and others. Harrison and Hirshleifer (1989) and Goeree et al. (2002) employed groups of 2 in their public goods experiments. Carpenter (2007) examines the interaction between punishment and group size, finding that, as in punishment-free settings, larger groups tend to elicit more contributions, but that the logistics of larger groups restricts the ability of punishment to discipline free riders.

<sup>11</sup> We vary the cohort size in the random strangers sessions from a smallest size of 6 to a largest size of 16. Two cohorts were present at the same time in a session, so the group size was a salient feature of the design. Subjects were given clear instructions on the sizes of their respective cohorts. When more than one cohort was present, the text of the instructions was changed to reflect the fact that the voting outcome was implemented separately for each cohort based on that cohort's vote.

After period 20 we ask subjects to vote on the environment they would like to participate in for one “Final Jeopardy!” round.<sup>12</sup> The instructions they received in one of the treatments are as follows:

We are now ready for your final task. This will consist of only one period. The task will be a repetition of one of the two tasks you have just completed. Which task this will be will be determined by a common vote in a moment. In this one period the stakes will be increased so that each token is now worth 50 cents, not just 5 cents. This is therefore 10 times the value that a token has had in each of the earlier periods. Before you play out this one period, you will be asked which environment you would like to participate in. You may choose either the one where you can reduce other participants’ earnings and they can reduce yours (**environment B**) or you may choose the environment in which there is no such opportunity (**environment A**). Everyone will be asked to vote for the environment that they prefer, and **we will implement the environment that a majority of the participants in this room vote for**. Thus, we will implement the same environment on all matched pairs.

In the event of a tied vote we will roll a ten-sided die for you all to see. If the die comes up 0–4 we will implement environment A, where earnings reductions are not available, if it comes up 5–9 we will implement environment B, where earnings reductions are available. Before you are asked to vote you will be shown a screen with a review of your earnings across the periods in both of the environments.

One variation of these instructions simply reverses the references to environment B and environment A since the order of the two in the first part of the experiment was reversed. Another variation is that in the random strangers design a voting outcome applies to a cohort rather than to everyone in the room. The matching protocol employed in the first part continues in the last period, so that in the perfect strangers design subjects once again meet somebody they have not met before. Once a decision is made, all subjects play the chosen environment for one period. In order to enhance the *relative* saliency of the voting decision, which is the main focus of our design, we tell subjects that their earnings in the last period will be ten times those of each of the first 20 periods. The one-shot design of the final round is precisely the environment that the earlier rounds are attempting to model, although the earlier rounds allow learning to occur over time. The question of interest, as in FG (2000, 2002), is whether punishments will be used in such anonymous one-shot environments, and what effect they then have on behavior.

Table 1 summarizes the experimental design. Thirteen sessions were conducted. The first 4 used perfect strangers designs, and the last 9 used random strangers designs. The return to the public good is discussed below, as are the votes.

### 3.2 Parameters and treatments

Parameters must be chosen carefully, and our parameter values are very similar to those used in FG (2000). All earnings and costs were presented to subjects as “tokens,” and they were told upfront that we would pay them 5 cents for every token they had at the end of the experiment.

<sup>12</sup> The popular TV game show *Jeopardy!* has three rounds of play: “Jeopardy!”, “Double Jeopardy!” and “Final Jeopardy!” The first two consist of three categories of three questions each, but “Double Jeopardy!” multiplies dollar values by two. “Final Jeopardy!” poses only one question and subjects can wager up to their accumulated earnings in that round.

In one treatment a relatively low return on contributions to the public good was offered and in another treatment a relatively high return was available. The low return was 0.6 of a token: every token contributed to the public good by one subject would reduce their private endowment by 1 token and return 0.6 of a token to him or her. Of course, it also would generate 0.6 of a token for the other player, so the social return was 1.2 tokens for every 1 token invested. In the high return treatment, we changed the public good return from 0.6 to 0.8, thereby increasing the social return from 20 to 60%. The objective of that treatment was to see the effects of making the environment more rewarding in the context of some other influence, such as a normative expectation imported from a “game of life” that would increase contributions to the public good. Table 1 shows that the low return was adopted in sessions 1 and 2, and the high return in all other sessions. We used a linear payoff schedule constant for all contributions, so the dominant strategy is simple: a subject who seeks only to maximize individual earnings in a single period should contribute nothing to the public good.<sup>13</sup>

In the punishment stage, each point allocated to punish the other player implied a 10% reduction in the other player’s earnings in that round. The cost to the subject inflicting the punishment is shown in Table 2. Each subject received an endowment of 20 tokens at the outset of each round; in addition, subjects received a one-time endowment of 25 tokens to cover possible losses. As Table 2 shows, *each* subject in *anyone* period was allowed to buy up to 9 punishment points without incurring a loss in that period (and before factoring in any profit from provision of the public good or the private good).

### 3.3 Procedures

We recruited 180 subjects from the University of Central Florida (UCF) in 2005.<sup>14</sup> Subjects were assigned randomly to each session, with no prior knowledge of the parameters or treatments. All sessions were conducted at the Behavioral Research Lab of UCF’s College of Business Administration. The facility is a standard, computerized laboratory: each station has a “sunken” monitor, and we employed personal “cubicle-style” screens to ensure even more privacy. Instructions were provided in written form and orally, and the experiment was implemented using version 2.1.4 of the *z-Tree* software developed by Fischbacher (2007). The same experimenter (Rutström) delivered the oral instructions for all sessions, to ensure comparability.<sup>15</sup> The oral instructions also utilized a large-screen display that easily could be seen by all subjects, to ensure that certain information was common knowledge. Training rounds were administered prior to each session to ensure that subjects understood the task.

Average earnings in the experiments were \$39, including a standard \$5 show-up fee. No session lasted more than 2 h, and most were at least 1½ h in length.

<sup>13</sup> Alternative assumptions about the factors motivating subjects to contribute in public goods’ experiments have long been studied. See, in particular, Palfrey and Prisbrey (1996, 1997) and Goeree et al. (2002).

<sup>14</sup> UCF is located in Orlando, Florida. It has a large student body, with a Fall 2004 enrollment of 42,837. The entering class in 2004 had an average SAT of 1186. The student body also is ethnically diverse: in 2004 8.5% stated that they were non-Hispanic black; 70% stated that they were non-Hispanic; white; 5.0% stated that they were Asian; and 12.2% stated that they were Hispanic.

<sup>15</sup> A digital recording of the oral instructions in one typical session is available in the data archive at <https://cear.gsu.edu/gwh>.



**Table 1** Experimental design

Session	Return to public good	Matching	N in session	History	Average profit per period			Vote for punishment (%)
					NP	P	NP	
1	Low	Perfect	26	NP-P	\$1.01	\$0.96	0	
2	Low	Perfect	24	P-NP		\$0.89	\$1.01	21
3	High	Perfect	26	NP-P	\$1.25	\$1.12	8	
4	High	Perfect	26	P-NP		\$1.22	\$1.16	42
1–4	Low or high	Perfect	102					18
5	High	Random	10	P-NP		\$1.26	\$1.05	60
6	High	Random	16	P-NP		\$0.98	\$1.08	19
7	High	Random	8	NP-P	\$1.34	\$1.34	25	
8	High	Random	6	NP-P	\$1.30	\$1.25	50	
9	High	Random	6	NP-P	\$1.28	\$1.06	0	
10	High	Random	8	NP-P	\$1.21	\$1.19	12	
11	High	Random	10	NP-P	\$1.42	\$1.44	40	
12	High	Random	8	P-NP		\$1.29	\$1.29	12
13	High	Random	6	P-NP		\$1.23	\$1.16	33

Each experiment had 10 rounds of one institution, followed by 10 rounds of the other institution

After round 20, all subjects voted on the institution for round 21

Round 21 had 10 times the payoffs of each of rounds 1–20

NP is the institution without voluntary punishment (no-punishment)

P is the institution with voluntary punishment (institutional punishment mechanism)

### 3.4 Observed outcomes

Table 1 shows the vote in each session, which is our “bottom line” result: when the chance of ever meeting any other person again is zero, in the perfect strangers design, no cohort voted for the institutional punishment mechanism. Overall, only 18% of participants in the perfect strangers treatment voted for that mechanism. The vote was close in one of the four perfect strangers sessions, but little doubt arose in the other three. In fact, in one session, all 26 subjects agreed unanimously to implement the institution that did not entail punishment. That result was robust to high or low returns to the public good and the order in which subjects experienced the institutions with or without punishment prior to voting.

