CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Endogenous choice of institutional punishment mechanisms to promote social cooperation
Authors	Botelho, Anabela;Harrison, Glenn W.;Pinto, Lígia M. Costa;Ross, Don;Rutstrom, Elisabet E.
Publication date	2021-01-03
Original Citation	Botelho, A., Harrison, G. W., Pinto, L/ M. C., Ross, D. and Rutstrom, E. E. (2021) 'Endogenous choice of institutional punishment mechanisms to promote social cooperation', Public Choice, 191, pp. 309-335. doi: 10.1007/s11127-020-00868-5
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://link.springer.com/article/10.1007/s11127-020-00868-5 - 10.1007/s11127-020-00868-5
Rights	© The Author(s), under exclusive licence to Springer Science +Business Media, LLC part of Springer Nature 2021. This is a post-peer-review, pre-copyedit version of an article published in Public Choice. The final authenticated version is available online at: https://doi.org/10.1007/s11127-020-00868-5
Download date	2026-11-05 22:41:57
Item downloaded from	https://hdl.handle.net/10468/10950



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Endogenous Choice of Institutional Punishment Mechanisms to Promote Social Cooperation

by

Anabela Botelho, Glenn W. Harrison, Lígia M. Costa Pinto, Don Ross & Elisabet E. Rutström †

September 2019

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.

† Department of Economics, Management and Industrial Engineering, University of Aveiro, Portugal (Botelho); Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Georgia State University, USA (Harrison); Department of Economics, University of Minho, Portugal (Pinto); School of Sociology, Philosophy, Criminology, Government and Politics, University College Cork, Ireland; School of Economics, University of Cape Town, Souuth Africa; Center for the Economic Analysis of Risk, Georgia State University, USA (Ross); Örebro University and Stockholm School of Economics, Sweden, and Center for the Economic Analysis of Risk, Georgia State University, USA (Rutström). Harrison and Rutström are also affiliated with the School of Commerce, University of Cape Town, South Africa. E-mail: anabela.botelho@ua.pt, gharrison@gsu.edu, pintol@eeg.uminho.pt, don.ross931@gmail.com, and erutstrom@gmail.com.
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. We are grateful to Ryan Brosette, Linnéa Harrison, James Monogan and Bob Potter for research assistance, and to Andreas Ortmann and Federica Pallente for helpful comments.

There is a growing consensus in the experimental literature that institutions that allow the use of voluntary punishments can reduce the free rider problem in public goods games. This behavioral phenomenon has been studied using variations of an experimental design introduced by Ostrom, Walker and Gardner [1992] and Fehr and Gächter [2000][2002].[1] The *inference* often drawn from these findings is that enforcement mechanisms tend to be welfare-enhancing to the extent that they align incentives. We offer a different perspective on the value of institutions by constructing laboratory experiments where we *directly elicit institutional preferences* from individuals that are experienced in institutions with voluntary punishment and institutions without voluntary punishment. Our analysis of these expressed preferences explicitly recognizes that imposing punishments is a costly endeavor, such that the value of an institution may not be positively correlated with its contribution to reducing free riding. We find that institutional preferences are not mysterious in this case: subjects are generally motivated by self-interest when choosing institutions.

Gürerk, Irlenbusch, and Rockenbach [2006] demonstrate that societies with rules that allow the voluntary use of costly sanctions, and that generate aggregate wealth *advantages* over societies without such rules, can grow over time in membership at the expense of the "more anarchic" societies. On the other hand, in experiments where subjects cannot switch between alternative institutional structures, we sometimes see voluntary sanctions leading to aggregate wealth disadvantage compared to those that do not allow sanctions. We review the data from Fehr and Gächter [2000][2002] and find that this is the case in their experiments. Many other studies report detrimental effects on earnings from unrestricted sanctions: see Carpenter and Matthews [2004], Page, Putterman and Unel [2005], Sefton, Shupp, and Walker [2005], Casari and Luini [2005], Anderson and Putterman [2006], Nikiforakis [2008], Nikiforakis and Normann [2008] and Egas and Riedl [2008]. We implement a new experimental design that allows us to identify the independent influence on institutional preferences from the incidence of free riding and the profitability of the institution.

---

[1] The efficacy of punishments is questioned in some studies. Gintis, Bowles, Boyd and Fehr [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 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."

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 reputation effects. This 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. There is a strong negative correlation between the vote for the sanctions institution and the loss in profits that it caused. Other motivations, such as fairness, may also have played a role, but effects on wealth are a key candidate for what motivated actions.[2]

1. The Value of a Punishment Rule

Relying on an established literature that argues 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.[3] More precisely, what is the net value of allowing the punishment technology that endogenously generates the cooperative behavior?

In the experiment of Fehr and Gächter (FG) [2000] subjects play a Voluntary Contribution (VC) game over 20 periods, where in one set of 10 periods they do not have the option to punish but in another set they do. They vary the ordering of these two within-subject treatments. Two between-subjects treatments are implemented based on how subjects are matched into groups of four. In one they employ a Partners design where the same subjects are matched throughout the full 20 periods, and in another they use a Random Strangers design where subjects are rematched into new groups before

[2] 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 greater with sanctions, and for a majority, but one could easily imagine that some in the majority might not want to live in such an inequitable world, and would vote against the world with sanctions.

[3] We take issue with the sense in which the previous literature in question actually identifies a social norm, and discuss that issue at the end of §3.

the start of each round. In the VC game all subjects are given an initial endowment of tokens, and they can choose to keep these 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 are also costly to the sender. The unique Nash Equilibrium prediction is for nobody to invest in the project or to send punishment points.

The top panels of Figure 1 display the relative gains from allowing sanctions, based on the data generated by FG [2000].[4] In the First Series subjects experienced ten periods with sanctioning after an initial ten periods without and in the Second Series this ordering was reversed. The earnings shown in the top panels of Figure 1 correspond to a pattern of contributions in the public goods game that converge to the cooperative outcome in periods 9 or 10, as shown in FG [2000]. But the accumulated cost of the use 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.[5] The experiment might trigger activation of pre-existing normative expectations in subjects that lead them to implement punishment. However, this 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 Figure 2 reports comparable calculations for the experiments in FG [2002].

---

[4] We only consider the "Strangers" design in FG [2000], since it controls for the possible role of strategic self-interest in employing the sanctions. In their "Partners" design the same subjects played against each other for ten periods, and in their "Strangers" design individuals were randomly assigned to groups after each period. In FG [2002] they only consider Strangers designs. We are grateful to Simon Gächter for providing the data from their experiments.

[5] 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, for a difference of 3.88 currency units or 17.1%.

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

These experiments show that sanctions can have a strong effect 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 this the punishment mechanism available if they were to make 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 Figures 1 and 2 consider this question, using two possible voting rules[6] for social choice:

- Majority Rule Referenda – would the median voter opt for the social technology?
- Super-Majority Rule Referenda – would 67% of the population vote for the punishment mechanism as am 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 each condition, 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 calculate the period-wise median and 33rd percentile. These are shown in the bottom panels.

The implication from Figure 1 and 2 is that, with two exceptions, *the institutional punishment*

[6] The results in the top panels imply what would happen if a Classical Utilitarian social choice rule was used, 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 to just use the results of the last period or two to calculate benefits or costs in the experiments of FG [2000] (Figure 1).

mechanism would not be adopted under either of these social choice criteria.[7] 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 socially chosen under these criteria. Our experiment was designed to investigate the question of social choice directly, by allowing participants to vote over institutional mechanisms after they have experienced the effects of each. This allows us to simply observe the choices made by the subjects and then infer whether they seem to be extrapolating into "games of life" or not.

2. Voting for a Punishment Institution

A. 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 there is a punishment mechanisms and also in games in which there are no such mechanisms. We examine order effects as they do by running one set of subjects through experiments in which the punishment option comes first, and then the no-punishment option is experienced, and running a separate set of subjects through the same game but with the reverse order. We allow 10 periods in each setting, so that each subject plays 20 periods prior to the vote.

To ensure that there are no confounding reputation effects, and to provide the cleanest possible test, we include a Perfect Strangers design in which no subject ever meets the same subject more than once. Virtually all previous public goods experiments use a Random Strangers design in which subjects are randomly re-assigned every period.[8] Although this design reduces the chance that the subject will

[7] These exceptions are period 8 of the First Series in Figure 1, and period 5 of the Second Series of experiments in Figure 2, where the median voter would just vote *for* the norm. The norm would not survive a constitutional referendum using a super-majority rule in these periods.

[8] 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 only considered one sequence

meet the same person to low levels, and is coupled with anonymity, the critical behavioral issue is whether the subjects believe that there is no reputation effect of their choices in a given round. In a Perfect Strangers design it is clear that subjects should hold a belief that there is a zero likelihood of meeting any other player again. In our experiments subjects participate in groups of 2 in each round.[9] We explain carefully to them how we ensure that there is no chance that they will meet the same person in any other round. Since this 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 reputation conditions.[10] Subjects are randomly matched up into pairs from within the same cohort.

After period 20 we ask subjects to vote on the environment they would like to participate in for one "Final Jeopardy!" round.[11] 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.

---

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, Harrison, Pinto and Rutström [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.

[9] Most public goods experiments use four subjects per group, although the effect of larger group sizes has been studied by Isaac and Walker [1988a] and others. Harrison and Hirshleifer [1989] and Goeree, Holt and Laury [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.

[10] We vary the cohort size in these 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 size of their respective cohort. When more than one cohort was present the text of the instructions was changed to reflect the fact that the vote outcome was implemented separately for each cohort based on that cohort's vote.

[11] In the popular TV game show *Jeopardy!* there are three rounds of play: "Jeopardy!," "Double Jeopardy!," and "Final Jeopardy!" The first two consist of three categories of three questions each, but "Double Jeopardy!" has doubled dollar values. There is only one question in "Final Jeopardy!," and subjects can wager their accumulated earnings in that round.

> 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 vote outcome applies to a cohort rather than to everyone in the room. The matching protocol employed in the first part continues in this last period, so that in the Perfect Strangers design they 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 this period will be ten times those of each of the first 20 periods. This one-shot design of the final round is precisely the environment that the earlier rounds are attempting to model, although they 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 ordinary Random Strangers designs. The return to the public good is discussed below, as are the votes.

B. 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 up front that we would pay them 5 cents for every token they had at the end of the experiment.

In one treatment we used a relatively low return on contributions to the public good, and in another treatment we used a relatively high return. The low return was 0.6 of a token: every token contributed to the public good by one subject would decrease their private endowment by 1 token and return 0.6 of a token for themself. Of course, it would also 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 this treatment was to see the effects of making the environment more rewarding to influence, such as a normative expectation imported from a "game if life," that would increase contributions to the public good. Table 1 shows that the low return was used in sessions 1 and 2, and the high return in all other sessions. We used a linear payoff schedule which was constant for all contributions, so the dominant strategy is simple: a subject that only seeks to maximize individual earnings in a single period should contribute nothing to the public good.[12]

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, and in addition subjects received a one time endowment of 25 tokens to cover possible losses. As Table 2 shows, this allowed *each* subject in *one* period to buy up to 9 punishment points without incurring a loss in that period (and before factoring in any profit from production of the public good or the private good).

[12] 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, Holt and Laury [2002].

C. Procedures

We recruited 180 subjects from the University of Central Florida (UCF) in 2005.[13] Subjects were randomly assigned to each session, with no prior knowledge of the parameters or treatments. The sessions were all conducted at the Behavioral Research Lab of the College of Business Administration of UCF. This 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].[14] The same experimenter (Rutström) delivered the oral instructions for all sessions, to ensure comparability.[15] The oral instructions also utilized a large-screen display that could be easily seen by all subjects, to ensure that certain information was common knowledge. Training rounds were included prior to each regime, to ensure that subjects understood the task.

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

D. Observed Outcomes

Table 1 shows the vote in each session, which is our "bottom line" result: when there was a zero chance of ever meeting any other person again, 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 the mechanism. The vote was close in one of the four Perfect Strangers sessions, but there was little doubt in the other three. In fact, in one session, all 26 subjects agreed unanimously

[13] UCF is located in Orlando, Florida. It has a large student body, with Fall 2004 enrollment of 42,837. The entering class in 2004 had an average SAT of 1,186. The student body is also ethnically divers: in 2004 8.5% stated that they were Black and Non-Hispanic; 70% stated that they were White and Non-Hispanic; 5.0% stated that they were Asian; and 12.2% stated that they were Hispanic.

[14] All instructions, scripts, and software are available at http://exlab.bus.ucf.edu. The latest version of the *z-Tree* software and documentation is available at http://www.iew.unizh.ch/ztree/index.php.

[15] A digital recording of the oral instructions in one typical session is available at the ExLab archive.

to implement the institutional mechanism that did not entail punishment. This result was robust to the use of high or low returns to the public good, and the order in which subjects experienced the institutions with or without punishment prior to the vote.

Our data replicates the finding reported in FG [2000][2002] that punishments lead to higher contributions on average. With punishments the average token contributions are 7.42 and without punishment they are 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. The joint significance of the observations here and in FG is that the use and effects of punishments vary across different groups of subjects, and one cannot say that they uniformly have a sustained positive effect on contributions, much less on profits.[16]

Figure 3 provides detailed results for session 1 to illustrate the outcomes. This is the session with 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 there was a punishment regime 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 that they are assumed to be motivated by. 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. This particular session almost collapsed to complete free riding, which is more extreme than our other sessions, but the decline was general. After round 10 there is a "re-start" effect, which is also 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.

[16] Such variability across groups of subjects has also been reported by Herrmann, Thöni and Gächter [2008] implementing the experimental design of Fehr and Gächter [2002] in different countries.

In terms of profits, the outcome in periods 11-20 for the punishment regime is striking. The pre-punishment profits of subjects were roughly comparable to the profits earned in periods 1-10, but the post-punishment profits were much lower. This reduction is particularly evident in the first 4 periods of the punishment regime, 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 no-punishment regime.

These results from session 1 are extreme, but illustrate the factors that went into 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 counter-argument 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 institutional regimes. One could also argue that the vote was stacked against the punishment mechanism by the return to the public good being low, but again a counter-argument would be that this 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.

For completeness, we also show in Figure 3 the average contributions after the vote, in period 21. The profits for this 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 online appendix displays similarly detailed outcomes for each of the other sessions.

Figures 4 though 9 show the average "take home profits" in each period and session, along with the percentage vote for the institutional punishment mechanism.[17] 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

[17] Figures 6 and 7 contain experiments of the same general type, but conducted in different physical sessions. The same is true for the experiments in Figures 8 and 9.

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 this 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. This 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 these 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 no-punishment regime.

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 the marked difference in contributions with the change in the return to the public good, across both regimes. Round 1 deserves comment, since we see a dramatic reduction in take home earnings in both sessions, due to extravagant use of the punishment mechanism. We conjecture that this is due to some subjects learning about the nature of the punishment technology "the hard way." In one session we had one subject privately ask the experimenter, "if I punish the other person, do I get their earnings?" Of course, this had been explained in the instructions, but as every experimenter knows there are always some subjects that gloss the written and oral instructions, or do not trust them, and use the actual session to try things out. It should also be noted that we included 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 this 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 Figure 3).

Nonetheless, these 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 (Figure 4), the major change is the sequence of the regimes, with the punishment regime being experienced first. In session 2 average take home profit under the punishment mechanism was consistently around 85 cents or 90 cents after the bloodbath of period 1, but average profit was steadily just above $1.00 for the no-punishment 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 personal history of the subject influenced the vote. That is, even if average profits were lower for all subjects under the punishment mechanism in session 2 compared to the no-punishment mechanism, maybe they were higher for those 21% that voted for the institutional punishment mechanism.

Session 4 was a voting cliff-hanger, of the kind that one only finds in Florida! Contributions started out relatively high, and apart from another period 1 bloodbath, the punishment was relatively efficient 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. These are each Random Strangers sessions, which share the same P-NP history and high returns to the public good. They differ in terms of the number of subjects that were in each cohort. In session 5 we had random draws from N=10, in session 6 we had random draws from N=16, in session 12 we had random draws from N=8, and in session 13 we had random draws from N=6. Subjects were made aware of the size of their cohort. This difference provides a nice bridge between the complete absence of re-encounters in the Perfect

Strangers design, and the perfect rematching in the Partners design. With N=6 there is a higher chance of meeting the same people in later rounds than with N=16.[18]

Session 6 provided results that matched those in sessions 2 and 4, consistent with subjects being aware that the larger cohort size implied a smaller chance of a rematch with 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 Figure 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 punishment regime.

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 the use of sporadic, efficient punishment. But this was a setting in which the mere threat of the use of sanctions seemed to have the desired effect: nobody needed to "pull the trigger" since contributions were generally high and profits robust, and there were no vandals engaging 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 resulting in 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 they had another, final period with the mechanism available. They were right: in period 21 of session 5 average contributions jumped from close to 0 in period 20 to over 5. This is in itself an interesting finding since the one-shot nature of the final round could easily have caused the cooperative equilibrium under the threat of punishment to unravel, but it did not.

---

[18] Sessions 5 and 6 were conducted in the same physical session, so there were 26 subjects in the room. Computer stations were previously logged on to 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 used for sessions 7 and 8.

Sessions 12 and 13 exhibited roughly the same average profit in each regime, but the variation in profitability in the institutional punishment mechanism was striking. These sessions led to very little support for the punishment regime in the voting stage, consistent with subjects being risk averse and wanting to avoid any variation in profits that is 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 regime. However, in session 9 this was due to a precipitous dip in profits in round 20, the last one of the punishment regime; 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 only find one session in which there is majority support for the institutional punishment mechanism. This is a Random Strangers session with a small cohort, where subjects experience the punishment regime first, and where the return to contributions is high. In a similar session, but where subjects experience the non-punishment regime 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.