Our results replicate the finding reported in FG (2000, 2002) that punishments lead to higher contributions on average. With punishments, the average token contribution was 7.42 and without punishment it was 5.53, which is significantly different according to a Wilcoxon–Mann–Whitney test with a  $p$  value below 0.001. Nevertheless, in our experiments, contributions decline over time even with punishments, and we therefore do not see the slight increase in profits over time reported in FG. We show the pattern of contributions and profits for each of our sessions in an appendix of Botelho et al. (2019). The joint significance of the observations here and in FG is that the use and effects of punishments

**Table 2** Punishment schedule

Points	0	1	2	3	4	5	6	7	8	9	10
Reduction of other person's earnings	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Cost to you of these points in tokens	0	1	2	4	6	9	12	16	20	25	30

vary across different groups of subjects; one therefore cannot say that they uniformly have a sustained positive effect on contributions, much less on profits.<sup>16</sup>

Figure 3 provides detailed results for session 1 to illustrate the outcomes. That was the session returning a unanimous vote against punishments. The top panel shows average token contributions in each period, and the bottom panel shows average dollar profits in each period. Since a punishment mechanism (P) operated in periods 11 through 20, we show pre-punishment profits as well as post-punishment profits. Of course, the latter were the “take home” profits to subjects, and the ones by which they are assumed to be motivated. In terms of contributions, we observe a now standard pattern in voluntary contribution experiments: subjects start out making some contributions, and then free riding sets in. Session 1 almost collapsed to complete free riding, which is more extreme than our other sessions, but the decline was general. After round 10, a “re-start” effect is observed, which also is a common behavioral effect, although not a universal one. We do not see that the punishment mechanism leads to sustained contributions in this particular session, although in some of our other sessions the results are more encouraging.

In terms of profits, the outcome in periods 11–20 for the institutional punishment mechanism is striking. The pre-punishment profits of subjects roughly were comparable to the profits earned in periods 1–10, but the post-punishment profits were much lower. The reduction is particularly evident in the first 4 periods of the punishment mechanism, with many subjects exercising their ability to punish others. If one compares the average profit in periods 1–10 with the average post-punishment profit in periods 11–20, it is not hard to see why every subject voted for the institution with no punishment mechanism (NP).

The results from session 1 are extreme; they nevertheless illustrate the factors underlying each vote. One could argue that the vote was stacked against the punishment mechanism in session 1 by it being second, when the standard decay in contributions had set in. But a counterargument is that it is precisely in such a setting where the punishment mechanism might be of value, since nobody needs a punishment mechanism if everyone is contributing heavily. And, of course, we test for such order effects from the sequencing of the two institutions. One also could argue that the vote was stacked against the punishment mechanism by the return to the public good being low, but again a counterargument would be that that is precisely when one needs some external device to get people to contribute, since the intrinsic returns are not high. We also considered higher returns to the public good in sessions 3 through 13.

<sup>16</sup> Such variability across groups of subjects also has been reported by Herrmann et al. (2008) implementing the experimental design of FG (2002) in different countries.

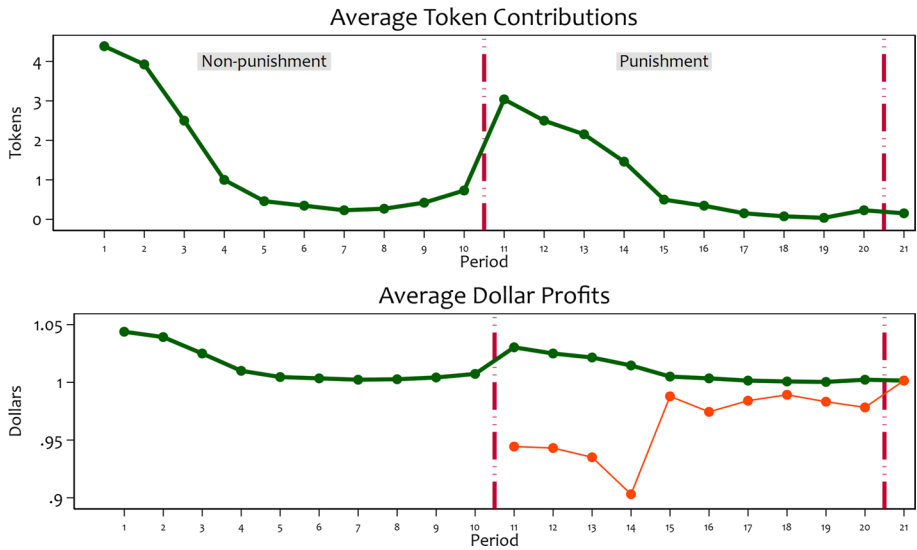
For completeness, we also show in Fig. 3 the average contributions after the vote, in period 21. The profits for that period were ten times the profits for each of the prior rounds, to increase the salience of the vote, but we display scaled-down levels of profits for comparability. An appendix of Botelho et al. (2019) reports similarly detailed outcomes for each of the other sessions.

Figures 4, 5, 6, 7, 8 and 9 show the average “take-home profits” in each period and session, along with the vote shares for the institutional punishment mechanism.<sup>17</sup> For comparability, each has the same vertical scale.

Figure 4 shows the results for sessions 1 and 3, which shared the same NP-P history and the perfect strangers design but differed in terms of the return to the public good being low or high. The unanimity of session 1 has been noted, but here we also see that only 8% of the subjects in the high return session 3 voted for the institutional punishment mechanism. In that case the contributions to the public good were relatively high in periods 1–10, were still around 7 or 8 tokens by period 10 and declined very slowly in periods 11–20. That is exactly what one would expect from the change from low returns to high returns to the public good, which is the only difference between the two sessions. The use of punishment in periods 10–20 of session 3 was relatively sparing. FG noted that some subjects also engaged in so-called “spiteful punishment.” Such punishment is said to occur when someone who was a free rider punishes a contributor and is extremely costly for the cohort. In those sessions we found very little “spiteful punishment” occurring. However, the punishment that did occur, along with the continued slow decay in contributions over time, resulted in take-home profits for session 3 that were systematically lower than those in the institution with no punishment mechanism.