E. Statistical Analysis

We complement the raw observations with a statistical analysis of individual subject votes. The dependent variable is the vote for the no-punishment regime. Explanatory variables include individual demographics and treatment effects. Binary dummy variables are included for the Perfect Strangers designs, the size of the cohort conditional on the use of a Random Strangers design,[19] the history

[19] This 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. Thus it 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 is always 2, but the number of people from which the pairings

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 regime (Profit_NP), and whether the *other* player contributed more *on average* in the NP regime (Cratio_NP). Demographics include a measure of age in years, binary indicators for sex, race, academic major, class standing, cumulative GPA below 3¼, cumulative GP above 3¾, number of people in the subject's household, and a binary indicator of those that work part-time or full-time. Table 3 lists descriptive statistics for these variables, and Table 4 shows the complete set of estimates.

We are concerned *a priori* that two of the explanatory variables of particular interest might be endogenous to the vote. These variable are the measures of relative profitability of the NP and P environments to the subject placing 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 these values are predetermined by the time the vote is taken, and cannot be endogenous. But our model of voting is implicitly a model of a *latent* propensity to vote for one regime over the other, and that latent propensity might well be correlated with either of these 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.].[20] The most natural identifying assumption for this procedure is that the experimental treatments determine the profits and contributions of others, but only affect the vote 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 there is evidence of endogeneity in the relative own-profit measure when tested independently of contributions by others,[21] and we report estimates under the assumption that it is endogenous. Since we

---

were selected.

[20] This is the **ivregress** command in *Stata*.

[21] 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

cannot reject exogeneity for contributions by others when independently tested, we use 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 determines the vote. Subjects that experienced higher profit in NP compared to P were 64 percentage points more likely to vote for NP, and this is a statistically significant effect (the 95% confidence interval is between 36 percentage points and 92 percentage points). There is no statistically significant effect on voting from 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 from absence of re-encounters in the Perfect Strangers are in the expected direction, and associated with an increase in the probability of voting for NP of 20 percentage points, and the effect is weakly significant using a one-tailed test (p-value = 0.077). Related to this effect, the size of the cohort of potential opponents in the Random Strangers environment also has an effect in the expected direction: every extra cohort member is associated in this environment with a 1.6 percentage point increase in the probability of voting for NP, and again this effect is weakly statistically significant using a one-tailed test. The history experienced by the subjects significantly affected their propensity to vote for the NP regime: 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, the use of low rewards encourages outcomes in which participants are significantly more likely to vote for the NP regime, since the return to encouraging cooperation by the *efficient use of punishment* is lower. The effect of this treatment is to increase the probability of voting for the NP regime by 15 percentage points on average (p-value = 0.01).

This analysis supports our hypothesis that preferences over institutions depend primarily on the

---

hypothesis of exogeneity (p-value of 0.32).

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.

3. Related Literature

There is already an emerging literature investigating endogenous institutions, such as in constitutional votes or "voting by feet". The extent to which participants make choices that involve punishment opportunities varies, supporting our conclusions that the circumstances favoring punishment are special, involving particular experiences and the extent to which repeated game characteristics are present. The findings in this literature are also supportive of our hypothesis that earnings are an important determinant of constitutional choices.[22]

Ehrhart and Keser [1999] examine the effects of allowing "Tiebout mobility" in a basic public goods contribution game. Their idea is to allow subjects to "vote with their feet" and decide which group they would like to be in, so that individuals that have a taste for the public good could associate with like individuals. Their experiments implemented this option in a simple manner, with 9 subjects in each session being able to choose which group they wanted to participate in at the outset of each of 29 rounds after the first. Migration was costly: 50% of the endowment 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.[23] Overall, average contributions to the public

[22] Appendix B (available on request) provides additional details on the studies referenced here, as well as reviews of the designs and primary findings of Gürerk, Irlenbusch and Rockenbach [2005], Gürerk, Irlenbusch and Rockenbach [2009], Kosfeld, Okada and Riedl [2009], Gürerk, Irlenbusch and Rockenbach [2009], Gürerk [2010], Putterman, Tyran and Kamei [2011], DeAngelo and Charness [2012], Markussen, Putterman and Tyran [2013] and Drouvelis and Jamison [2015].

[23] 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-biding social contract. They implement a "Hobbesian" framework where non-cooperation involves taking the property of another person, finding that in the only one instance (out of 31

decayed steadily.

Page, Putterman and Unel [2005] extend this idea in several ways. In each session 16 subjects participate in a voluntary public goods contribution game in groups of 4 for 20 rounds. After round 3, they are allowed to individually rank the other 15 individuals, who have anonymous labels. The information available after each round is 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 the similarity in the rankings of each other. Four environments were examined. One was a *baseline* in which there was 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. They 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 when the baseline and combined environments were compared. Compared to the design in Ehrhart and Keser [1999], this design appears less vulnerable to free-riders exploiting cooperative sub-groups.

Gürerk, Irlenbusch, and Rockenbach [2006] use 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 subjects 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 Fehr and Gachter [2000][2002]. After the contribution stage, subjects receive an additional amount of 20 tokens. These extra tokens are simply retained by those in the SFI, but they may be used to punish or reward other in-group members by those in the SI. At the end of each round, subjects receive information concerning contributions, tokens given and

---

possibilities) where individuals passed the social contract deviation soon set in.

received as punishments or rewards (if in the SI), and profits for *every* subject in *both* institutions on an anonymous basis.

The overall results from 7 sessions implementing this design are striking. Initially less than 40% of the subjects join the SI, but this percentage increases steadily and after eighteen rounds over 90% of the subjects have joined. Contribution levels are substantially higher in the SI than in the SFI throughout the experiment. High contributors in the SI achieve substantially higher earnings 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. These results support our findings that choices over institutions with or without punishment mechanisms will depend on earnings. They would also 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 institutons are preferred are very special, depending on the exact experiences that participants have.

Ertan, Page and Putterman [2009], Noussair and Tan [2009] and Sutter, Haigner and Kocher [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 other alternative institutions.

Ertan, Page, and Putterman [2009] experimentally investigate 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. Then they voted for the rule that would govern their in-group interaction over the next eight rounds, and 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 in case their contribution was 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 were in favour of some punishment rule, with 72% of these allowing for punishment of lower-than-average contributors. The vast majority of the individual votes (67%) were against at least one of the possible punishment rules. Overall, 61% of the groups allowed punishment only of lower-than-average contributors, and 35% of the groups did not allow any punishment whatsoever.

The results are supportive of our finding that the particular experience that participants have affect their votes. There is a decline in the number of groups prohibiting punishment in favor of groups allowing for punishment of low contributors over time, but it seems to be more pronounced in the 3-Vote treatment than in the 5-Vote treatment. This may be due to 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 [2009] adapt the design developed by Ertan, Page and Putterman [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. Then they voted for the rule that would govern their in-group interaction 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 upon 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 these four ballot items. Considering both treatments and all voting stages, no group ever voted in favour 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). Average earnings of the subjects in the latter groups are 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 when compared to the groups in Ertan, Page and Putterman [2009].

Sutter, Haigner and Kocher [2010] investigate whether subjects prefer to interact in institutions that allow punishments, that allow rewards, or neither. They exogenously varied the intensity of the reward and punishment options. 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 takes place before participants gain 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 for 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). These results again lend support to our conclusion that the circumstances under which a constitution with costly punishments is chosen are very special.

Gintis, Bowles, Boyd and Fehr [2005; ch.1] and Boyd, Gintis, Bowles and Richerson [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 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 generic 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, it is crucial to this evolutionary argument that there be lots of such mistakes. If we insist on interpreting the presence of sanctions as evidence of an evolutionary process in cases where the outcomes are not adaptiveness peaks, then we must be observing the blind watchmaker on one of his bad days, or else there may be some other latent metric of adaptiveness we cannot (currently) observe. 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.

Fehr and Gächter [2000][2002], Gintis, Bowles, Boyd and Fehr [2005; ch.1] and Boyd, Gintis, Bowles and Richerson [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 behavior of their subjects might be motivated by "norms of fairness" or "norms of reciprocity" that might be to some extent shared. This 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 as used 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 is incompatible with revealed preference theory. Binmore [2010] defends an alternative conception of norms according to which they are evolved social conditions, roughly, unformalised institutions, that serve as equilibrium selection mechanisms by assigning asymmetric bargaining weights to occupants of different social roles.

This leaves the relationship between norms and cognitive structures, such as beliefs, in a black box. Bicchieri [2006] offers a theory of the contents of this black box. According to her, norms are best understood as shared expectations, and fall into two types. A norm can be merely *descriptive* if people in

a group all (or mostly) expect others to make choices in accord 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 Fehr and Gächter [2000][2012], and the experiment we designed, we understand the role of norms, compatibly 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." They may or may not believe that the game they play in the lab is a domain in which these norms are in force. Observed play of others in the experiment may or may not provoke revisions of these beliefs between rounds of play. And subjects might or might not operate utility functions that reflect optimal play in the game of life, which can diverge from optimal play in the experimental game.

In light of this indeterminacy, we deliberately discussed our experimental design and results in terms of observed *preferences over institutions* revealed by explicit, incentivised choices with which subjects are presented. Though we assume that these 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 these unobserved conditions.

4. Conclusions

Our experiment addresses the question of whether the desirability of social institutions, such as those that allow punishments to enforce cooperative plays, depends on the extent to which cooperative expectations are upheld or on the profits generated. Earlier studies in this domain have already 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 the use of sanctions has been largely interpreted as the individuals' desire to retaliate against those who do not comply with a cooperative norm of behavior

(e.g., Falk, Fehr and Fischbacher [2005]), the strength of this 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 the 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*[24] and the principles that apply to actions *guided by institutions*. Although an analysis of the latter enables an understanding of how institutions work, it leaves completely open questions pertaining to their origin and evolution. In the specific setting examined here the simple maximization of expected profit appears to explain the choices made by subjects when they are allowed to vote on the institution.

[24] That is, the set of rules, combined with their enforcement mechanisms, that constrain the choices of individuals.

Figure 1: Value of the Institutional Punishment Mechanism in *AER* Experiments

Data from Fehr and Gächter [2000]

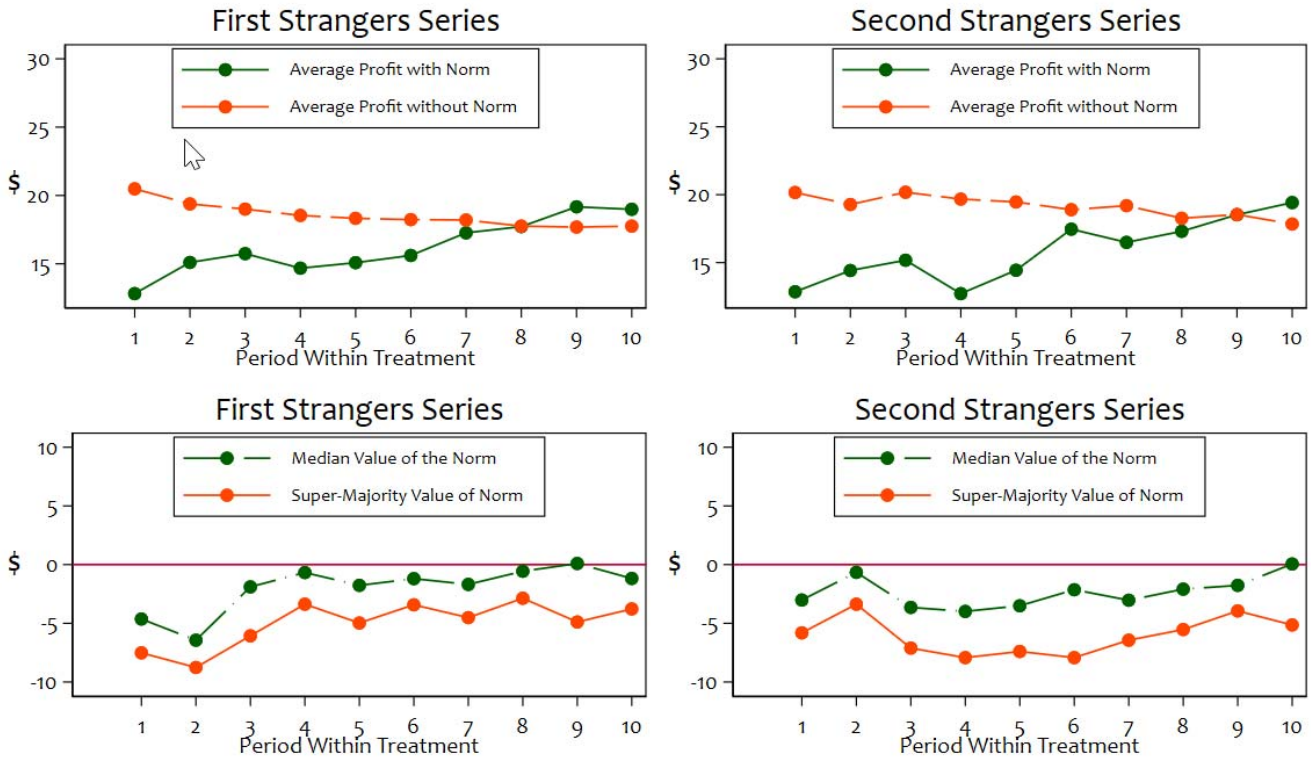


Figure 2: Value of the Institutional Punishment Mechanism in *Nature* Experiments

Data from Fehr and Gächter [2002]

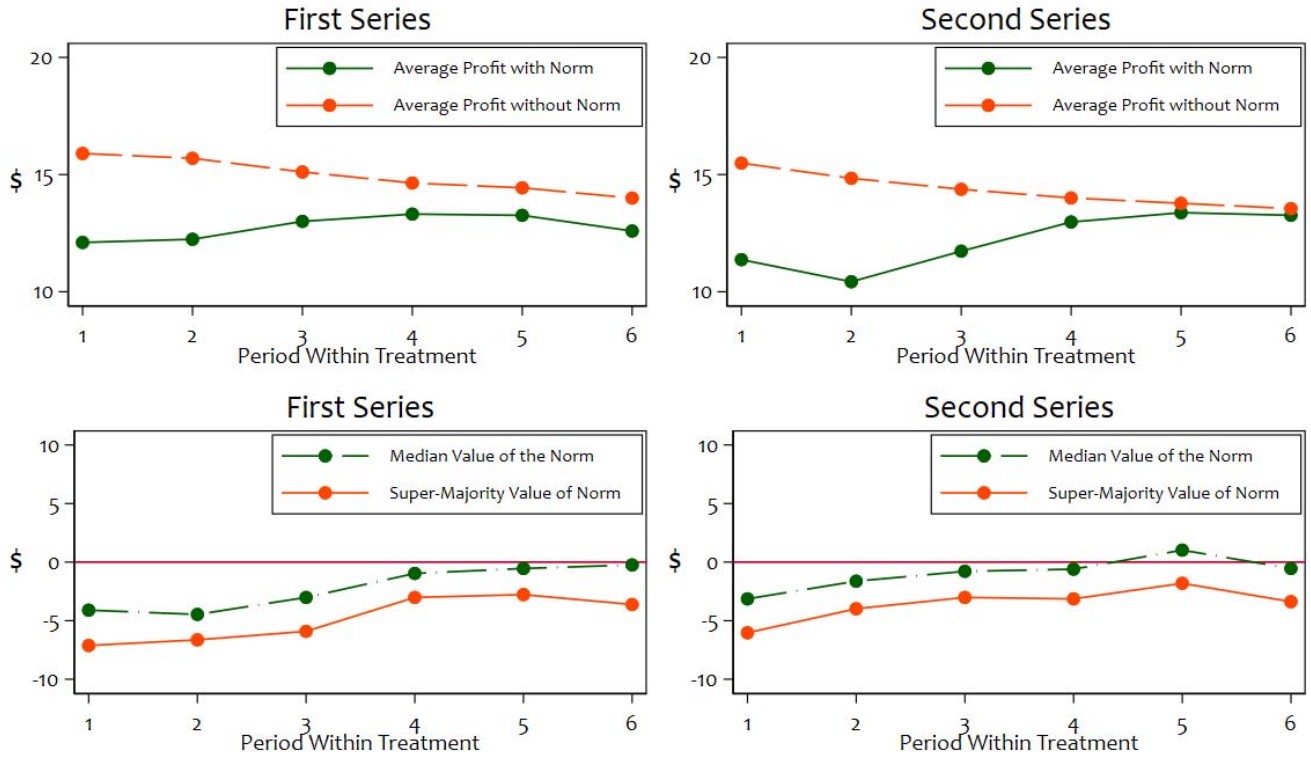


Table 1: Experimental Design

Each experiment had 10 rounds of one regime, followed by 10 rounds of the other regime
After round 20, all subjects voted on the regime for round 21
Round 21 had 10 times the payoffs of each of rounds 1-20

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%
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%

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

Figure 3: Results in Session 1

N=26 Perfect Strangers in Groups of 2
Low Return to Public Good

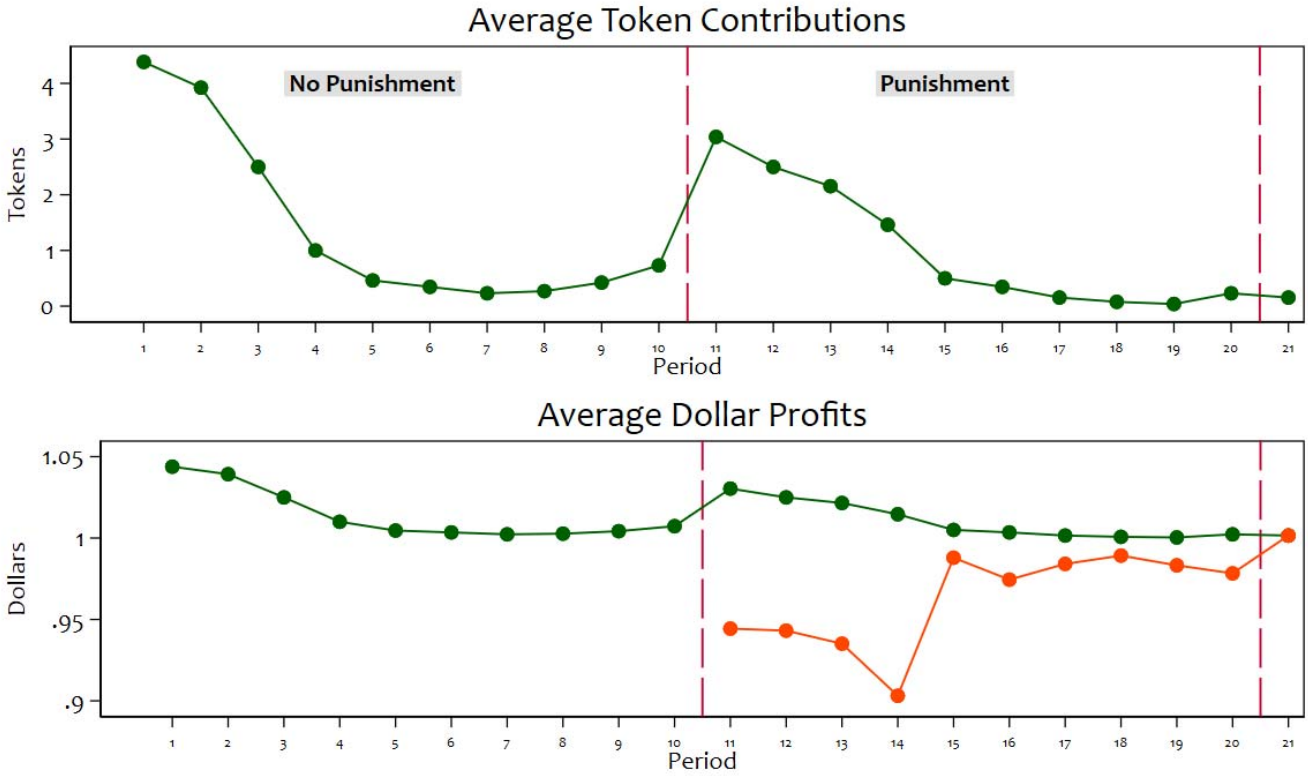


Figure 4: Average Profits With Perfect Strangers and NP-P History

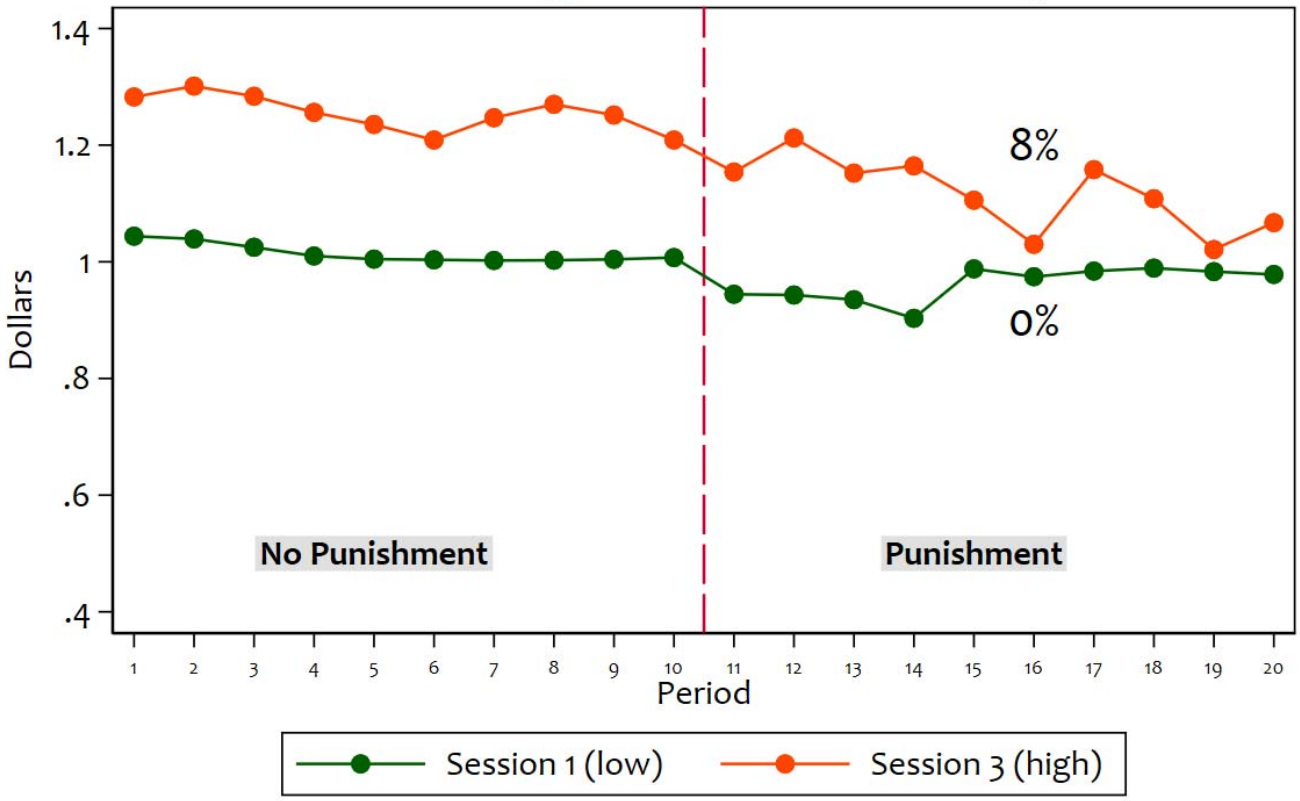


Figure 5: Average Profits With Perfect Strangers and P-NP History

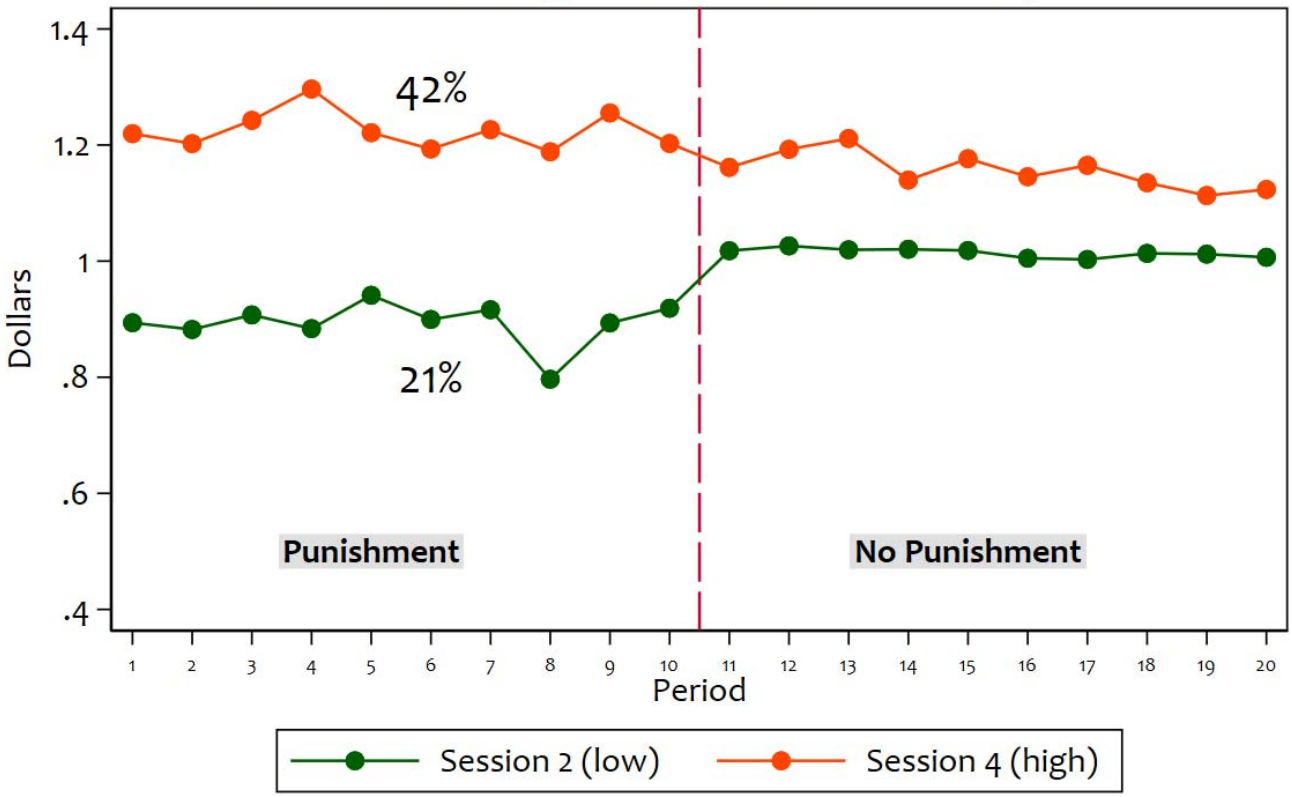


Figure 6: Average Profits With Random Strangers and P-NP History

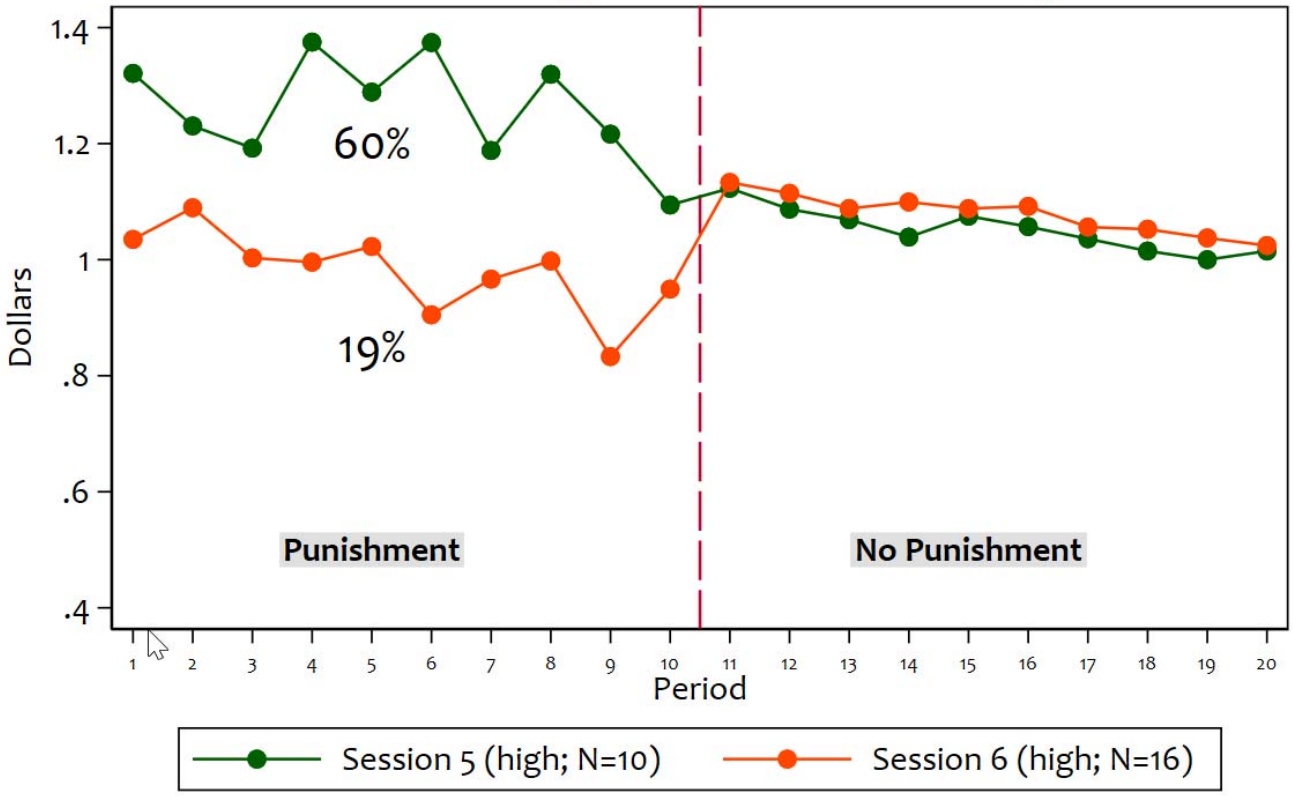
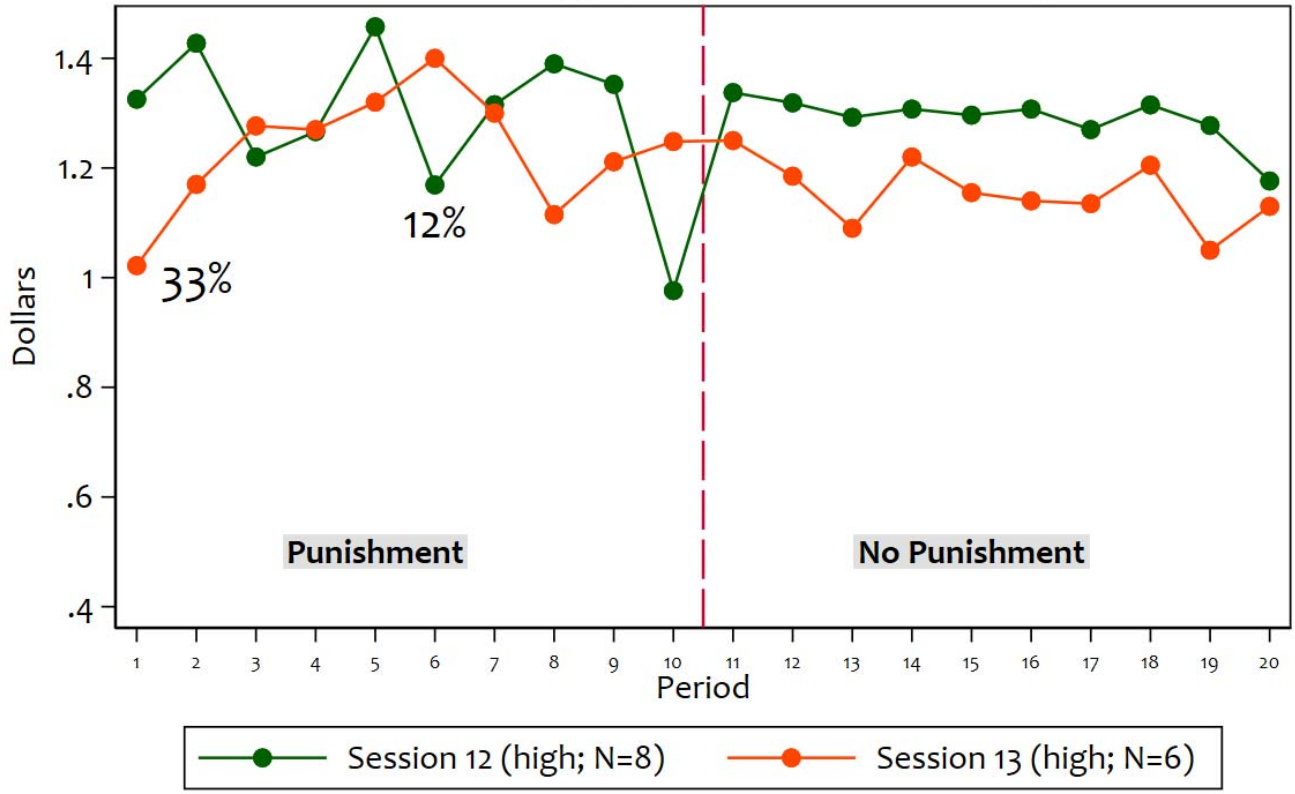