Figure 5 shows the results for sessions 2 and 4, also perfect strangers sessions, which shared the same P-NP history and differed in terms of low or high returns to the public good. We again see a marked difference in contributions, with the change in the return to the public good, across both institutions. Round 1 deserves comment, since we see a dramatic reduction in take-home earnings in both sessions, owing to extravagant use of the punishment mechanism. We conjecture that Round 1’s outcome can be explained by some subjects learning about the nature of the punishment mechanism “the hard way.” In one session, one subject asked the experimenter privately, “if I punish the other person, do I get their earnings?” Of course, the question had been answered in the instructions, but as every experimenter knows some subjects always gloss the written and oral instructions, or do not trust them, and use the actual session to try things out. It also should be noted that we conducted two periods of non-paid training prior to each session. Nonetheless, the behavior in period 1 in sessions 2 and 4 (and sessions 5 and 6, discussed below) is consistent with that conjecture. The fact that the reduction stopped being so dramatic after round 1 is consistent with the subjects learning the rules of the game, as distinct from experimenting with the right dose of punishment (as one observed in periods 11–14 of session 1, shown in the bottom panel of Fig. 3). Nonetheless, the two sessions provided a stronger vote in favor of the institutional punishment mechanism than the other two perfect strangers sessions. Compared to sessions 1 and 3 (Fig. 4), the major change is the sequence of the institutions, with the institution of voluntary punishment being experienced first. In session 2, average take-home profit under the institutional punishment mechanism was consistently around 85 cents or 90 cents after the bloodbath of period 1, but average profit was steadily just

<sup>17</sup> Figures 6 and 7 contain experiments of the same general type but conducted in different physical sessions. The same is true of the experiments in Figs. 8 and 9.



**Fig. 3** Results in session 1.  $N=26$  perfect strangers in groups of 2 low return to public good

above \$1.00 for the institutions with no punishment in rounds 11–20. Thus, only 21% of the subjects voted for the institutional punishment mechanism. We undertake a formal statistical analysis of individual votes below, to see if the subject’s personal history influenced the vote. That is, even if average profits were lower for all subjects under the institutional punishment mechanism in session 2 than on the no-punishment institutions, maybe they were higher for those 21% that voted for the institutional punishment mechanism.

Session 4 was a voting cliff-hanger. Contributions started out relatively high, and apart from another period 1 bloodbath, the punishment was relatively efficient<sup>18</sup> and non-spiteful. Average contributions actually increased from around 10 tokens in period 1 to 11 or 12 in periods 4–10, with take-home profits around \$1.25 after period 2. The happy bubble crashed in period 11, with a dramatic fall in contributions. However, free riding did not take over completely, subjects continued to contribute around 5 tokens per period on average, and profits averaged about \$1.16 in periods 11–20. When the vote came, 42% voted for the institutional punishment mechanism.

Figures 6 and 7 show the results for sessions 5, 6, 12 and 13, all random strangers sessions sharing the same P-NP history and high returns to the public good. They differ in terms of the number of subjects in each cohort. In session 5, we drew randomly from  $N=10$ , in session 6 we drew randomly from  $N=16$ , in session 12 we drew randomly from  $N=8$ , and in session 13 we drew randomly from  $N=6$ . Subjects were made aware of the sizes of their cohorts. That feature provides a nice bridge between the complete absence of re-encounters in the perfect strangers design and the perfect rematching in the partners

<sup>18</sup> We refer heuristically to “efficient punishment” as occurring when it is targeted at individuals who have exhibited free-riding behavior, with a presumed goal of encouraging them to help the group return to the socially efficient outcome of no punishment.

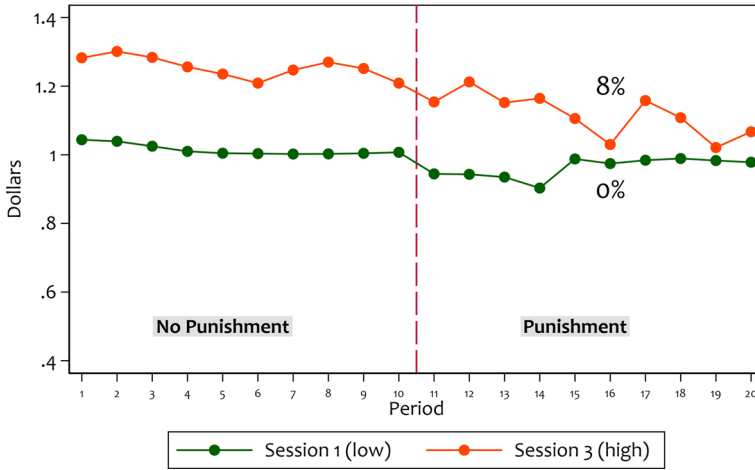


Fig. 4 Average profits with perfect strangers and NP-P history

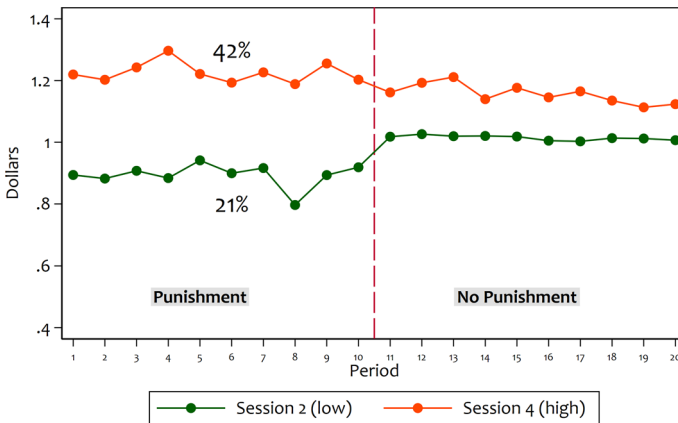
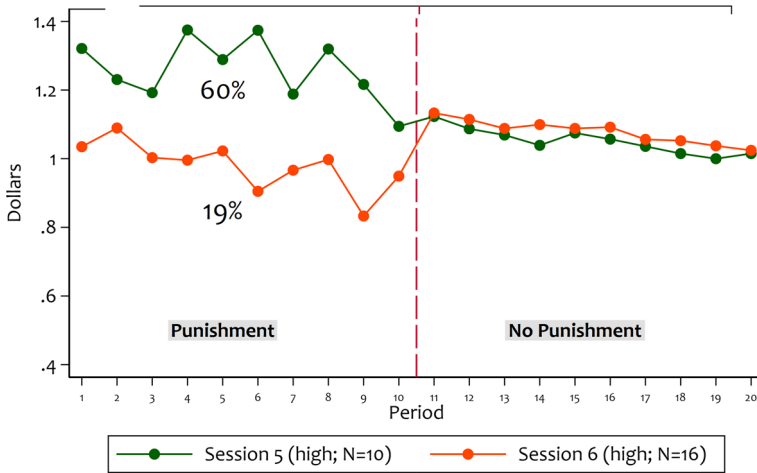


Fig. 5 Average profits with perfect strangers and P-NP history

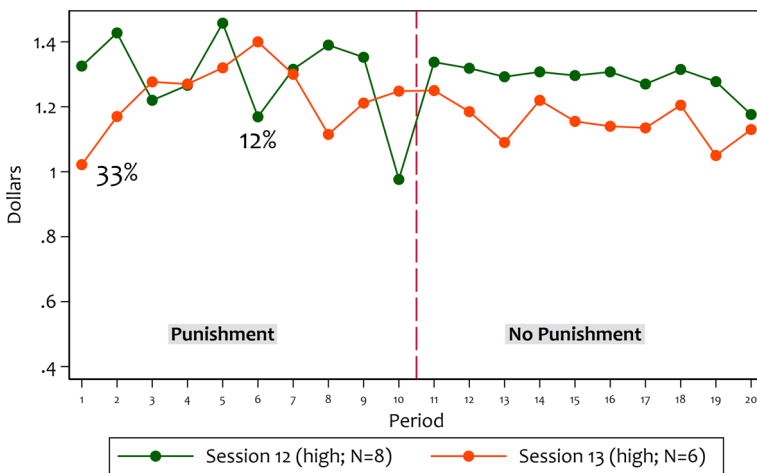
design. With  $N=6$ , the chance of meeting the same people in later rounds is higher than with  $N=16$ .<sup>19</sup>

Session 6 provided results that matched those in sessions 2 and 4, consistent with subjects being aware that a larger cohort size implied a smaller chance of a rematching with

<sup>19</sup> Sessions 5 and 6 were conducted in the same physical session, so 26 subjects were in the room. Computer stations previously were logged onto two different servers running two different sessions. Upon entering the lab subjects chose their seats. After seating, subjects were handed cards showing the number of subjects in the cohort. The same procedure was adopted for sessions 7 and 8.