Figure 7: Average Profits With Random Strangers and P-NP History



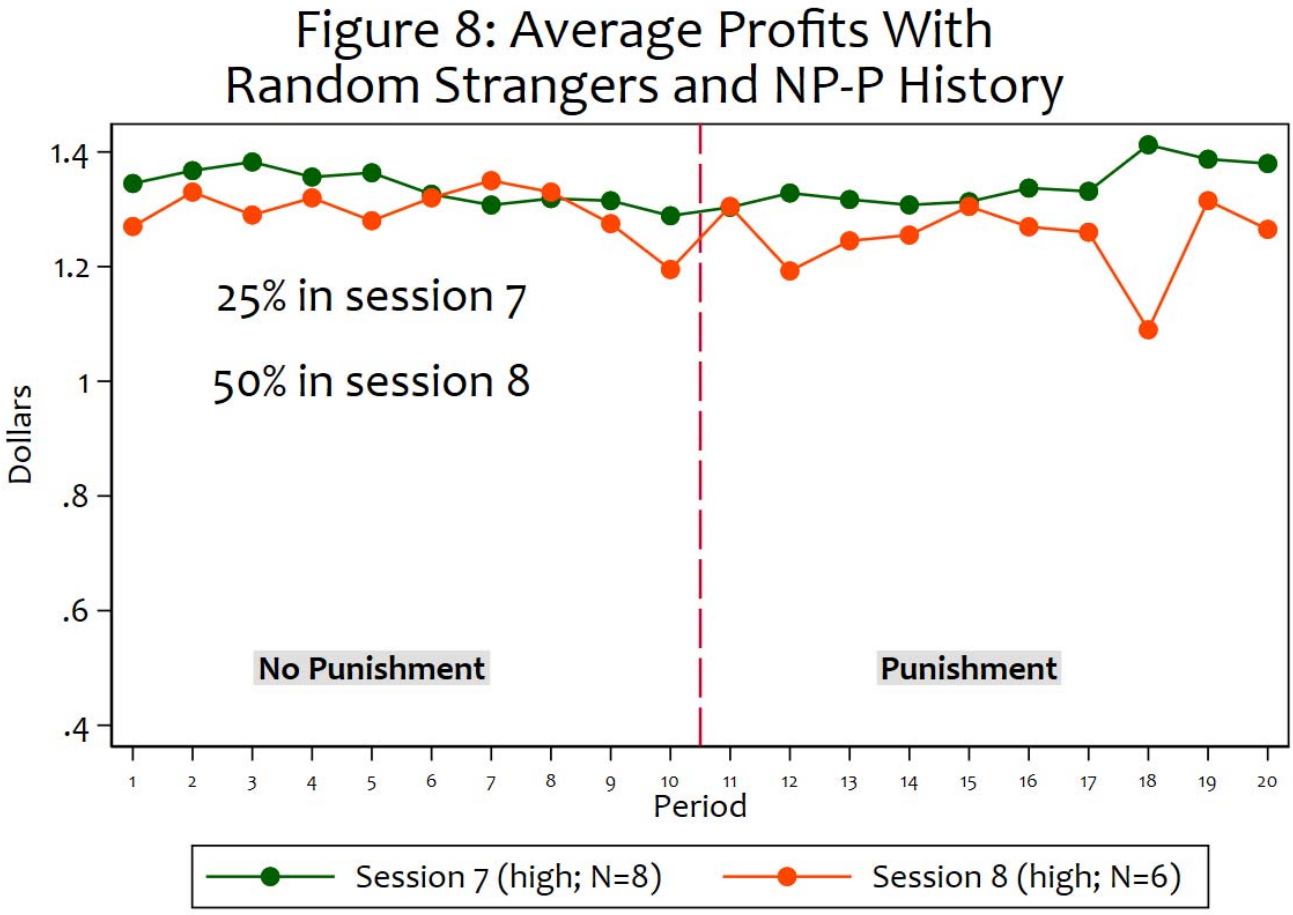

Figure 8: Average Profits With
Random Strangers and NP-P History
Dollars
1.4
1.2
1
.8
.6
.4
25% in session 7
50% in session 8
No Punishment
Punishment
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Period
Session 7 (high; N=8)
Session 8 (high; N=6)

Figure 9: Average Profits With Random Strangers and NP-P History

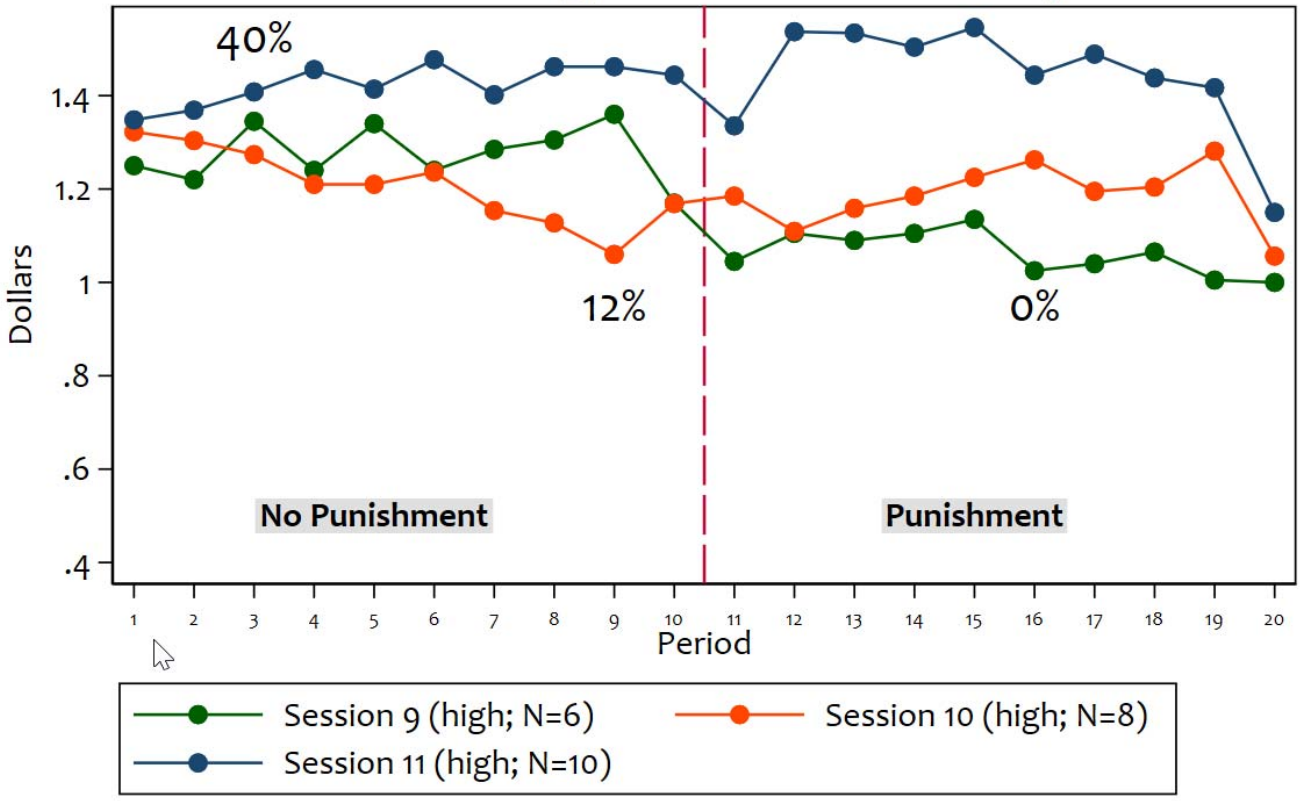


Table 3: Descriptive Statistics for Variables in Voting Model

Variable	Mean (Standard Deviation) Full Sample	High Returns Sample	Description
VoteNP	0.778	0.731	Dummy variable, 1 if vote for the no-punishment (NP) regime, 0 otherwise
Profit_NP	0.628	0.546	Dummy variable, 1 if subject received higher take home profit in the NP regime, 0 otherwise
Cratio_NP	0.311	0.308	Dummy variable, 1 if other player contributed more in the NP regime, 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 regime 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	p-value	95% Confidence Intervals	
A. Full Sample (N=180; 78% vote for NP; Wald χ^2_{14} =77.49; *p*-value = 0.0000)					
Profit_NP	0.639	0.144	0.000	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
GPAlow	-0.038	0.071	0.592	-0.178	0.101
GPAhigh	-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
B. High Returns Sample (N=130; 73% vote for NP; Wald χ^2_{14} =41.46; *p*-value = 0.0002)					
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
GPAlow	-0.021	0.097	0.828	-0.211	0.169
GPAhigh	-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

Table 5: Marginal Effects Estimated With Reduced Form Probit Model

Variable	Estimate	Standard Error	p-value	95% Confidence Intervals	
A. Full Sample (N=180; Wald χ^2_{16} =39.66; *p*-value = 0.0009)					
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	0.000	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
B. High Returns Sample (N=130; Wald χ^2_{15} =21.49; *p*-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

References

Anderson, Christopher M., and Putterman, Louis, "Do Non-Strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism," *Games and Economic Behavior*, 54(1), 2006, 1-24.

Andreoni, James, and Croson, Rachel T.A., "Partners versus Strangers: Random Rematching in Public Goods Experiments," in C.R. Plott and V.L. Smith (eds.), *Handbook of Experimental Economics Results* (North-Holland: Amsterdam, 2005).

Bicchieri, Cristina, *The Grammar of Society* (Cambridge: Cambridge University Press, 2006).

Binmore, Ken, *Game Theory and the Social Contract Volume 1: Just Playing* (Cambridge, MA: MIT Press, 1994).

Binmore, Ken, *Game Theory and the Social Contract Volume 2: Playing for Real* (Cambridge, MA: MIT Press, 1998).

Binmore, Ken, "Social Norms or Social Preferences?" *Mind and Society*, 9(2), 2010, 139-157.

Bischoff, Ivo, "Institutional Choice versus Communication in Social Dilemmas - An Experimental Approach," *Journal of Economic Behavior & Organization*, 62, 2007, 20-36.

Bochet, Olivier; Page, Talbot, and Putterman, Louis, "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior & Organization*, 60, 2006, 11-26.

Botelho, Anabela; Harrison, Glenn W.; Pinto, Lígia M.Costa and Rutström, Elisabet E., "Testing Static Game Theory with Dynamic Experiments: A Case Study of Public Goods," *Games and Economic Behavior*, 67(1), 2009, 253-265.

Boyd, Robert; Gintis, Herbert, Bowles, Samuel, and Richerson, Peter J., "The Evolution of Altruistic Punishment," in H. Gintis, S. Bowles, R. Boyd and E. Fehr (eds.), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (Cambridge, MA: MIT Press, 2005).

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

Carpenter, Jeffrey, and Matthews, Peter, "Social Reciprocity," *Working Paper 0229r*, Department of Economics, Middlebury College, 2004.

Casari, Marco, and Luini, Luigi, "Group Cooperation Under Alternative Peer Punishment Technologies: An Experiment," *Working Paper #1176*, Krannert School of Management, Purdue University, 2005.

Dawkins, Richard, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design* (New York: Norton, 1986).

DeAngelo, Greg, and Charness, Gary, "Deterrence, Expected Cost, Uncertainty and Voting: Experimental Evidence," *Journal of Risk and Uncertainty*, 44(1), 2012, 73-100.

Drouvelis, Michalis, and Jamison, Julian C., "Selecting Public Goods Institutions: Who Likes to Punish and Reward?" *Southern Economic Journal*, 82(2), 2015, 501-534.

Egas, Martijn, and Riedl, Arno, "The Economics of Altruistic Punishment and the Maintenance of Cooperation," *Proceedings of the Royal Society B – Biological Sciences*, 275, 2008, 871-878.

Erhart, Karl-Martin, and Keser, Claudia, "Mobility and Cooperation: On the Run," *Working Paper 99s-24*, CIRANO, University of Montreal, June 1999.

Ertan, Arhan; Page, Talbot, and Putterman, Louis, "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free Rider Problem," *European Economic Review*, 53(5), 2009, 495-511.

Ertan, Arhan; Page, Talbot, and Putterman, Louis, "Can Endogenously Chosen Institutions Mitigate the Free-Rider Problem and Reduce Perverse Punishment?" *Working Paper No. 2005-13*, Department of Economics, Brown University, 2015.

Falk, Armin; Fehr, Ernst, and Fischbacher, Urs, "Driving Forces Behind Informal Sanctions," *Econometrica*, 73(6), 2005, 2017-2030.

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

Fehr, Ernst, and Gächter, Simon, "Altruistic Punishment in Humans," *Nature*, 415, 10 January 2002, 137-140.

Fehr, Ernst, and Rockenbach, Bettina, "Detrimental Effects of Sanctions on Human Altruism," *Nature*, 422, 13 March 2003, 137-140.

Fischbacher, Urs, "z-Tree - Zurich Toolbox for Readymade Economic Experiments," *Experimental Economics*, 10(2), June 2007, 171-178.

Gintis, Herbert; Bowles, Samuel; Boyd, Robert, and Fehr, Ernst (eds.), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (Cambridge, MA: MIT Press, 2005).

Goeree, Jacob K.; Holt, Charles A., and Laury, Susan K., "Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior," *Journal of Public Economics*, 83, 2002, 255–276.

Gürerk, Özgür; Irlenbusch, Bernd, and Rockenbach, Bettina, "On the Evolvement of Institutions in Social Dilemmas," *Working Paper*, University of Erfurt, 2005.

Gürerk, Özgür; Irlenbusch, Bernd, and Rockenbach, Bettina, "The Competitive Advantage of Sanctioning Institutions," *Science*, 312, April 7, 2006, 108-111.

Gürerk, Özgür, Irlenbusch, Bernd, and Rockenbach, Bettina, "Motivating Teammates: The Leader's Choice Between Positive and Negative Incentives," *Journal of Economic Psychology*, 30(4), 2009a, 591-607.

Gürerk, Özgür, Irlenbusch, Bernd, and Rockenbach, Bettina, "Voting with Feet: Community Choice in Social Dilemmas," *Working Paper 4643*, Institute for the Study of Labor, 2009b.

Harrison, Glenn W., and Hirshleifer, Jack, "An Experimental Evaluation of Weakest-Link/ Best-Shot Models of Public Goods," *Journal of Political Economy*, 97, February 1989, 201-225.

Herrmann, Benedikt; Thöni, Christian, and Gächter, Simon, "Antisocial Punishment Across Societies," *Science*, 319, 7 March 2008, 1362-1367.

Isaac, R. Mark and Walker, James M., "Communication and Free-Riding Behavior: the Voluntary Contribution Mechanism," *Economic Inquiry*, 26(4), 1988a, 585-608.

Isaac, R. Mark and Walker, James M., "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics*, 53, 1988b, 179-200.

Kosfeld, Michael; Okada, Akira, and Riedl, Arno, "Institution Formation in Public Goods Games," *American Economic Review*, 99(4), September 2009, 1335-1355.

Masclet, David; Noussair, Charles; Tucker, Steven, and Villeval, Marie-Claire, "Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism," *American Economic Review*, 93(1), March 2003, 366-380.

Markussen, T., Putterman, Louis, & Tyran, J. R., "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes," *Review of Economic Studies*, 2013.

Nikiforakis, Nikos, "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics*, 92, 2008, 91-112.

Nikiforakis, Nikos, and Normann, Hans-Theo, "A Comparative Statics Analysis of Punishment in Public Good Experiments," *Experimental Economics*, 11(4), 2008, 358-369.

Noussair, Charles N. and Tan, Fangfang, "Voting on Punishment Systems within a Heterogeneous Group," *Discussion Paper No. 2009-19*, CentER, Tilburg University, March 2009.

Ostrom, Elinor; Walker, James, and Gardner, Roy, "Covenants With and Without a Sword: Self-Governance Is Possible," *American Journal of Political Science*, 86(2), June 1992, 404-417.

Page, Talbot; Putterman, Louis, and Unel, Bulent, "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency," *Economic Journal*, 115, October 2005, 1037-1058.

Palfrey, Thomas R., and Prisbrey, Jeffrey E., "Altruism, Reputation, and Noise in Linear Public Goods Experiments," *Journal of Public Economics*, 61, 1996, 409-427.

Palfrey, Thomas R., and Prisbrey, Jeffrey E., "Anomalous Behavior in Linear Public Goods Experiments: How Much and Why?" *American Economic Review*, 87, 1997, 829–846.

Powell, Benjamin, and Wilson, Bart J., "An Experimental Investigation of Hobbesian Jungles," *Journal of Economic Behavior & Organization*, 66(3-4), 2008, 669-686.

Putterman, Louis; Tyran, J. R., and Kamei, K., "Public Goods and Voting on Formal Sanction Schemes," *Journal of Public Economics*, 95(9), 2011, 1213-1222.

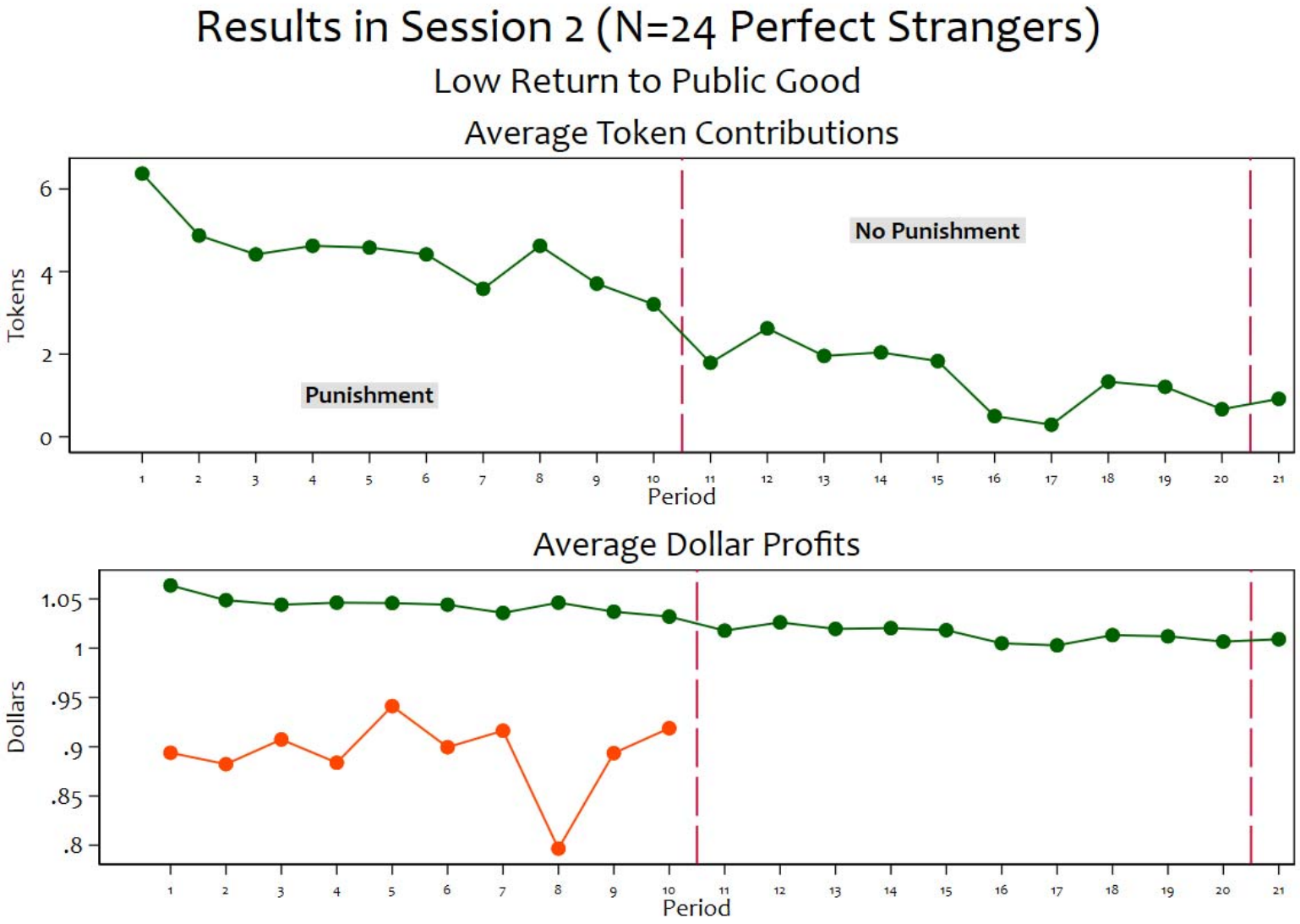
Sefton, Martin; Shupp, Robert S., and Walker, James, "The Effect of Rewards and Sanctions in Provision of Public Goods," *Economic Inquiry*, 45(4), 2007, 671-690.

Simonsohn, Uri, "Review of *Moral Sentiments and Material Interests*," *Journal of Economic Literature*, XLIV, September 2006, 745-747.

StataCorp, *Stata Base Reference Manual: Release 15* (College Station, TX: Stata Corp. LLC, 2017).

Sutter, Matthias; Haigner, Stefan, and Kocher, Martin,. "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations," *Review of Economic Studies*, 77(4), 2010, 1540-1566.

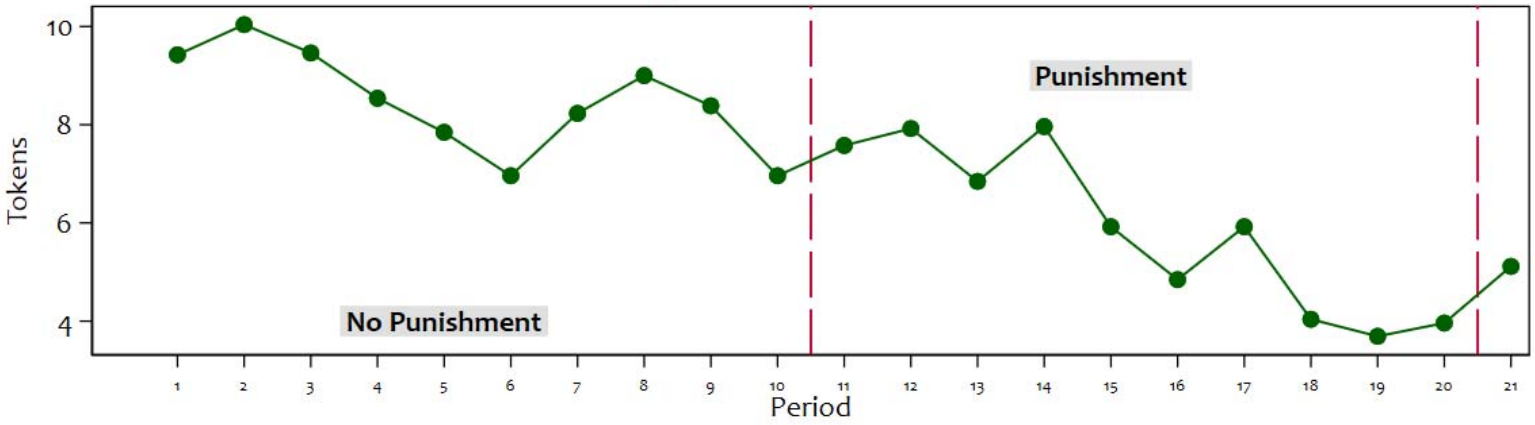
Appendix A: Results for Each Session (NOT FOR PUBLICATION)



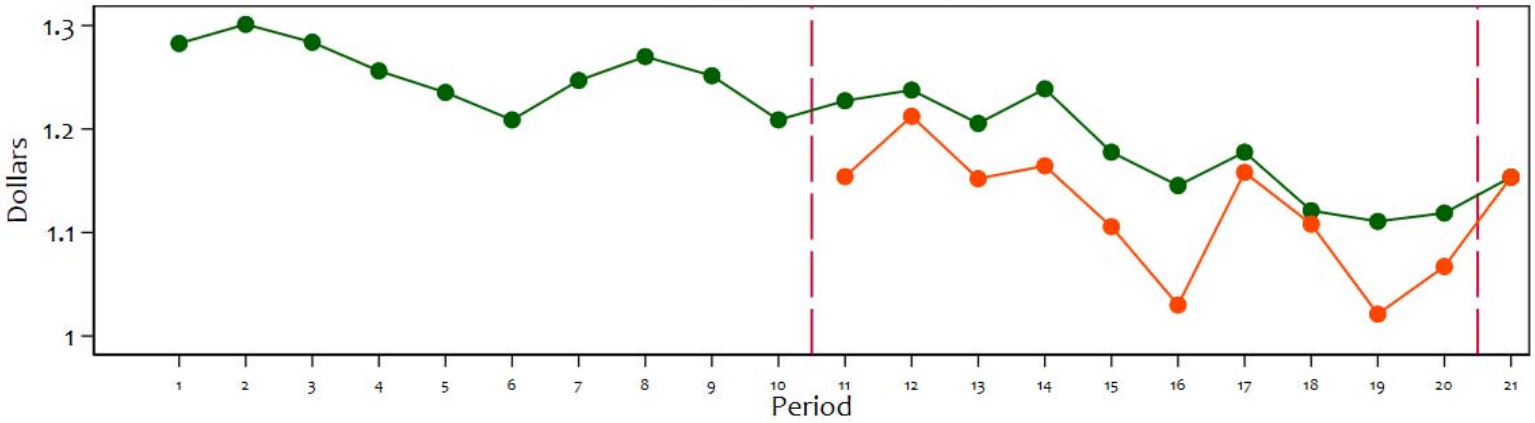
Results in Session 3 (N=26 Perfect Strangers)

High Return to Public Good

Average Token Contributions



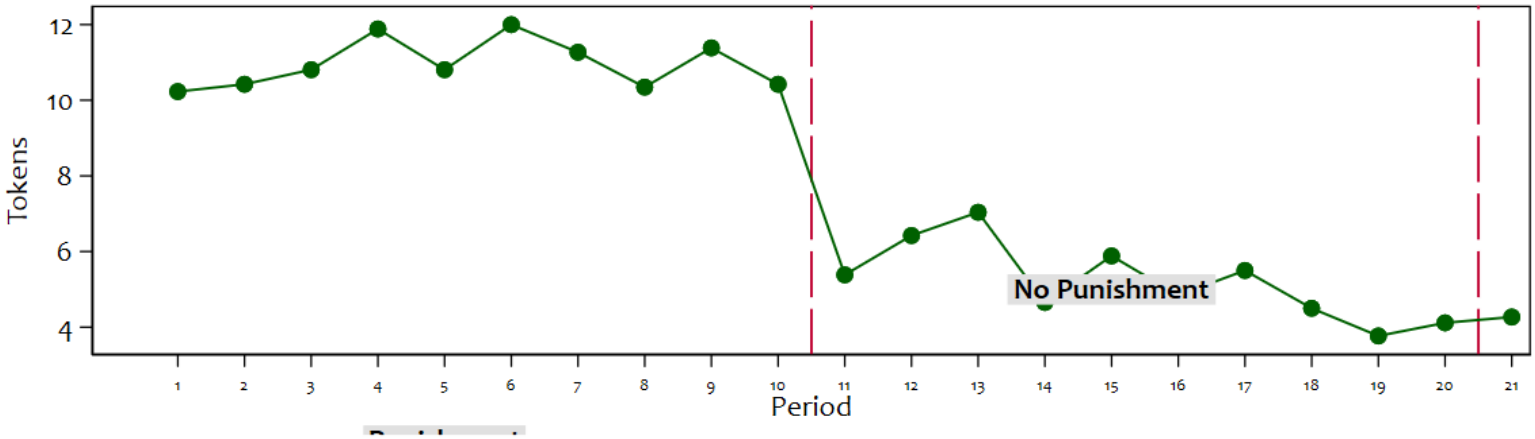
Average Dollar Profits



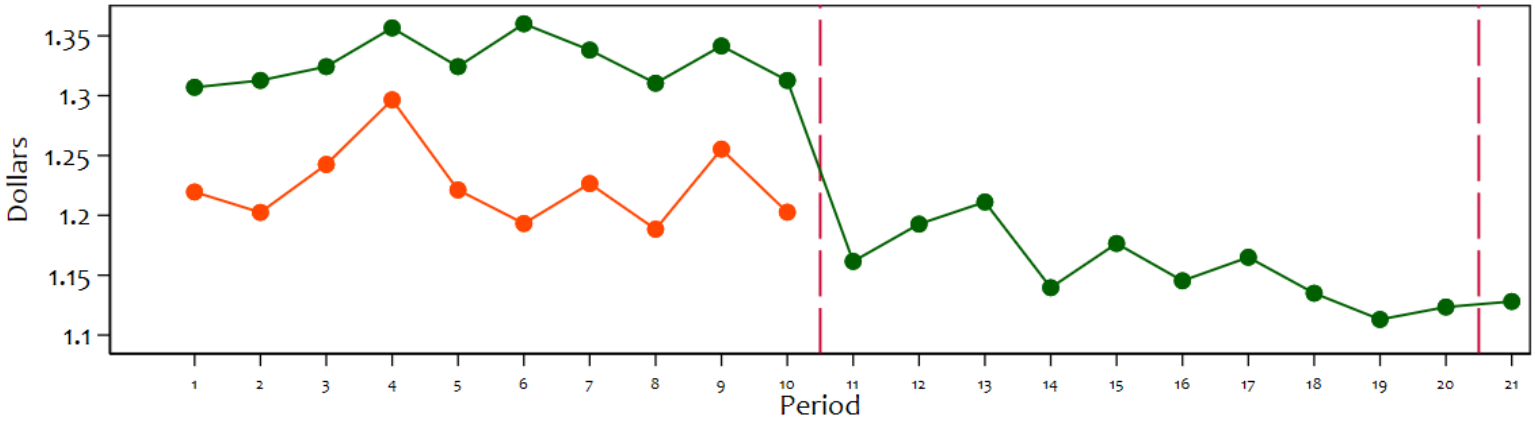
Results in Session 4 (N=26 Perfect Strangers)

High Return to Public Good

Average Token Contributions



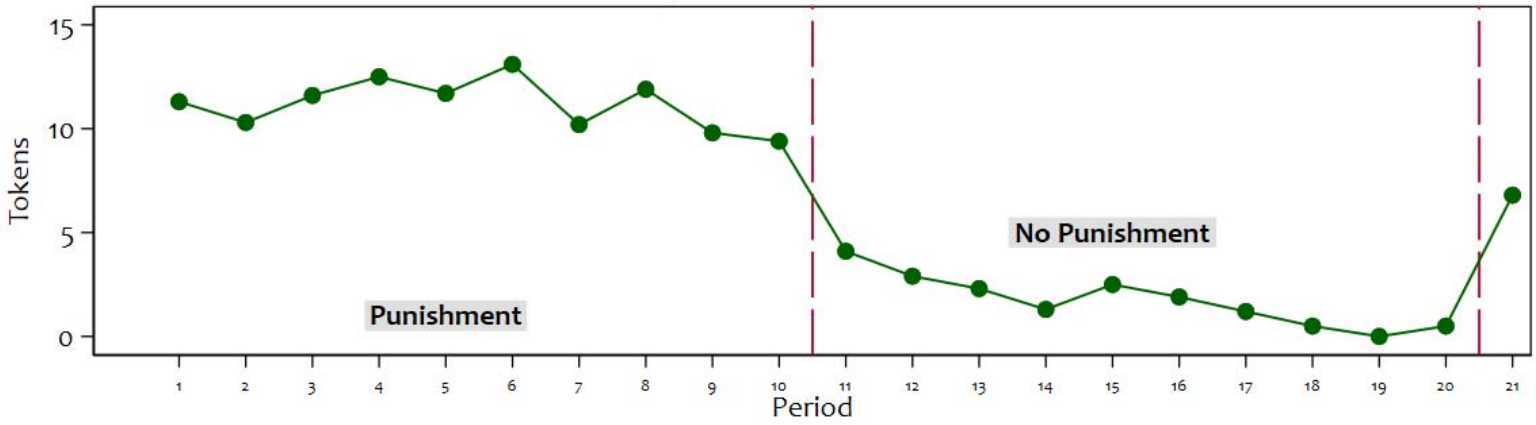
Average Dollar Profits



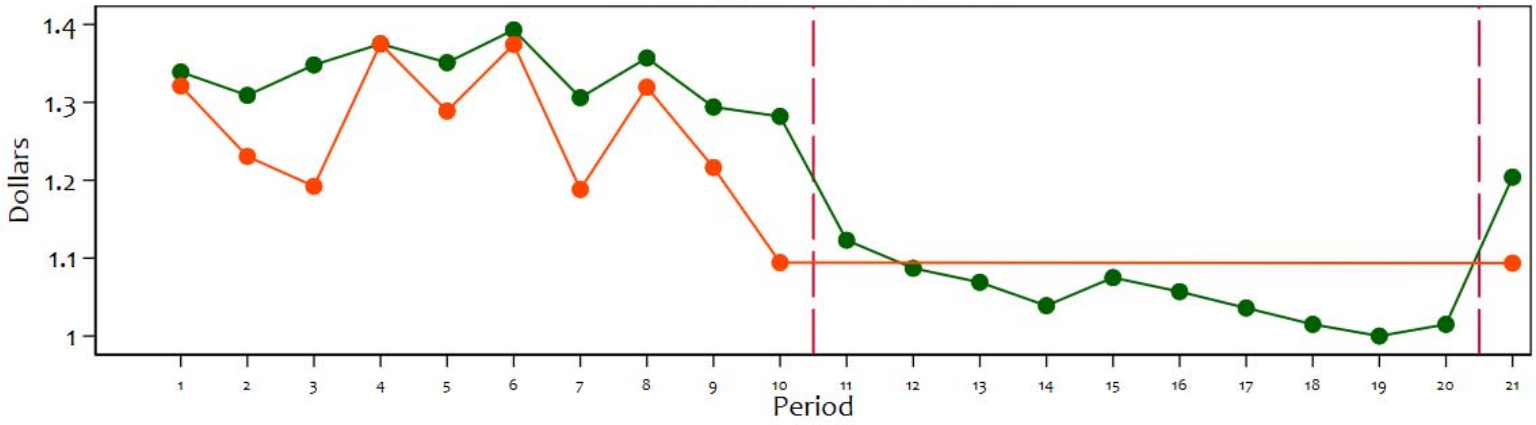
Results in Session 5 (N=10 Random Strangers)

High Return to Public Good

Average Token Contributions



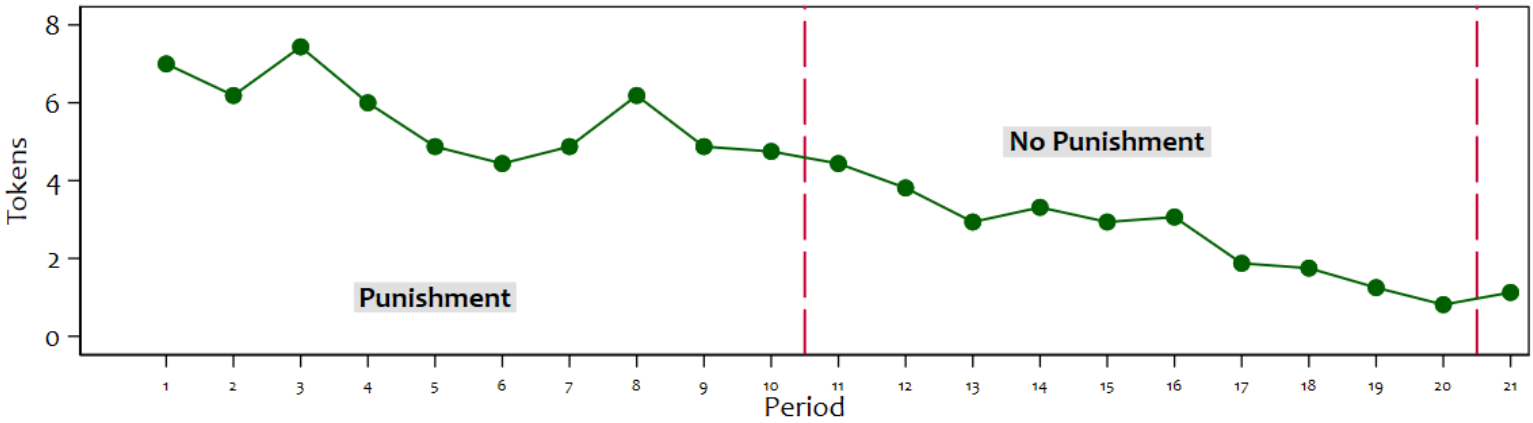
Average Dollar Profits



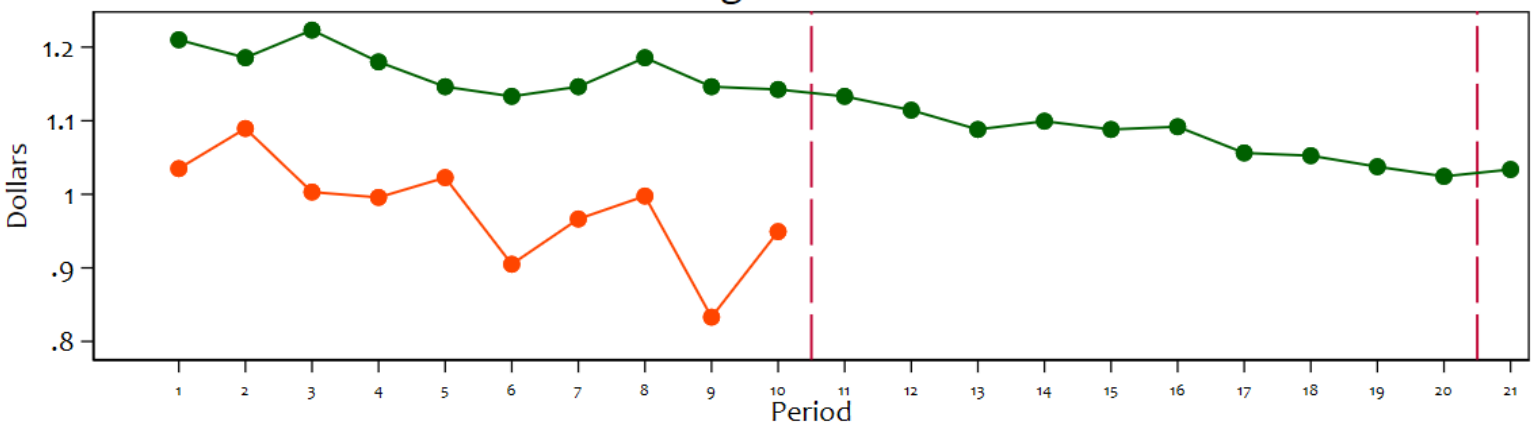
Results in Session 6 (N=16 Random Strangers)