**Fig. 6** Average profits with random strangers and P-NP history, Sessions 5 and 6



**Fig. 7** Average profits with random strangers and P-NP history, Sessions 12 and 13

the same person. Contributions started out around 7 tokens per period and decayed very slowly. They were around 4.5 tokens by period 10 and declined slowly through period 20. Punishment in periods 1–10 was costly, even after the customary period 1 bloodbath: average profits were lower by over 20 cents in each period because of the punishment. As Fig. 6 shows, average profits were systematically higher and less variable in periods 11–20 of session 6, so it was no surprise that only 8% voted for the institutional punishment mechanism.

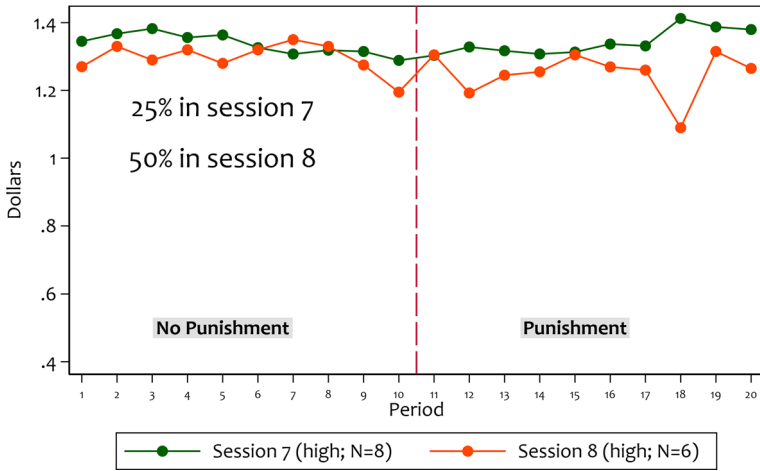


Fig. 8 Average profits with random strangers and NP-P history, Sessions 7 and 8

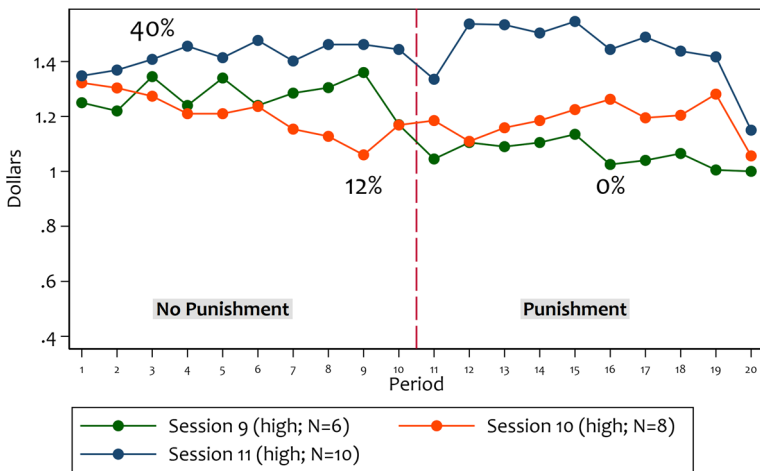


Fig. 9 Average profits with random strangers and NP-P history, Sessions 9, 10 and 11

Session 5 was a “poster boy” for the interpretation suggested by the observations in FG alone. Contributions started high, around 10.5 tokens, and generally remained at that level with sporadic, efficient punishment. But that was a setting in which the mere threat of sanctions seemed to have the desired effect: nobody needed to “pull the trigger” since contributions generally were high and profits robust; no “vandals” engaged in spiteful punishment that would undermine cooperative equilibria. With only 10 subjects in the cohort, it is likely that the subjects perceived the higher rematching probability as yielding reputation effects. Although the usual period 1 bloodbath occurred and might have weighed against

the vote for the institutional punishment mechanism, 60% of the subjects presumably viewed that as an outlier from the promise of things to come if another, final period with the mechanism was available. They were right: in period 21 of session 5 average contributions jumped from close to 0 in period 20 to more than 5. That is in itself an interesting finding since the one-shot nature of the final round easily could have caused the cooperative equilibrium under the threat of punishment to unravel, but it did not.

Sessions 12 and 13 exhibited roughly the same average profit in the two institutions, but the variation in profitability in the institutional punishment mechanism was striking. Those sessions led to very little support for the institutional punishment mechanism at the voting stage, consistent with subjects being risk averse and wanting to avoid any risk not associated with a clear “return” in the form of substantially higher average profits.

Figures 8 and 9 report results from random strangers sessions 7 through 11, all sharing an NP-P history and a high return to the public good. The institutional punishment mechanism fails to increase average profits over time compared to the prior no-punishment institution. However, in session 9, the explanation is a precipitous dip in profits in round 20, the last one of the institution with the punishment mechanism; ignoring that round, average profits were higher in this session. We generally see little support for the institutional punishment mechanism in the voting stage.

In summary, we find only one session in which a majority supports the institutional punishment mechanism. That is a random strangers session drawn from a small cohort when subjects experienced the institution with the punishment mechanism first and when the return to contributions was high. In a similar session, but wherein subjects experienced the non-punishment institution first, we find almost majority support, but in all other sessions the majority of the subjects prefer to live in a world without a punishment mechanism.

### 3.5 Statistical analysis

We complement the raw observations with a statistical analysis of individual subject votes. The dependent variable is the vote for the institution without the punishment mechanism. Explanatory variables include individual demographics and treatment effects. Binary dummy variables are entered for the perfect strangers designs, the size of the cohort conditional on the use of a random strangers design,<sup>20</sup> the history during periods 1–10, whether the subject received a low rate of return for contributions to the public good, whether the subject received a higher take-home profit in the NP institution (Profit\_NP), and whether the *other* player contributed more *on average* in the NP institution (Cratio\_NP). Demographic controls are subject’s age in years, binary indicators for sex, race, academic major, class standing, cumulative GPA below 3.25, cumulative GP above 3.75, number of people in the subject’s household, and a binary indicator of part-time or full-time employment. Table 3 reports descriptive statistics for the variables and Table 4 shows the complete set of coefficient estimates.

<sup>20</sup> The variable takes on the value 0 for the perfect strangers treatment and the size of the cohort (the “N in session” column from Table 1) for the random strangers treatments. It thus can be viewed as an interaction between the perfect strangers treatment and cohort size. Cohort size here is not the number of players in each particular public good game, which always is 2, but the number of people from which the pairings were selected.



**Table 3** Descriptive statistics for variables in voting model: Mean (standard deviation)

Variable	Mean (standard deviation)		Description
	Full sample	High returns sample	
VoteNP	0.778	0.731	Dummy variable, 1 if vote for the no-punishment (NP) institution, 0 otherwise
Profit_NP	0.628	0.546	Dummy variable, 1 if subject received higher take home profit in the NP institution, 0 otherwise
Cratio_NP	0.311	0.308	Dummy variable, 1 if other player contributed more in the NP institution, 0 otherwise
Pstrangers	0.567	0.4	Dummy variable, 1 if Perfect Strangers designs, 0 otherwise
Csize	4.2 (5.340)	5.815 (5.487)	Interaction between Pstrangers and cohort size in Random Strangers designs
np_p	0.5	0.492	Dummy variable, 1 if NP institution in the first 1–10 periods, 0 otherwise
high	0.722		Dummy variable, 1 if high rate of return to contributions, 0 otherwise
Age	21.517 (2.648)	21.285 (2.553)	Age, in years
Male	0.639	0.623	Dummy variable, 1 if male, 0 otherwise
Black	0.083	0.085	Dummy variable, 1 if black, 0 otherwise
Asian	0.083	0.085	Dummy variable, 1 if asian, 0 otherwise
Hispanic	0.128	0.115	Dummy variable, 1 if hispanic, 0 otherwise
OtherRace	0.044	0.054	Dummy variable, 1 if race other than white, black, asian, hispanic; 0 otherwise
Business	0.433	0.462	Dummy variable, 1 if academic major is business, 0 otherwise
PreSenior	0.472	0.5	Dummy variable, 1 if pre senior, 0 otherwise
GPAlow	0.467	0.431	Dummy variable, 1 if cumulative GPA below 3.25, 0 otherwise
GPAhigh	0.15	0.146	Dummy variable, 1 if cumulative GPA above 3.75, 0 otherwise
HHsize	1.650 (1.257)	1.638 (1.276)	Number of people in household
Work	0.722	0.738	Dummy variable, 1 if work part-time or full-time, 0 otherwise
N	180	130	Sample Size

**Table 4** Marginal effects of instrumental variables probit model of vote

Variable	Estimate	Standard error	<i>p</i> value	95% Confidence Intervals	
<i>A. Full Sample</i> (N = 180; 78% vote for NP; Wald $\chi^2_{14} = 77.49$ ; <i>p</i> value < 0.001)					
Profit_NP	0.639	0.144	< 0.001	0.357	0.921
Cratio_NP	-0.033	0.151	0.825	-0.329	0.262
Age	-0.017	0.013	0.177	-0.042	0.008
Male	-0.028	0.067	0.670	-0.160	0.103
Black	-0.024	0.126	0.852	-0.270	0.223
Asian	0.027	0.117	0.815	-0.202	0.257
Hispanic	-0.024	0.108	0.822	-0.235	0.187
OtherRace	-0.319	0.203	0.117	-0.717	0.079
Business	-0.036	0.075	0.631	-0.184	0.112
PreSenior	0.015	0.071	0.828	-0.123	0.154
GPA <sub>low</sub>	-0.038	0.071	0.592	-0.178	0.101
GPA <sub>high</sub>	-0.041	0.104	0.697	-0.245	0.164
HHsize	0.062	0.032	0.054	-0.001	0.124
Work	-0.016	0.069	0.812	-0.152	0.119
<i>B. High Returns Sample</i> (N = 130; 73% vote for NP; Wald $\chi^2_{14} = 41.46$ ; <i>p</i> value < 0.001)					
Profit_NP	0.604	0.267	0.024	0.081	1.128
Cratio_NP	-0.089	0.289	0.759	-0.655	0.477
Age	-0.016	0.019	0.408	-0.054	0.022
Male	0.024	0.091	0.794	-0.154	0.202
Black	0.077	0.139	0.578	-0.195	0.350
Asian	-0.040	0.166	0.809	-0.366	0.285
Hispanic	-0.154	0.175	0.380	-0.496	0.189
OtherRace	-0.378	0.227	0.096	-0.823	0.067
Business	-0.115	0.096	0.230	-0.302	0.072
PreSenior	0.045	0.094	0.629	-0.139	0.230
GPA <sub>low</sub>	-0.021	0.097	0.828	-0.211	0.169
GPA <sub>high</sub>	-0.024	0.143	0.866	-0.304	0.256
HHsize	0.082	0.040	0.041	0.003	0.161
Work	-0.021	0.100	0.836	-0.216	0.175

We are concerned a priori that two of the explanatory variables of particular interest might be endogenous to the vote. Those variables are the measures of relative profitability of the NP and P institutions to the subject casting the vote (Profit\_NP), and the measure of the relative cooperativeness of the other player in the NP and P environments (Cratio\_NP). One might argue that the values are predetermined by the time the vote is taken and cannot be endogenous. But our model of voting implicitly is a model of a *latent* propensity to vote for one institution over the other; that latent propensity might well be correlated with either of the two variables since it could reflect unobserved characteristics of the individual (e.g., “I like to free ride and punish people, no matter what” or “I like to contribute and avoid punishment”).

We checked for endogeneity using tests based on a maximum likelihood instrumental variables procedure documented in StataCorp (2017, p. 1185ff). The most natural

identifying assumption for the endogeneity test is that the experimental treatments determine the profits and contributions of others, but only affect voting through their impact on profits. When we allow both variables of interest to be endogenous, we *cannot* reject the null hypothesis of exogeneity using a Wald test ( $p$  value of 0.17). But evidence of endogeneity is found in the relative own-profit measure when tested independently of contributions by others<sup>21</sup>; we report estimates under the assumption that it is endogenous. Since we cannot reject exogeneity for contributions by others when tested independently, we enter instruments only for the own-profit measure.

Using the entire sample, we find strong evidence that it is the individual's relative profits in the NP versus P treatments that determine the vote. Subjects who experienced higher profit in NP than in P were 64 percentage points more likely to vote for NP; the effect is statistically significant (the 95% confidence interval is between 36 percentage points and 92 percentage points). No statistically significant effect on voting is found for the contribution levels of the other players. Using the sample with high returns only one comes to the same qualitative conclusion.

Although the treatment variables are used as instruments in the main statistical model, we can see the treatment effects in a reduced form model shown in Table 5. The effects of the absence of re-encounters in the perfect strangers design are in the expected direction. They are associated with an increase in the probability of voting for NP of 20 percentage points and are weakly significant on a one-tailed test ( $p$  value = 0.077). Related to that effect, the size of the cohort of potential opponents in the random strangers environment also is in the expected direction: every extra cohort member is associated in that environment with a 1.6 percentage point increase in the probability of voting for NP; the effect again is weakly statistically significant on a one-tailed test. The history experienced by the subjects affected their propensities to vote for the NP institution significantly: moving from the P-NP sequence to the NP-P sequence increases the probability of voting for NP by 22 percentage points.

As expected from our prior discussion of results, low rewards encourage outcomes in which participants are significantly more likely to vote for the NP institution, since the return to encouraging cooperation by the *efficient use of punishment* is lower. The effect of the low-returns treatment is to increase the probability of voting for the NP institution by 15 percentage points on average ( $p$  value = 0.01).

That analysis supports our hypothesis that preferences over institutions depend primarily on the earnings that subjects have experienced in them and not on the extent to which cooperative play is supported. We also find that the institutional preference is sensitive to the circumstances of the experience, as modeled by the experimental treatments.

## 4 Related literature

An emerging literature investigates endogenous institutions, such as in constitutional votes or “voting by feet” (Tiebout 1956). The extent to which participants make choices that involve punishment opportunities varies, supporting our conclusions that the circumstances favoring punishment are special, relating to particular experiences and the extent to which

---

<sup>21</sup> In an independent test of exogeneity of own profits, we *can* reject the null hypothesis of exogeneity ( $p$  value of 0.031). Similarly, in an independent test of other-player contributions, we cannot reject the null hypothesis of exogeneity ( $p$  value of 0.32).

repeated game characteristics are present. The findings in that literature also are supportive of our hypothesis that earnings are an important determinant of constitutional choices.<sup>22</sup>

Erhart and Keser (1999) examined the effects of allowing “Tiebout mobility” in a basic public goods contribution game. Their idea was to allow subjects to “vote with their feet” and decide which group they would like to join, so that individuals who have a taste for the public good could associate with like-minded others. Their experiments implemented that option in a simple manner, with 9 subjects in each session being able to choose the group in which they wanted to participate at the outset of each of 29 rounds after the first. Migration was costly: 50% of the endowment had to be sacrificed each period. The results were disappointing in the sense that endogenous migration did not generate the homogeneous groups one might expect. Basically, free-riding individuals behaved as if seeking out cooperating individuals. That would not be so bad for public good provision if they changed their self-interested ways, but after joining the group they exploited it and the process cycled.<sup>23</sup> Overall, average contributions to the public good decayed steadily.