High Return to Public Good

Average Token Contributions



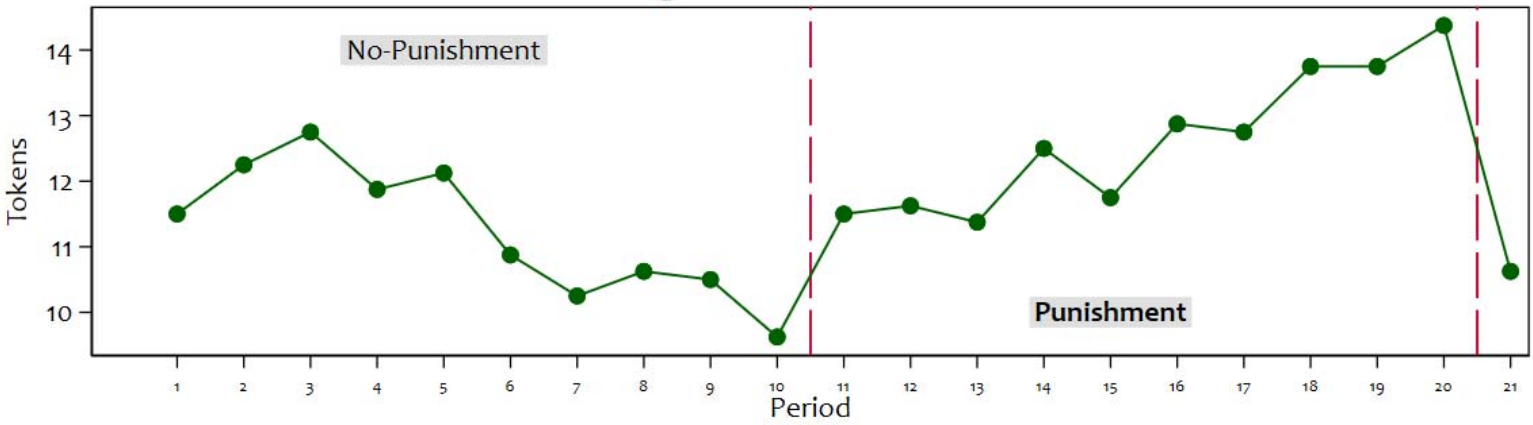
Average Dollar Profits



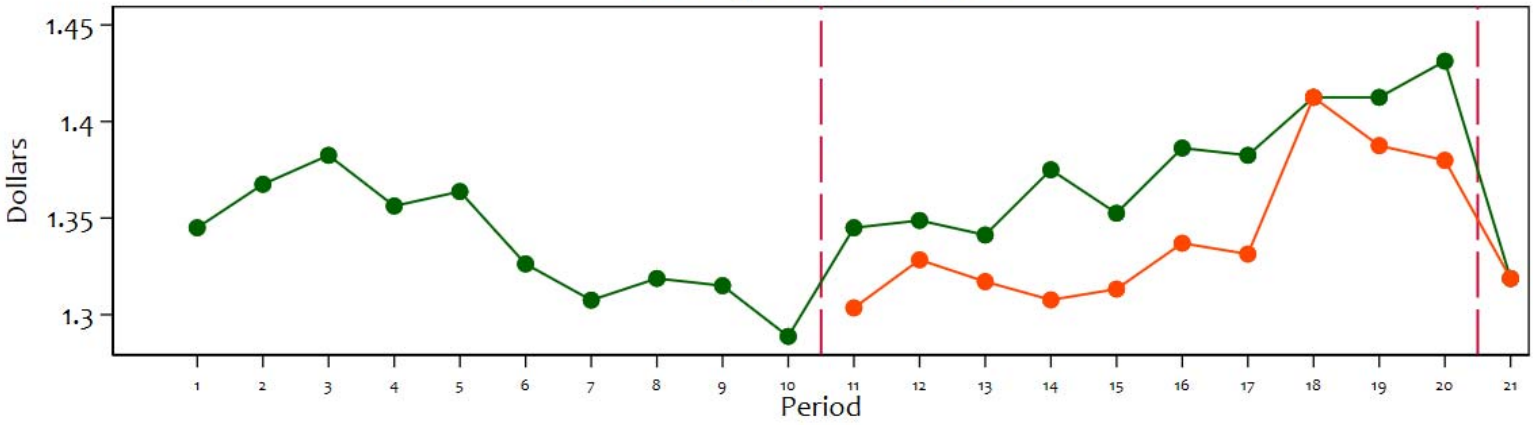
Results in Session 7 (N=8 Random Strangers)

High Return to Public Good

Average Token Contributions



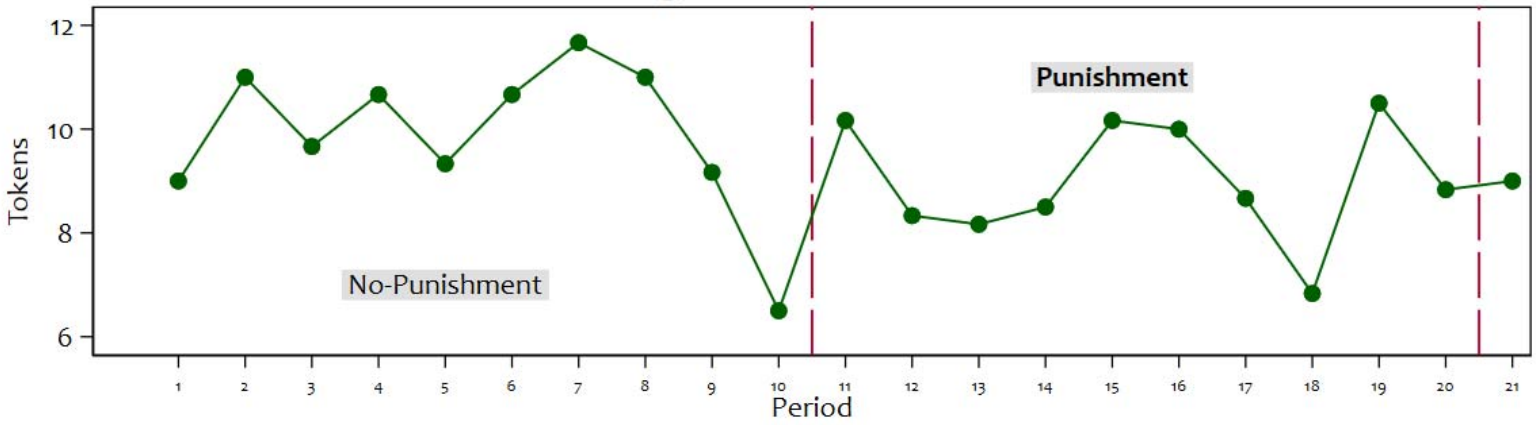
Average Dollar Profits



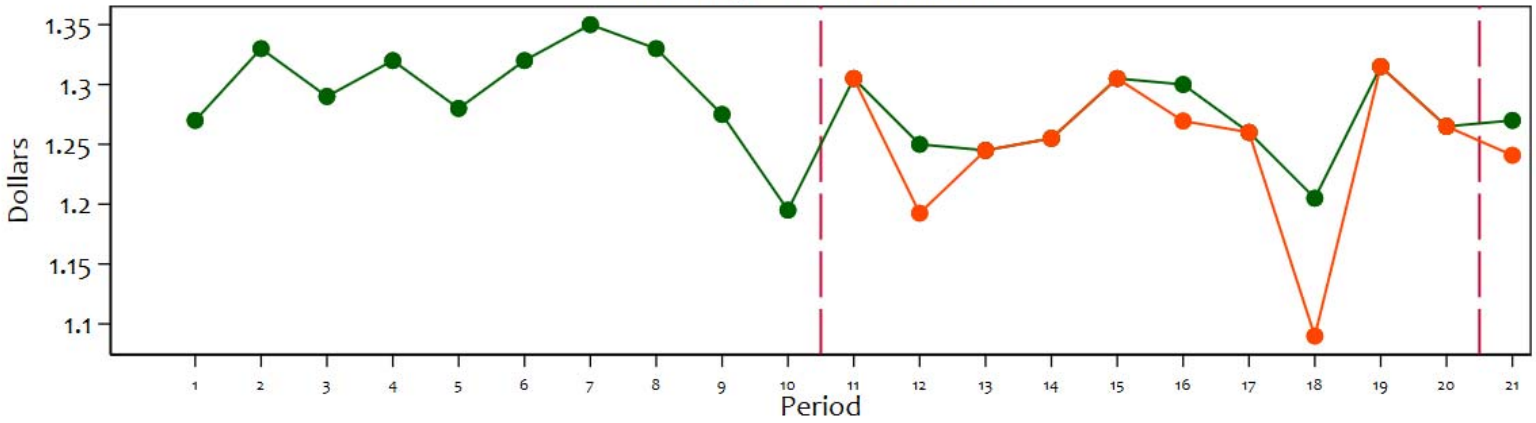
Results in Session 8 (N=6 Random Strangers)

High Return to Public Good

Average Token Contributions



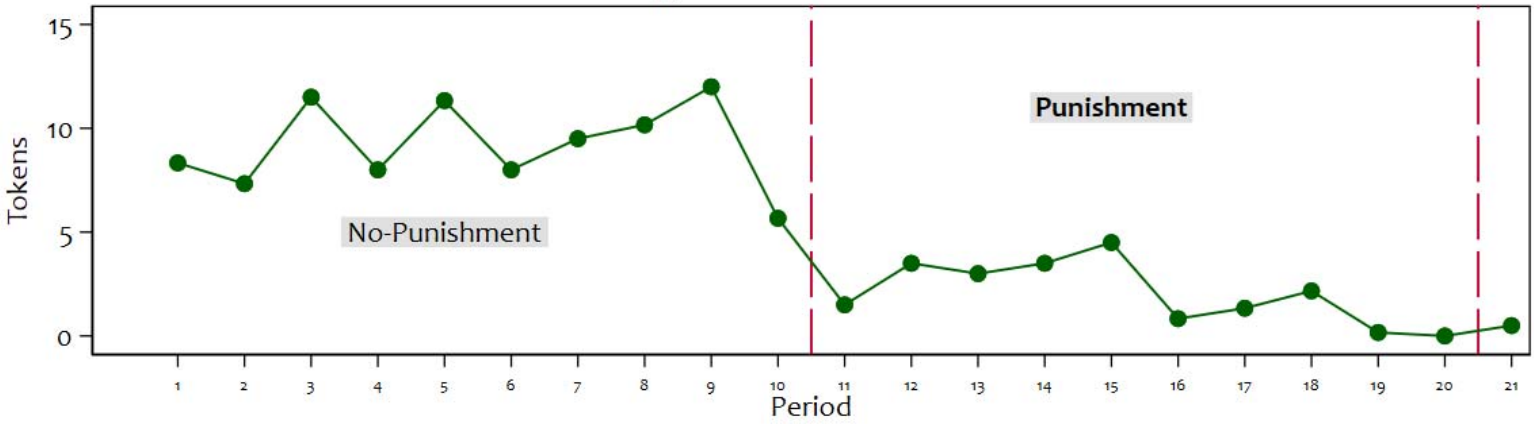
Average Dollar Profits



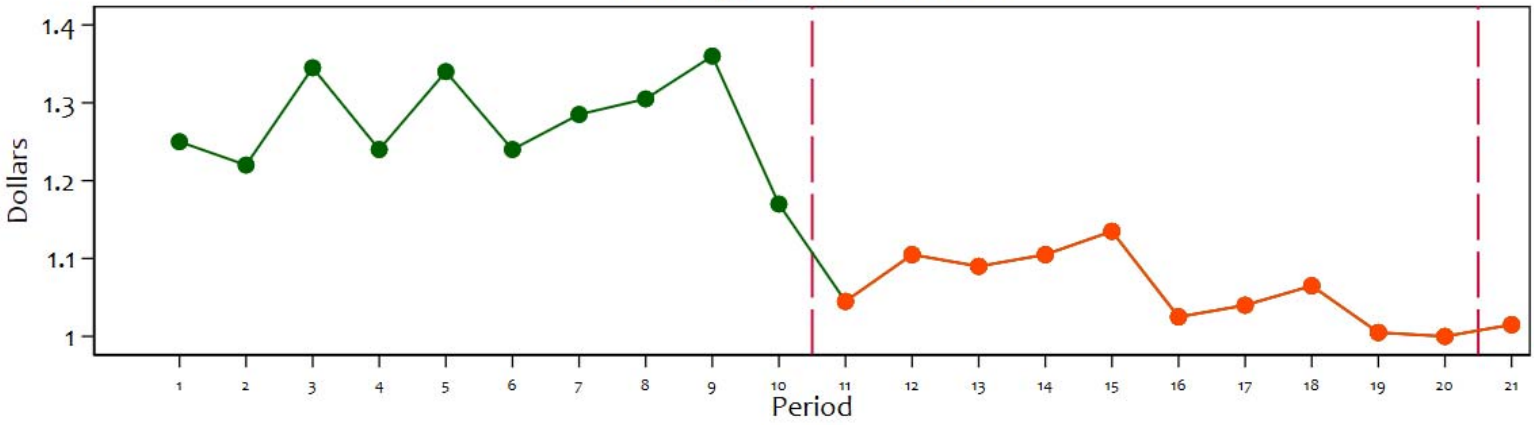
Results in Session 9 (N=6 Random Strangers)