Page et al. (2005) extended the same idea in several ways. In each session, 16 subjects participated in a voluntary public goods contribution game in groups of 4 for 20 rounds. After round 3, they were allowed to rank the other 15 individuals, who had anonymous labels. The information available after each round was the *average* contribution of the other individual over the experiment up to the previous round. Ranking activities are costly, but the cost is minimal. An algorithm assigned subjects to groups of four based on similarity in the rankings of each other. Four environments were examined. One was a *baseline* with no punishment option or ranking. The second was a *punishment* environment, akin to the one studied by FG. The third was a *regrouping* environment in which subjects were placed into groups in rounds 4–20 based on the rankings submitted. The fourth was a *combination* of the punishment and regrouping treatments. Page and coauthors found no significant pairwise differences in contributions or earnings between the last three environments. However, they did find a statistically significant increase in contributions and earnings when the baseline and regrouping environments were compared and, moreover, when the baseline and combined environments were compared. Relative to the Erhart and Keser (1999), Page et al.’s design appears less vulnerable to free riders exploiting cooperative subgroups.

Gürek et al. (2006) adopt a “voting by feet” design to examine the effects of allowing subjects to self-select into groups operating under different institutions in a public goods game repeated over 30 rounds. Each subject in a group of 12 chooses at the beginning of each round between being in a sanction-free institution (SFI) or a sanctioning institution (SI), knowing that they will then interact with subjects who also choose the same institution in that round. The design of the contribution stage follows FG (2000, 2002). After the contribution stage, subjects receive an additional 20 experimental tokens. The extra tokens simply were retained by those in the SFI, but they could be used to punish or reward other

<sup>22</sup> Appendix B of Botelho et al. (2019) provides additional details on the studies referenced here, as well as reviews of the designs and primary findings of Gürek et al. (2005, 2009, 2014), Kosfeld et al. (2009), Putterman et al. (2011), DeAngelo and Charness (2012), Gürek (2013), Markussen et al. (2014) and Drouvelis and Jamison (2015).

<sup>23</sup> In a different experimental setup than the one used in much of the recent punishment literature, Powell and Wilson (2008) investigate whether individuals deviate from “cooperative” behavior after agreeing unanimously to a cooperative non-binding social contract. They implement a “Hobbesian” framework wherein non-cooperation involves taking the property of another person, finding that in only one instance (out of 31 possibilities) after individuals approved the social contract did deviation soon set in.

**Table 5** Marginal effects estimated with reduced form probit model

Variable	Estimate	Standard error	<i>p</i> value	95% Confidence Intervals	
<i>A. Full Sample</i> (N = 180; Wald $\chi^2_{16} = 39.66$ ; <i>p</i> -value < 0.001)					
Pstrangers	0.199	0.140	0.153	-0.074	0.473
Csize	0.016	0.011	0.149	-0.006	0.039
np_p	0.225	0.062	<0.001	0.102	0.347
Low	0.153	0.059	0.010	0.037	0.270
Age	-0.008	0.011	0.456	-0.031	0.014
Male	-0.036	0.063	0.565	-0.160	0.087
Black	0.081	0.084	0.333	-0.083	0.246
Asian	0.073	0.083	0.379	-0.090	0.235
Hispanic	-0.122	0.109	0.265	-0.336	0.092
OtherRace	-0.139	0.171	0.415	-0.473	0.195
Business	-0.127	0.066	0.053	-0.256	0.002
PreSenior	0.027	0.062	0.670	-0.096	0.149
GPAlow	-0.051	0.069	0.462	-0.185	0.084
GPAhigh	-0.022	0.091	0.812	-0.199	0.156
HHsize	0.066	0.026	0.011	0.015	0.116
Work	-0.041	0.060	0.499	-0.159	0.077
<i>B. High Returns Sample</i> (N = 130; Wald $\chi^2_{15} = 21.49$ ; <i>p</i> value = 0.1220)					
Pstrangers	0.180	0.137	0.188	-0.088	0.449
Csize	0.017	0.014	0.208	-0.010	0.045
np_p	0.217	0.081	0.007	0.058	0.375
Age	-0.004	0.018	0.811	-0.039	0.030
Male	-0.037	0.086	0.670	-0.205	0.132
Black	0.150	0.103	0.147	-0.052	0.352
Asian	0.061	0.126	0.631	-0.187	0.309
Hispanic	-0.246	0.156	0.115	-0.553	0.060
OtherRace	-0.229	0.212	0.281	-0.644	0.187
Business	-0.183	0.084	0.029	-0.348	-0.019
PreSenior	0.042	0.092	0.651	-0.139	0.222
GPAlow	-0.053	0.094	0.573	-0.238	0.132
GPAhigh	-0.084	0.130	0.518	-0.339	0.171
HHsize	0.066	0.033	0.046	0.001	0.130
Work	-0.070	0.080	0.383	-0.227	0.087

in-group members in the SI. At the end of each round, subjects received information concerning contributions, tokens given and received as punishments or rewards (if in the SI), along with profits for every subject in both institutions on an anonymous basis.

The overall results from 7 sessions implementing that design are striking. Initially, less than 40% of the subjects join the SI, but the percentage increases steadily and, after 18 rounds, more than 90% of the subjects join. Contribution levels were substantially higher in the SI than in the SFI throughout the experiment. High contributors in the SI earned substantially more than free riders in the SFI after the fifth period, suggesting that subjects self-select into the institution that yields higher profits and mimic the behavior prevalent

under that institution. Those results support our findings that choices over institutions with or without punishment mechanisms will depend on earnings. They likewise would suggest that subjects should vote in favor of institutional punishment mechanisms, which is not confirmed in our data. Our data instead suggest that the conditions under which such institutions are preferred are very special, depending on the exact experiences of the participants.

Ertan et al. (2009), Noussair and Tan (2011) and Sutter et al. (2010) employ designs that are similar to ours. Rather than allowing individuals to migrate between institutions, they implement a constitutional choice in which individuals vote on whether to adopt one or another alternative institution.

Ertan et al. (2009) investigate experimentally how the adoption of sanctioning rules evolves over a series of votes. In one treatment, named “3-Vote,” subjects played a three-round contribution game without punishment opportunities, followed by another three rounds with unrestricted punishment in the spirit of FG. They then voted for the rule that would govern their in-group interactions over the next eight rounds; the vote was repeated two more times at the end of each sequence of eight rounds. The other treatment, named “5-Vote,” was similar except that subjects started by voting on the rules without any prior experience. Subjects could vote for reducing other in-group members’ earnings when their contributions were lower than, equal to, or higher than the average group contribution. A majority voting rule was applied to each of these three ballot items.

Across both treatments and all voting stages, only 30% of the individual votes favored some punishment rule, with 72% of them allowing for punishment of lower-than-average contributors. The vast majority of the individual votes (67%) opposed at least one of the possible punishment rules. Overall, 61% of the groups allowed punishment only of lower-than-average contributors; 35% of the groups did not allow any punishment whatsoever.