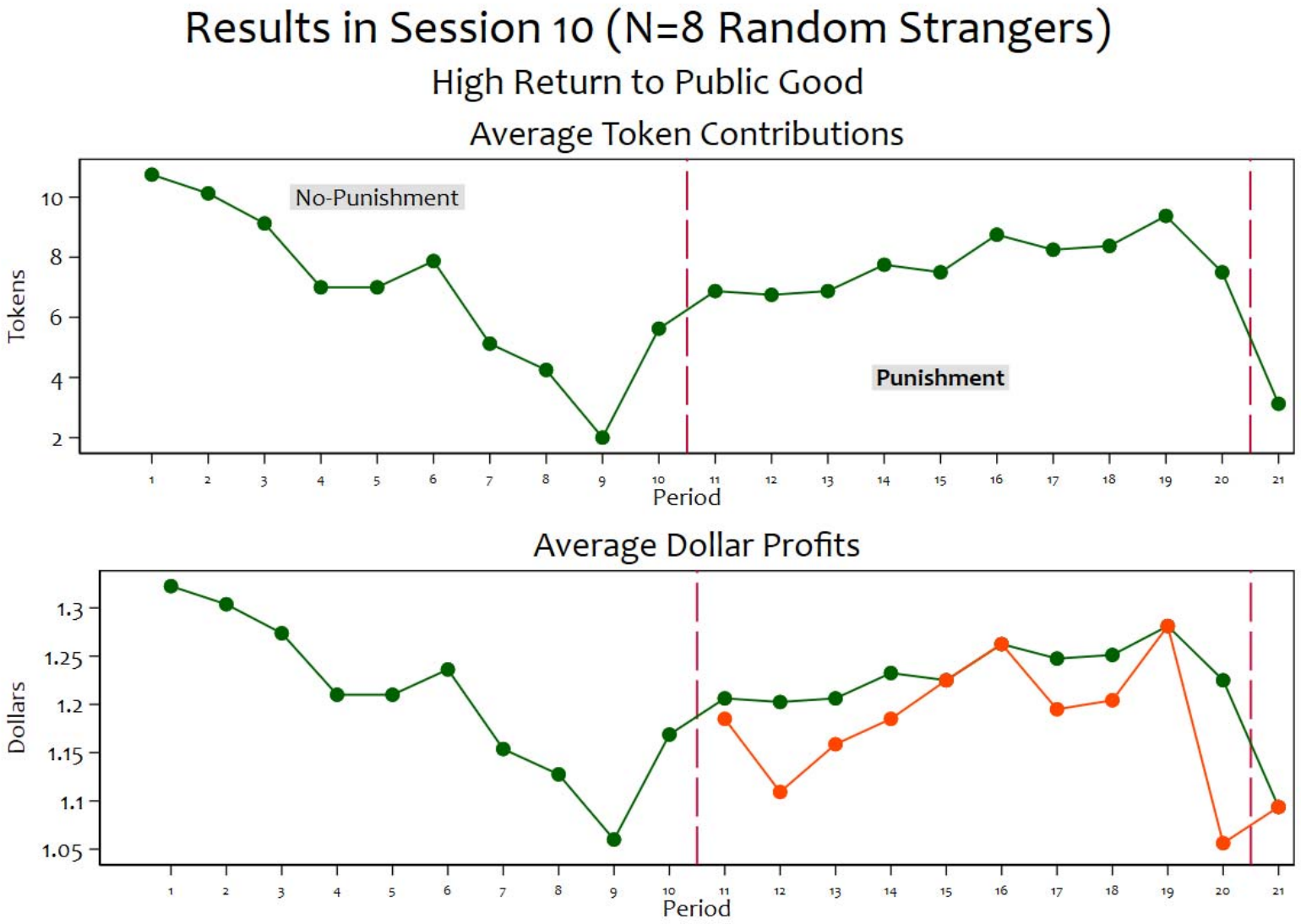
High Return to Public Good

Average Token Contributions



Average Dollar Profits

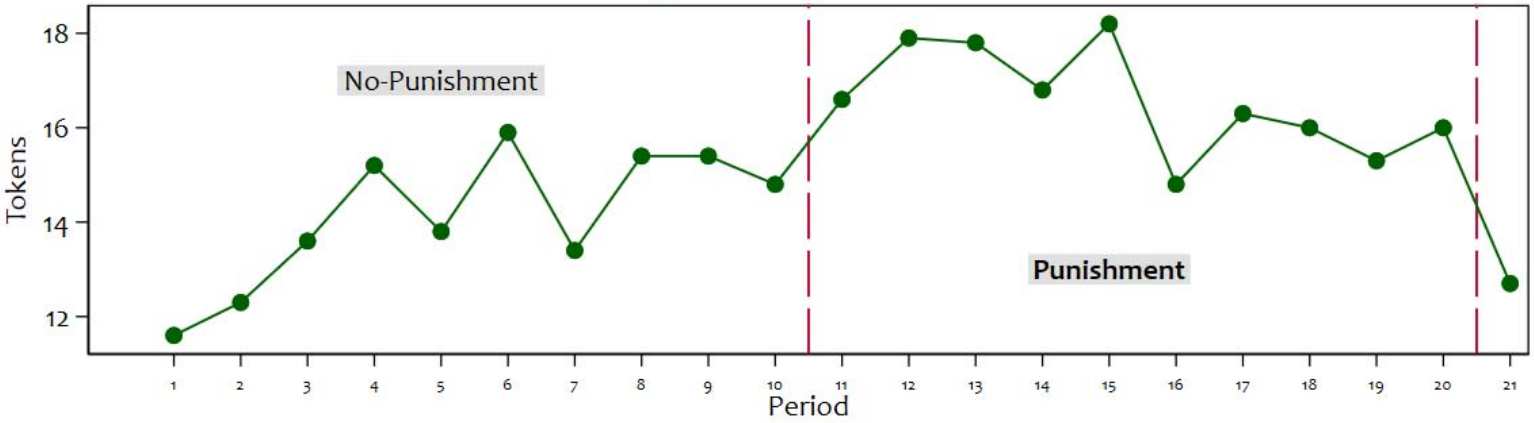



Results in Session 10 (N=8 Random Strangers)
High Return to Public Good
Average Token Contributions
No-Punishment
Punishment
Tokens
Period
Average Dollar Profits
Dollars
Period

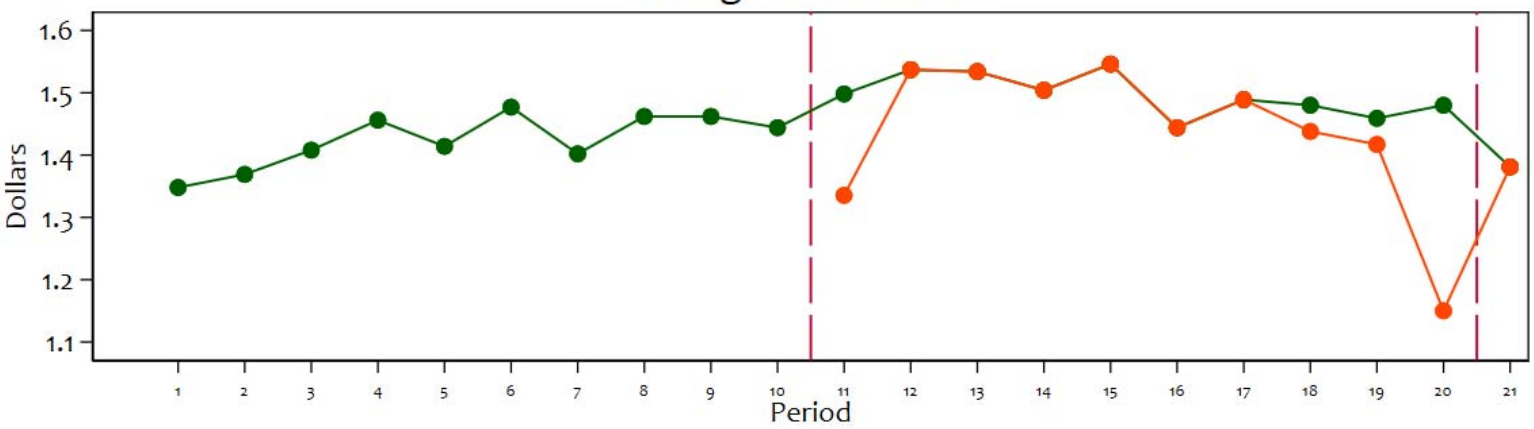
Results in Session 11 (N=10 Random Strangers)

High Return to Public Good

Average Token Contributions



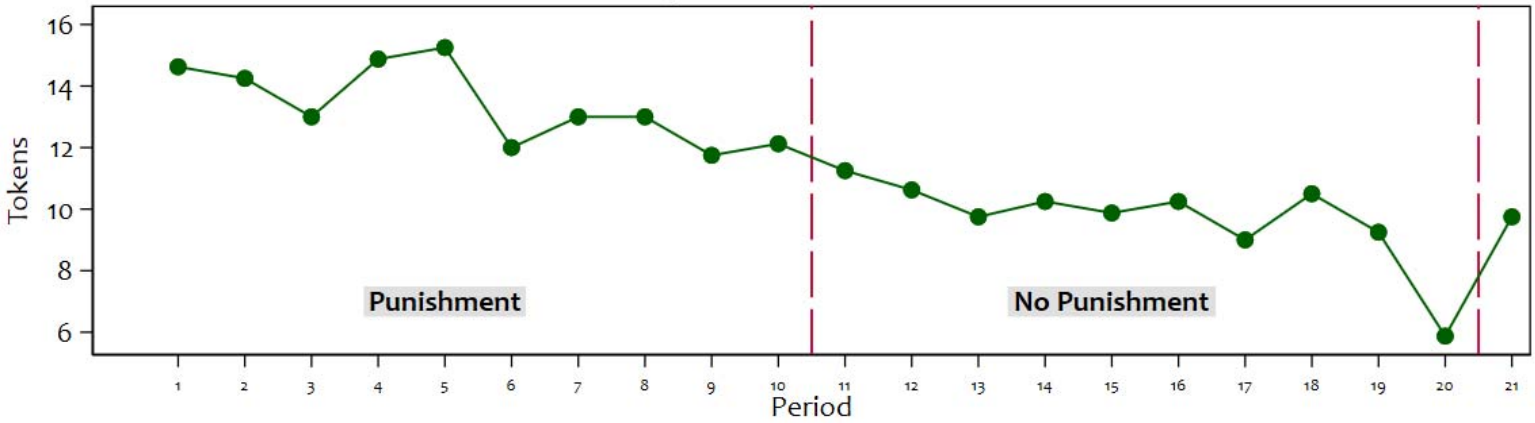
Average Dollar Profits



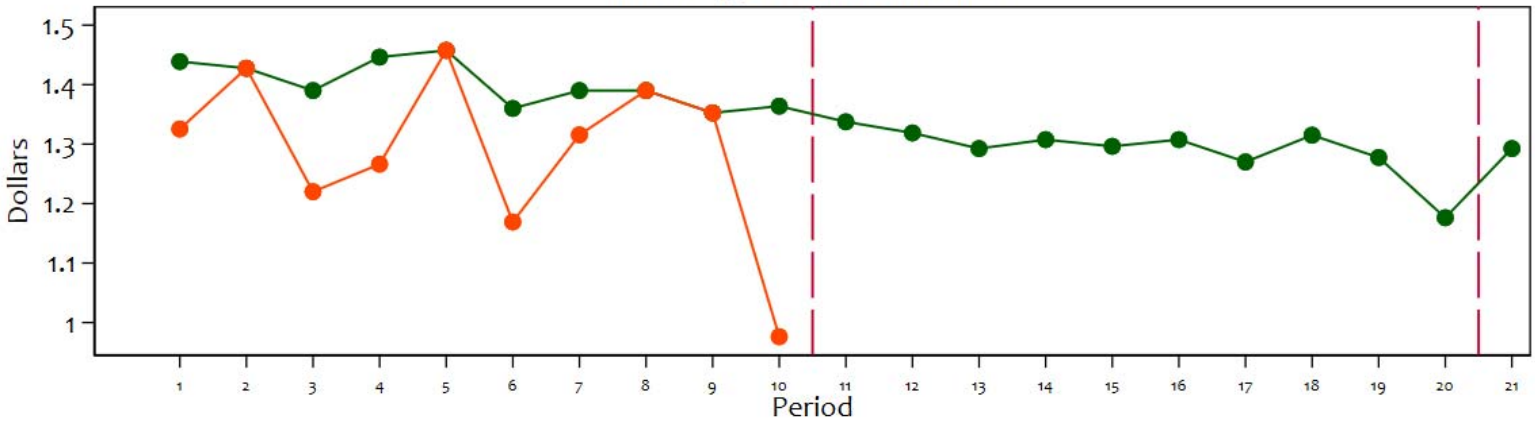
Results in Session 12 (N=8 Random Strangers)

High Return to Public Good

Average Token Contributions



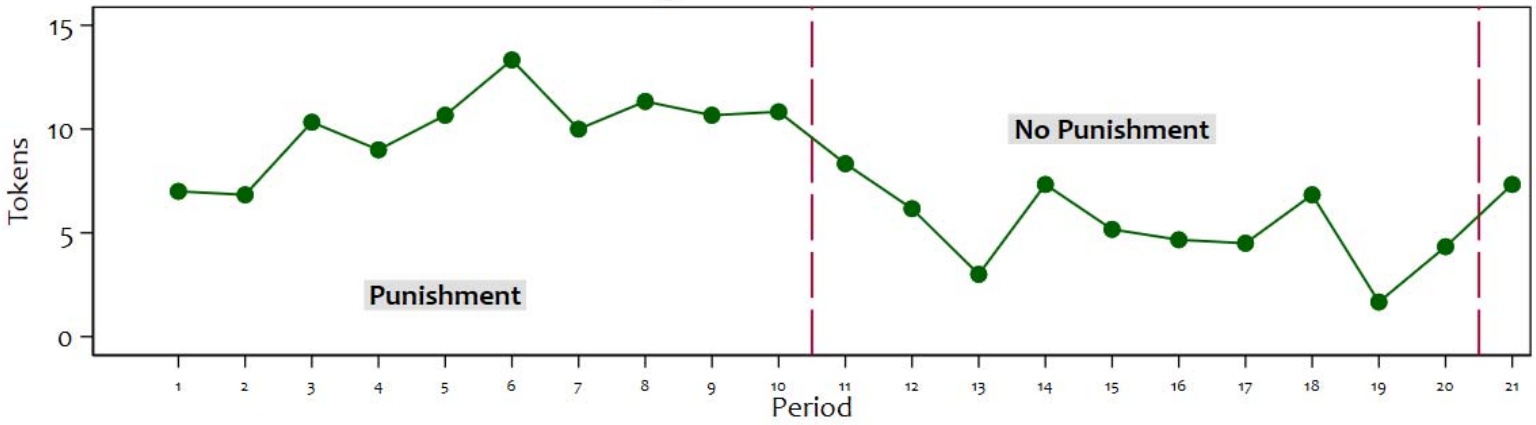
Average Dollar Profits



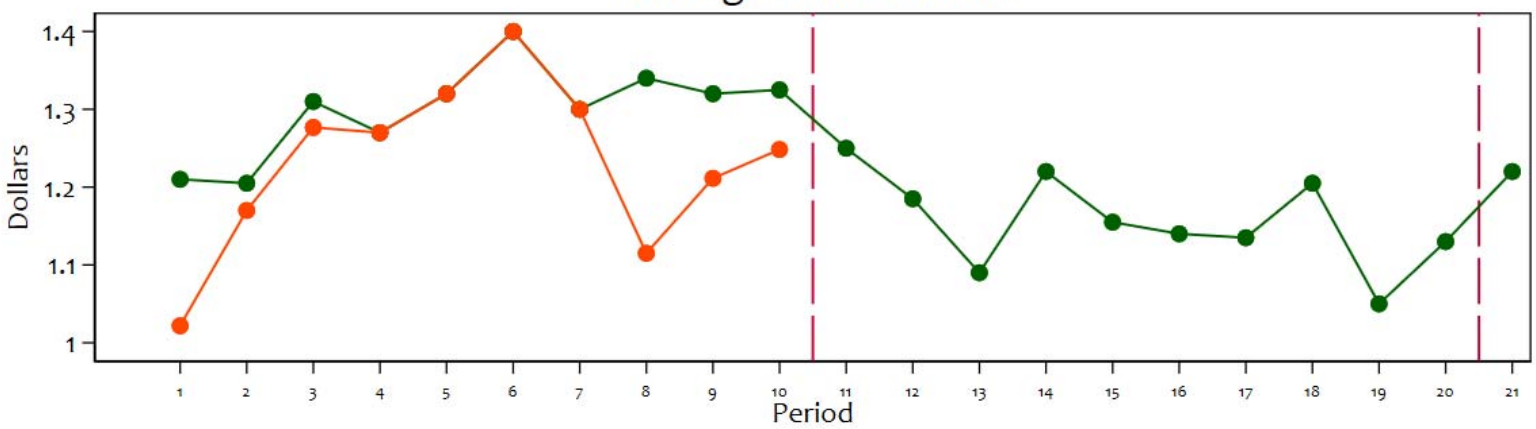
Results in Session 13 (N=6 Random Strangers)

High Return to Public Good

Average Token Contributions



Average Dollar Profits



Appendix B: Additional Literature Review (NOT FOR PUBLICATION)

Gürerk, Irlenbusch and Rockenbach [2005] study the evolution of different institutions with different sanctioning opportunities. They analyze the ability of these institutions to sustain cooperation and achieve high levels of efficiency, and whether these levels are higher under the endogenous institutional choice by each subject with respect to the case where the type of institution is decided exogenously by a central authority. They consider three types of institutions: the standard Voluntary Contribution Mechanism (VCM), the VCM with a possibility of punishing, and the VCM with reward opportunities.

The size of each institution is exactly the number of votes received in the endogenous institutional choice case. Conversely, under the exogenous institution case, each group consists of 6 members. The experiment lasts 30 periods, and each period consists of three stages under the endogenous choice and two stages under the exogenous case. In the first stage of the endogenous institutional choice, each subject has to vote between the standard VCM and VCM with punishing possibility (in this case the treatment is called FPN) or between the standard VCM and VCM with reward opportunities (in this case the treatment is called FRN), depending on the treatment in which the subject is randomly assigned.

Analyzing the institution choices, in the first round of FPN the majority of the subjects (69%) chose the standard VCM. But over time, subjects opted for the VCM with a possibility of punishing reaching a majority of 90% by the last round. On the other hand, under FRN, the majority of participants (76%) chose the VCM with reward opportunities, and this percentage remained quite stable throughout the experiment. Analyzing average contributions, contributions are higher when the punishment is allowed compared to the case in which the reward is allowed, under both the endogenous choice and the exogenous case. The contributions under endogenous punishment achieve higher levels than those in exogenous punishment, leading to the claim that endogenous choice was a crucial driver for evolution of high cooperation. Conversely, the contributions in the endogenous reward case are lower than the contributions in the exogenous reward case. Looking at the

punishment activities, it can be observed that they diminish over time because of higher levels of contribution. In case of reward opportunities, a diminishing path is shown because of lower level of contributions. Analyzing the efficiency levels, initially they are lower in case of punishment with respect to the standard VCM in both matching schemes (endogenous and exogenous). But as the experiment proceeds, the efficiency of VCM with endogenous punishment opportunities reaches the highest level among all the other treatments.

Gürerk, Irlenbusch and Rockenbach [2006] study how individuals behaved in a particular institution chosen endogenously by them. The subjects could choose between a sanction-free institution and an institution with the possibility of punishing and/or rewarding. The experiment lasts 30 periods and each subject interacts with the other subjects who have chosen the same institution. Analyzing the initial institutional choices, there is a preference for the sanction-free institution. Analyzing the contributions levels in the first periods, those made in the institution with the possibility of punishing and/or reward are higher than those made under the sanction-free institution. Specifically, half of the subjects are high contributors in the first period in the former institution; conversely, in the latter institution, almost half of the players are free-riders. Analyzing the contribution levels when subjects switch from an institution to another, a large majority of subjects increase their contributions when the sanctioning opportunities are introduced; conversely, when those opportunities are removed, the contributions are reduced by a large majority of the subjects. Efficiency approaches zero under the sanction-free institution; conversely when there are sanctioning opportunities, efficiency almost reaches 100%.

Gürerk, Irlenbusch and Rockenbach [2009] investigate endogenous institutional choices and their effect on cooperation among group members. They design three types of institutions: a non-sanctioning community, a punishment community, and a reward community (ReC). Conditional on which type of treatments they are randomly assigned, subjects have to choose between the non-sanctioning community and punishment community, in one case. In the other case, the community choice set is composed of the non-sanctioning community and the reward community.

The Partners matching protocol was used, and the experiment lasts 30 periods. Once each subject has chosen the community, she interacts with other subjects who have opted for the same community. Analyzing community choices, a majority of players prefer the non-sanctioning community over the punishment community; on the other hand, a majority of subjects prefer the reward community over the non-sanctioning community. Contribution levels are higher in sanctioning communities (punishment and reward communities) than in the non-sanctioning community at the beginning of the experiment. Over time, contributions in the punishment community increase, while those in the reward community decrease. Hence, punishment activities are more effective in reaching higher levels of cooperation with respect to the reward activities. Earnings are lower in the punishment community compared to the non-sanctioning community in the first half of the experiment, but over time steadily increase towards the social optimum. Therefore, the opportunities for punishment lead to a higher level of efficiency in the long run. On the other hand, earnings in the reward community decrease over time. Analyzing the sanctioning activities, high contributors heavily punish low contributors in order to increase their levels of contribution or to encourage them to leave the punishment community; high contributors also tend to reward other high contributors. Looking at the contribution levels when subjects change their communities, a large majority of subjects increase their contributions when punishment opportunities are introduced; however, when punishment opportunities are removed a majority of subjects decrease their contribution levels.

Kosfeld, Okada and Riedl [2009] investigate behavior in an institution formation setting. They consider two levels of marginal return to the public good and the possibility or not of institution formation. They implement the Partners matching protocol and each experiment lasts 30 periods. One feature of this experiment is that it relied on the subject's duty to contribute the full endowment if she chooses to participate in the organization. Specifically, each subject has to first decide if she wants to belong to a specific organization. If so, she will be treated as a participant, otherwise as a nonparticipant. Each participant then has to vote over the implementation of the organization, and the outcome of this vote requires unanimity. If the organization is implemented, then the participants

become members; otherwise they are involved in a standard public good game. In case of implementation, each member has to contribute her full endowment to the public good and in that case she cannot be punished. Conversely, if the member contributes an amount different from the entire endowment, she will be sanctioned depending on her actual level of contribution. All the other players (nonparticipants and nonmembers) can decide on any amount of tokens they want to allocate to the public good.

Analyzing results from the first and the second stages, when the marginal return is *low* and there is possibility of institution formation, there is at least one subject per group who is willing to implement an organization and this organization is actually implemented in almost the majority of the cases. When the marginal return is *high*, in almost all cases there is at least one subject who wants to implement the organization and that organization is actually implemented most of the time. Looking at implementation rates, the probability of an organization of being implemented is high when all the subjects in a group decide to participate in institution formation, independently of the level of marginal return to the public good. Examining contribution levels under low marginal return, when there is the possibility of institution formation compared to the case in which there is not this possibility, the contributions of the latter decrease over time. Therefore the possibility of institution formation leads to an increase in contributions when the marginal return is low. Conversely, when the marginal return is high, there is no difference in terms of contributions. Average efficiency reached under the possibility of institution formation is higher than that when this is not possibile, at least when the marginal return is low. A high marginal return leads to higher levels of efficiency, independently of whether or not there is a possibility of institution formation.

Ertan, Page and Putterman [2009] investigate the effects on contributions, earnings and efficiency of the possibility of endogenous choices rules of punishment. These rules provide an opportunity to vote on the possibility of prohibiting the punishment of high contributors. The Partners matching protocol and they design two experiments of 30 periods each. The difference between the two experiments relies on the number of times the participants are asked to vote. In the

so called "3 – Vote" experiment the subjects vote 3 times on the punishment rules; in the so called "5 – Vote" experiment they vote 5 times. In the latter case subjects do not experience the non-punishment situation. In both cases each subject had to vote to allow a reduction of individual's payoff indicating "yes," "no," or "no preference." She has to vote over 3 cases: if that individual contributes less than the average, if she contributes exactly the average, or if she contributes more than the average.

In the first 3 periods of the "3 – Vote" design subjects are not allowed to punish; from period 4 to period 6 unrestricted punishment is allowed; and at the beginning of period 7 subjects are asked to vote on who can be punished in the next 8 rounds. At the beginning of round 15, and at the beginning of round 23, subjects are asked again for a second and a third vote respectively. In the "5 – Vote" design subjects are asked to vote on who can be punished at the beginning of the 1st, 7th, 13th, 19th and 25th rounds. Hence each chosen rule lasted 6 periods in this design.

In both experiments none of the groups allowed punishment of higher-than-average contributors. A a large number of groups voted to prohibit all types of punishment (low contributors, high contributors and the medium contributors) in their first vote, and just a minority of groups voted to allow punishment of low-but-not-high contributors (combination of low contributors and medium contributors). However, this behavior is reversed as the periods go on. Indeed, in the last vote the majority of groups voted to punish the low-but-not-high contributors, and only a minority voted to prohibit all punishment. Looking at contribution levels, the groups that allow punishment of low-but-not-high contributors achieve a higher levels of contributions compared to to the groups that prohibit all punishment. The contributions associated with groups that allow punishing the low-but-not-high contributors increase over time. Efficiency reached by the groups that allow punishing the low-but-not-high contributors is higher than that achieved by the groups that prohibit all punishment, and in both experiments. In the "3 – Vote" experiment the groups with no punishment have the highest frequency of free riding, followed by the groups with unrestricted punishment, and in the end by the groups that allow punishment of the low-but-not-high contributors. The subjects who contribute

more than the average contribution are more likely to vote to allow punishment of low-but-not-high contributors. Hence, introducing the possibility of endogenous choices rules of punishment mitigates the free rider problem.

Gürerk, Irlenbusch and Rockenbach [2009] examine the behavior of a team leader who can choose a particular incentive scheme. They analyze how a specific incentive scheme affects team performance. The role of the leader is assigned randomly to one of the group members. They implement the Partners matching protocol over 30 periods, with 3 three phases of 10 periods each. At the beginning of each phase the team leader chooses between the reward institution and the punishment institution. Once she has voted, the chosen institution is implemented throughout the phase. Within the group, only the team leader can choose how many tokens she wants to assign to her teammates given the specific incentive scheme chosen.

Analyzing the incentive scheme choices, almost all leaders choose the positive incentive scheme during the first phase. In the second phase a smaller number of leaders opt for the positive incentive scheme, but it is still chosen by the majority. In the third phase, however, the majority of leaders opt for the negative incentive scheme. Overall contributions are higher in the punishment institution than in the reward institution. Contributions made under the punishment institutions are still higher than those achieved under the reward institution, even if the contributions of leaders and the contributions of teammates are considered separately. The relationship between the contributions of leaders and the contributions of teammates remains the same when leaders switch from the reward institution to the punishment institution. On the other hand, when leaders move from the punishment institution to the reward institution, the contributions of both leaders and teammates in the latter case are higher. Teammates tend to imitate their leader's behavior with a time lag of one period. And teammates who contribute as much as leader are heavily rewarded by her under the reward incentive scheme, and are virtually never sanctioned under the punishment scheme. Earnings are higher in the reward institution, but over time the difference diminishes and at the end the punishment institution leads to higher payoffs. The leaders' earnings are higher than teammates' earnings in both incentive

schemes, but te difference is less pronounced in the reward scheme. The leaders' earnings are higher in the punishment institution.

Gürerk [2010] studies if the lack of information about high cooperation levels in a public game with punishment opportunities is one of the reasons behind reluctance to join that community. He implements an experiment that is a replication of Gürerk et al. [2009], but with the provision of experience-based information. He provides a social history, with the main results of the previous experiment, to the subjects. Analyzing institutional choices when the additional information is provided, a slight majority of the subjects opt for the punishment community at the beginning of the experiment. Hence social history diminishes the reluctance of the community to punish. Higher levels of cooperation are reached in the punishment community with additional information, starting in early rounds. Therefore the punishment activity is less strong under the provision of social history compared to the case of no social history information. Earnings of punishment community are lower than those in the non-sanctioning community at the outset, independent of the provision of information. Over time, subjects' payoffs increase in the punishment community: and those with the social history information experience an immediate increase in payoffs, eventually reaching and exceeding the payoffs in the non-sanctioning community.

Sutter, Haigner and Kocher [2010] examine cooperation in social dilemma situations with the possibility given to subjects to decide endogenously which type of institutions they want to participate in: a standard VCM, a VCM with a punishing possibility, and a VCM with a reward opportunities. They analyze how the choice of a particular institution affects the behavior and interaction among group members. They compare behavior with endogenous institutional choice compared to exogenous imposition of a specific institution. Two levels of "leverage" (high and low) capture the effectiveness of punishment or reward opportunities: how large the impact is on the monetary payoff. A Partners matching protocol is used, and the experiment lasts 10 periods. When institutions are exogenously imposed, contributions are lowest in the standard VCM, and there is not a large difference compared to VCM with punishment and reward given the low level of leverage. In the

high leverage case contributions under punishment and reward are higher than those achieved in the standard VCM. Therefore, when the effectiveness of sanctioning opportunities is relatively high, the introduction of these opportunities leads to an increase in contributions. In case of endogenous institutional choice, the percentage of subjects who want to participate in a costly vote to determine in which type of institutions they want to belong, is higher in the high leverage scenario compared to low leverage. When the effectiveness of sanctioning possibility is relatively high, the majority chooses the VCM with a possibility of rewarding; on the other hand, when the effectiveness is relatively low the majority chooses the standard VCM. When the effectiveness is relatively low, contributions are lowest in the standard VCM, followed by the VCM with reward opportunities, and highest level in the VCM with punishment possibilities. Conversely, when the effectiveness is relatively high, punishment is not chosen at all, and the highest contributions are reached under the VCM with a reward possibility. Earnings are always higher under sanctioning opportunities than in the standard VCM. Contributions when the institutional choice is endogenously determined in comparison to when it is exogenously imposed, contributions are always greater in the former.

Putterman, Tyran and Kamei [2011] investigate if groups are able to choose efficient formal sanction schemes among several options by voting without any external guidance. Eight political preference questions are taken from the World Values Survey, and subjects take a short intelligence test in order to analyze how political preferences and cognitive ability might affect the choice of efficient sanction scheme. There are two experimental conditions: a baseline and a penalty treatment. A Partners matching protocol is used, and the experiment lasts 24 periods, divided into six blocks of four periods. In the baseline treatment the blocks are identical. However, in the penalty treatment, at the end of each block subjects have to vote over penalization of private contributors or public contributors, a maximal penalty level, and a penalty exemption level. The outcomes of the vote are then implemented for the next four periods. The first block of the penalty treatment replicates the baseline treatment.

Average contributions levels decrease over time in the baseline treatment; on the other hand,

there is an increase in contributions when the possibility to vote over penalty schemes is introduced. There is considerable support for penalizing contributors to private accounts. A near-majority of votes are in favor of both the maximal penalty level and setting the penalty exemption level equal to the value of the initial endowment. Earnings are higher in the penalty treatment compared to the levels reached in the baseline. The preference to penalize low contributors to the public good is positively correlated with a high level of IQ and cooperation orientation, but negatively correlated with political conservatism and being female.

DeAngelo and Charness [2012] investigate the effect of different deterrence mechanisms in order to find the most effective mechanism, and analyze the effect of uncertainty over the regime implemented on individual behavior. The proscribed activity considered is "speeding." Speeding can be defined as a public good (or bad, since it is a violation of laws) because it is very difficult to completely prohibit, and if a subject exceeds speed limits that does not preclude another subject exceeding speed limits as well. Each subject has to decide if to speed or not speed under different conditions. Two experiments are designed with 30 periods in each experiment, each experiments consisting of three phases. In each phase there are two alternatives regimes with different probabilities of being caught, and different fines such that the expected fine is the same over the two regimes.

In the first phase of the first experiment, expected earnings associated with the decision to speed are higher than those associated with the decision to not speed. In this phase subjects have a 50% probability of being randomly assigned to one of the two regimes. In the second phase subjects vote over the regime they prefer in each round. The regime with a majority of the votes is implemented and then each subject decides to speed or not. In the third phase the probabilities and the fines associated with the two regimes change such that the expected earnings are lower under the decision to speed. Once each subject votes for one of these two regimes in each period, the regime with the majority of votes is implemented and subjects decide to speed or not.

The first and the third phases of the second experiment are a replication of the first and second phases in the first experiment, respectively. In the second phase the regime is imposed and

changes period-by-period.

Speeding rates are higher when the subjects vote over their preferred regime than when there is 50% of chance to be assigned to one of the regimes. Speeding rates are clearly higher when the preferred regime is implemented according to a majority rule. Speeding rates are also higher when the cost if a subject is caught is low, indicating that individuals are sensitive to expected costs. Subjects perceive larger expected cost under the regime where the probability to be caught is higher. An implication is that policymakers have to implement the regime that is the *opposite* of the regime preferred by subjects in order to make the decision to speed unattractive.

Markussen, Putterman and Tyran [2013] study a collective action dilemma and compare the performance of formal and informal sanctioning, as well as the option to join a sanction-free scenario. They analyze how these different environments affect behavior, and if there exists some difference between endogenous institutional choice and exogenous imposition of a particular institution. Three types of regimes are considered: no sanction, formal sanction and informal sanction. A Partners matching protocol is used, and the experiment lasts 28 periods, divided into seven phases of four periods each.

In the first phase the NS regime was imposed exogenously. At the beginning of the other six phases, each subject had to vote on the regime preferred over two alternatives. In phase 2 each subject had to decide between no-sanction and informal sanction regimes, in phase 3 between no-sanction and formal sanction regimes, in phase 4 between informal sanction and formal sanction regimes, and from phase 5 to phase 7 this cycle is repeated. Under the formal sanction regime each subject is penalized at a rate of 0.4 or 0.8 (these regimes are called "non-deterrent" or "deterrent," respectively) for each token allocated in the private account. Each subject incurs in a cost of 2 or 8 (these regime are called "cheap" or "expensive," respectively) to implement the formal sanction regime.

Analyzing the voting outcomes in the endogenous treatments, the majority of subjects prefer the no-sanctions regime compared to the informal sanction regime when voting for the first time in phase 2. However, when the subjects have to vote again over this choice set, in phase 5, the majority

prefers the informal sanction regime. Almost 75% of the groups choose the formal sanction regime rather than the no-sanction regime in the deterrent (penalization rate of 0.8) and cheap (implementation cost of 2) treatment, but only 25% of groups chooses the formal sanction regime over the no-sanctions regime in the deterrent and expensive (implementation cost of 8) treatment. This means that the cost of implementation has a significant impact on preferences and regime popularity. In terms of voting over the formal sanction and informal sanction regimes, nearly 75% of the groups prefer the informal regime over the formal regime except in the deterrent and cheap treatment, where the preference is reversed. In the no sanctions regime average contributions start at 50% of the endowment and then decline over time, but are always above zero. Contribution levels under the deterrent formal sanction regime are between 80% and 100% of the endowment. Contributions in the informal sanctions regime are similar to those achieved under deterrent sanctions. Hence informal punishment has a disciplining effect by inducing higher cooperation. Comparing results between the deterrent and cheap treatment and its exogenous correspondent, holding experience constant, average contributions and earnings are higher when the informal sanctions regime is chosen compared to when this type of regime is exogenously established. Therefore, endogenous choice leads to an increase in contributions in the informal sanctions regime. Subjects use the informal sanctions regime in a deterrent way, leading to an increase in terms of efficiency.

Drouvelis and Jamison [2015] study how institutions are selected and implemented and what characteristics of each subject predict the effect of the institutional choice on behavior. Four institutions are considered: the standard Voluntary Contribution Mechanism (VCM), the VCM with a possibility of punishing, the VCM with a reward opportunity, and the VCM with both punishment and reward possibilities. A Partners matching protocol is used. The experiment lasts 25 periods and consists of several parts. In the first part the levels of risk and ambiguity aversion are elicited, in each case over gains or losses. In the second part preferences over the four different public goods institutions are elicited, and then subjects are randomly assigned to one of the four institutions. After the elicitation of risk, loss and ambiguity aversion, subjects are asked to indicate their confidence about

subjects' earnings with respect to the maximum earnings achievable from the 25 periods. Subjects are asked also to indicate in which institution they prefer to participate by assigning a monetary amount, between -£5 and +£5, for each institution. If the subject is assigned to a preferred institution, according to willingness to pay, then that amount indicated is subtracted from the final payment. Conversely, if she is assigned to the institution where she has indicated that she wants to get paid, then the amount indicated is added to the final payment. Once decisions have been submitted, subjects are randomly assigned to one of the four institutions. At that point, subjects participate in a public good game.

There appears to be a positive, significant correlation among all the preference measures (risk, loss and ambiguity aversion) and individual characteristics (such as age and nationality). Subjects do not prefer to participate in institutions with punishment opportunities, and these institutional choices are not affected by individual preference measures. The two institutions with punishment opportunities show an increasing contributions path over time, while the other two institutions display a decreasing contributions path over time. The treatments with punishment lead to a higher level of efficiency in the long run due to higher level of cooperation. In terms of punishment and reward points assigned, free riders are heavily punished and high contributors are rewarded by other high contributors.