The results are supportive of our finding that the particular experiences of participants affect their votes. The number of groups prohibiting punishment in favor of groups allowing punishment of low contributors declines over time, but it seems to be more pronounced in the 3-Vote treatment than in the 5-Vote treatment. That result may be explained by the initial institutional experience that participants have in the former. Groups that vote for punishments of low contributors generally realize significantly higher average contributions; however, despite the earnings advantage, it is a small advantage in comparison to the cost.

In a related study, Noussair and Tan (2011) adapt the design developed by Ertan et al. (2009) to allow for the heterogenous composition of groups with respect to the productivity of the members’ contributions: half of the members in each group were assigned a high return on contributions to the public good (named type A players), and the other half a low return (named type B players). All subjects played a three-round contribution game without punishment opportunities, followed by another three rounds with unrestricted punishment. They then voted for the rule that would govern their in-group interactions over the next rounds. Voting occurred every two rounds in a designated short-term treatment and every eight rounds in a designated long-term treatment. Subjects could vote for reducing other in-group members’ earnings depending on their types and their contribution levels to the group account (contribution level lower or higher than the average group contribution). A majority voting rule was applied to each of the four ballot items. Considering both treatments and all voting stages, no group ever voted in favor of unrestricted punishment, 31% of the groups explicitly voted against any punishment whatsoever, and 58% of the groups allowed punishment only of lower-than-average contributors (from both types of players simultaneously or just from one of the types). The average earnings of the subjects

in the latter groups were substantially higher than the average earnings of subjects in the other groups, but the heterogeneity of the players makes it more difficult for the groups to achieve consensus on which particular punishment system to implement relative to the groups in Ertan et al. (2009).

Sutter et al. (2010) investigate whether subjects prefer to interact in institutions that allow punishments, that allow rewards, or neither. They varied the intensity of the reward and punishment options exogenously. In a “low-leverage” treatment, it cost a subject 1 token to increase (reduce) the earnings of another group member by 1 token; in a “high-leverage” treatment, it cost a subject 1 token to increase (reduce) the earnings of another group member by 3 tokens. The vote took place before participants gained any experience in either institution. Subjects incurred a one-time fee to participate in the vote and could abstain from the costly vote knowing that the decision of the voters would still be binding on them. Roughly 44% and 60% of the subjects in the low-leverage and high-leverage treatments participated in the costly vote, respectively. No group ever opted for the punishment institution in the high-leverage treatment and only 12.5% opted for it in the low-leverage treatment. The vast majority of groups opted for the institution with rewards in the high-leverage condition (85% of these groups) and for neither rewards nor punishments in the low-leverage condition (62.5% of these groups). Those results again lend support to our conclusion that the circumstances under which an institution with costly punishments is chosen are very special.

Gintis et al. (2005, ch.1) and Boyd et al. (2005) have argued that the desirability of sanctions, and norms that encourage their use, is the product of evolution. Our results, and those of others that show detrimental effects from sanctions, suggest that where wealth outcomes are the metric of fitness, this evolutionary argument cannot hold without serious qualification. Dawkins (1986) famously introduced the metaphor of a blind watchmaker to make the point that complex objects could be produced by an evolutionary process that had no intention of producing the object. Dawkins (1986) referred to genetic evolution, whereas the hypothesis in the economics literature we cited is about cultural evolution. However, both kinds of evolutionary process rely on some adaptiveness filter to be applied to weed out the many mistakes that random deviation tosses out as part of such an undirected process. In fact, lots of mistakes are crucial to the evolutionary argument. If we insist on interpreting the presence of sanctions as evidence of an evolutionary process in cases for which the outcomes are not adaptiveness peaks, then we must be observing the blind watchmaker on one of his bad days; otherwise, some other latent metric of adaptiveness we cannot (currently) observe may be at work. If sanctions lead to net earnings losses compared to processes without sanctions, then sanctions cannot be part of an evolutionarily stable outcome in which relative earnings are the adaptiveness filter.

FG (2000, 2002), Gintis et al. (2005, ch.1) and Boyd et al. (2005) follow other literature in behavioral economics in understanding norms as arguments in individual utility functions that rank social distributions of utility. Thus, they wonder whether the behaviors of their subjects might be motivated by “norms of fairness” or “norms of reciprocity” that might be to some extent shared. Such an understanding of norms is “individualistic” in the sense that it views them as emerging strictly from social preferences.

Such an understanding of norms is not compatible with the concept of norms adopted in other social sciences, which model norms as social structures that are independent of any individual’s preferences. Binmore (2010) argues that interpreting norms as emerging from social preferences, with the attendant purpose of explaining strategy choices in games by reference to them, misconstrues utility functions as descriptions of motivational structures instead of as summaries of choice patterns and, as such, incompatible with revealed

preference theory. Binmore (2010) defends an alternative conception of norms according to which they are evolved social conditions, roughly, unformalized institutions, that serve as equilibrium selection mechanisms by assigning asymmetric bargaining weights to occupants of different social roles.

That argument leaves the relationship between norms and cognitive structures, such as beliefs, in a black box. Bicchieri (2006) offers a theory of the black box's contents. According to her, norms are best understood as shared expectations and fall into two types. A norm can merely be *descriptive* if all or most people in a group expect others to make choices consistent with the norm; and it is a *social* norm if in addition all or most people in a group believe that others will think that all or most are obliged, as a matter of social responsibility, to make choices in accord with the norm.

In the context of laboratory experiments such as those of FG (2000, 2002), and the experiment we designed, we understand the role of norms as compatible with the complementary conceptions of Bicchieri (2006) and Binmore (2010) as follows. Subjects may come into the lab with normative expectations, which might be descriptive or social, that they have learned to apply in what Binmore (1994, 1998, 2010) calls “the game of life.” In other words, they may or may not believe that the game they play in the lab is a domain in which the same norms are in force. Observed play of others in the experiment may or may not provoke revisions in those beliefs between rounds of play. And subjects might or might not operationalize preferences that reflect optimal play in the game of life, which can diverge from optimal play in the experimental game.

In light of such indeterminacy, we deliberately discussed our experimental design and results in terms of observed *preferences over institutions* revealed by explicit, incentivized choices with which subjects are presented. Although we consider that those preferences are partly conditional on subjects' norms and on their beliefs about the applicability of norms to the lab, we do not take our evidence as sufficient to license inferences about the unobserved conditions.

## 5 Conclusions

Our experiment addresses the question of whether the desirability of social institutions, such as those that allow punishments to enforce cooperative play, depends on the extent to which cooperative expectations are upheld or on the profits generated. Earlier studies in the same domain already have shown the pervasiveness of costly, informal sanctioning behavior. The crucial question we address is not, therefore, whether individuals will sanction if given the option, but whether they want to have the option available at all in the first place. While sanctions largely have been interpreted as the individual's desire to retaliate against those who do not comply with a cooperative norm of behavior (e.g., Falk et al. 2005), the strength of that interpretation for the success and stability of “self-governing” institutions rests on the assumption that choices made within *imposed* sets of constraints or values coincide with the *endogenous* choice of those constraints and values themselves.

Our results provide a case study in which observed sanctioning behavior within an imposed institutional framework does not translate into the acceptance by the same individuals of that institutional framework. Thus, a distinction is required between the principles that guide the *choice of institutions* and the principles that apply to actions *guided by institutions*. Although an analysis of the latter facilitates an understanding of how institutions work, it leaves completely open questions pertaining to their origin and evolution. In



the specific setting examined herein, the simple maximization of expected profit appears to explain the choices made by subjects when they are allowed to vote on the institution.

**Acknowledgements** Harrison and Rutström thank the U.S. National Science Foundation for research support under Grants NSF/IIS 9817518, NSF/HSD 0527675 and NSF/SES 0616746. Botelho and Pinto thank the Fundação para a Ciência e Tecnologia for sabbatical scholarships SFRH/BSAB/489/2005 and SFRH/BSAB/491/2005, respectively. Pinto acknowledges financial support from the Portuguese Foundation for Science and Technology within the project UID/ECO/03182/2019. We are grateful to Ryan Brosette, Linnéa Harrison, James Monogan and Bob Potter for research assistance, and to Andreas Ortmann, Federica Palente, Greg DeAngelo and a referee for helpful comments. All experimental instructions, data and computer code for statistical analyses are available at <https://cear.gsu.edu/gwh/>.

## References

- Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, 54(1), 1–24.
- Andreoni, J., & Croson, R. T. A. (2005). Partners versus strangers: Random rematching in public goods experiments. In C. R. Plott & V. L. Smith (Eds.), *Handbook of experimental economics results*. Amsterdam: North-Holland.
- Bicchieri, C. (2006). *The grammar of society*. Cambridge: Cambridge University Press.
- Binmore, K. (1994). *Game theory and the social contract volume 1: Just playing*. Cambridge: MIT Press.
- Binmore, K. (1998). *Game theory and the social contract volume 2: Playing for real*. Cambridge: MIT Press.
- Binmore, K. (2010). Social norms or social preferences? *Mind & Society*, 9(2), 139–157.
- Botelho, A., Harrison, G. W., Pinto, L. M. C., Ross, D., & E. Rutström, E. E. (2019). Endogenous choice of institutional punishment mechanisms to promote social cooperation. CEAR Working Paper 2019-04. Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.
- Botelho, A., Harrison, G. W., Pinto, L. M. C., & Rutström, E. E. (2009). Testing static game theory with dynamic experiments: A case study of public goods. *Games and Economic Behavior*, 67(1), 253–265.
- Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2005). The evolution of altruistic punishment. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.
- Carpenter, J. (2007). Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, 60, 31–51.
- Carpenter, J., & Matthews, P. (2004). Social reciprocity. Working Paper 0229r, Department of Economics, Middlebury College.
- Casari, M., & Luini, L. (2009). Group cooperation under alternative peer punishment technologies: An experiment. *Journal of Economic Behavior & Organization*, 71, 273–282.
- Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, 12, 187–218.
- Dawkins, R. (1986). *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. New York: Norton.
- DeAngelo, G., & Charness, G. (2012). Deterrence, expected cost, uncertainty and voting: Experimental evidence. *Journal of Risk and Uncertainty*, 44(1), 73–100.
- Drouvelis, M., & Jamison, J. C. (2015). Selecting public goods institutions: Who likes to punish and reward? *Southern Economic Journal*, 82(2), 501–534.
- Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B Biological Sciences*, 275, 871–878.
- Erhart, K. M., & Keser, C. (1999). Mobility and cooperation: On the run. Working Paper 99 s-24, CIRANO, University of Montreal.
- Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5), 495–511.
- Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 73(6), 2017–2030.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137–140.
- Fischbacher, U. (2007). z-Tree—Zurich toolbox for readymade economic experiments. *Experimental Economics*, 10(2), 171–178.
- Fréchette, G. R. (2012). Session-effects in the laboratory. *Experimental Economics*, 15, 485–498.

- Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (Eds.). (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.
- Goeree, J. K., Holt, C. A., & Laury, S. K. (2002). Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics*, 83, 255–276.
- Güererk, Ö. (2013). Social learning increases the acceptance and the efficiency of punishment institutions in social dilemmas. *Journal of Economic Psychology*, 34, 229–239.
- Güererk, Ö., Irlenbusch, B., & Rockenbach, B. (2005). On the evolution of institutions in social dilemmas. Working Paper, University of Erfurt.
- Güererk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312, 108–111.
- Güererk, Ö., Irlenbusch, B., & Rockenbach, B. (2009). Motivating teammates: The leader's choice between positive and negative incentives. *Journal of Economic Psychology*, 30(4), 591–607.
- Güererk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics*, 120, 220–230.
- Harrison, G. W., & Hirshleifer, J. (1989). An experimental evaluation of weakest-link/best-shot models of public goods. *Journal of Political Economy*, 97, 201–225.
- Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(7), 1362–1367.
- Isaac, R. M., & Walker, J. M. (1988). Communication and free-riding behavior: The voluntary contribution mechanism. *Economic Inquiry*, 26(4), 585–608.
- Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, 99(4), 1335–1355.
- Markussen, T., Putterman, L., & Tyran, J. R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *Review of Economic Studies*, 81(1), 301–324.
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves. *Journal of Public Economics*, 92, 91–112.
- Nikiforakis, N., & Normann, H. T. (2008). A comparative statics analysis of punishment in public good experiments. *Experimental Economics*, 11(4), 358–369.
- Noussair, C. N., & Tan, F. (2011). Voting on punishment systems within a heterogeneous group. *Journal of Public Economic Theory*, 13(5), 661–693.
- Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Journal of Political Science*, 86(2), 404–417.
- Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: Reciprocity, mimicry, and efficiency. *Economic Journal*, 115, 1037–1058.
- Palfrey, T. R., & Prisbrey, J. E. (1996). Altruism, reputation, and noise in linear public goods experiments. *Journal of Public Economics*, 61, 409–427.
- Palfrey, T. R., & Prisbrey, J. E. (1997). Anomalous behavior in linear public goods experiments: How much and why. *American Economic Review*, 87, 829–846.
- Powell, B., & Wilson, B. J. (2008). An experimental investigation of Hobbesian Jungles. *Journal of Economic Behavior & Organization*, 66(3–4), 669–686.
- Putterman, L., Tyran, J. R., & Kamei, K. (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics*, 95(9), 1213–1222.
- Sefton, M., Shupp, R. S., & Walker, J. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, 45(4), 671–690.
- Simonsohn, U. (2006). Review of “Moral sentiments and material interests”. *Journal of Economic Literature*, XLIV, 745–747.
- StataCorp. (2017). *Stata base reference manual: Release 15*. College Station, TX: Stata Corp. LLC.
- Sutter, M., Haigner, S., & Kocher, M. (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies*, 77(4), 1540–1566.
- Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64(5), 416–424.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Anabela Botelho<sup>1</sup>  · Glenn W. Harrison<sup>2,5</sup>  · Lígia M. Costa Pinto<sup>3</sup>  ·  
Don Ross<sup>4,5,6</sup>  · Elisabet E. Rutström<sup>6,7</sup> 

Anabela Botelho  
anabela.botelho@ua.pt

Lígia M. Costa Pinto  
pintol@eeg.uminho.pt

Don Ross  
don.ross931@gmail.com

Elisabet E. Rutström  
erutstrom@gmail.com

- <sup>1</sup> Department of Economics, Management, Industrial Engineering and Tourism, Research Unit on Governance, Competitiveness, and Public Policy (GOVCOPP), University of Aveiro, Aveiro, Portugal
- <sup>2</sup> Department of Risk Management and Insurance and Center for the Economic Analysis of Risk, Georgia State University, Atlanta, USA
- <sup>3</sup> NIPE, Centre for Research in Economics and Management, University of Minho, Braga, Portugal
- <sup>4</sup> School of Society, Politics and Ethics, University College Cork, Cork, Ireland
- <sup>5</sup> School of Economics, University of Cape Town, Cape Town, South Africa
- <sup>6</sup> School of Business, Örebro University, Örebro, Sweden
- <sup>7</sup> Center for the Economic Analysis of Risk, Georgia State University, Atlanta, USA