

Social Norms and Social Choice

by

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

February 2007

ABSTRACT

Experiments can provide rich information on behavior conditional on the institutional rules of the game being imposed by the experimenter. We consider what happens when the subjects are allowed to choose the institution through a simple social choice procedure. Our case study is a setting in which sanctions may or may not be allowed to encourage “righteous behavior.” Laboratory experiments show that some subjects in public goods environments employ costly sanctions against other subjects in order to enforce what appears to be a social norm of contribution. We show that such an artificial society would not be an attractive place to live in, by standard social choice criteria. If such societies came about because of evolutionary forces, as speculated in the literature, then we argue that The Blind Watchmaker was having one of his many bad days at the workbench. In fact, none of our laboratory societies with perfect strangers matching ever chose to live in such a world. Our findings suggest that the conditions under which a group or a society would choose a constitution that is based on voluntary costly sanctions are very special. More fundamentally, they demonstrate the importance of naturally endogenizing the choice of institution if one is to make reliable inferences about treatment effects from different institutions.

† Department of Economics, University of Minho and NIMA (Botelho and Pinto) and Department of Economics, College of Business, University of Central Florida (Harrison and Rutström). E-mail: botelho@eeg.uminho.pt, gharrison@research.bus.ucf.edu, pintol@eeg.uminho.pt, and erutstrom@bus.ucf.edu. 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 for helpful comments. All data, instructions, and statistical code is available at the *ExLab* Digital Library at <http://exlab.bus.ucf.edu>.

A popular thesis is that reciprocal behavior has arisen out of human evolution, whether it be biological or social evolution. Reciprocity in the form of punishments has been shown to be effective in creating and preserving cooperative norms in experimental 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 findings from these experiments have generated the conclusion that social processes that involve sanctioning are superior to those that do not involve sanctions, and that this is a result of evolutionary forces. We believe this conclusion to be premature. Although we understand that it is tempting to argue that we are “genetically coded” to support cooperation through sanctions, we argue that an evolutionarily valid argument cannot be based simply on the observed *use* of sanctions, but must be based on the demonstrated *fitness* of the behavior. We take the natural next step and extend these experiments to also investigate how the fitness of cooperative norms depends on the presence of punishment options.

It is important to recognize that fitness is not a necessary condition for some behavior to be present at a point in time during an evolutionary process. Fitness is necessary for behavior to survive over time, but has no impact on the emergence of behavior. Apart from the weeding out of relatively unfit behavior, evolutionary processes also embed an important element of experimentation, i.e. the random introduction of new behaviors. 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 conscious intention of producing the object. That process relies on some criteria of fitness and survival to be applied to weed out the many mistakes that random deviation tosses out as part of such a mindless process. In fact, it is crucial to this

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.”

evolutionary argument that there be lots of such mistakes. To make inferences regarding the survival probabilities of certain behaviors that are present in a population, it is therefore critical to identify what one means by fitness in terms of the underlying evolutionary process hypothesized, and to measure that.

In economic games one appropriate measure of fitness is simply the earnings of the individuals in the group.² Groups with behavior that generates low earnings are then hypothesized to disappear over time, either because individuals leave the group, or because the group is extinguished by other relatively fitter groups in the competition for resources. Since sanctions are costly to impose, generating and sustaining cooperation through sanctions may be a losing proposition for some groups. If we interpret the presence of sanctions as evidence of an evolutionary process in cases where the outcomes are relatively unfit, then we must be observing the blind watchmaker on one of his bad days.³ If sanctions lead to net earnings losses compared to processes without sanctions, the former cannot be part of an evolutionarily stable outcome in which earnings are the measure of fitness.

The selection of sanctions as a social norm can be undertaken through numerous social mechanisms. We study selections by the use of a majority vote mechanism, where participants in laboratory experiments vote over norms that differ in their use of sanctions. We use this mechanism to focus directly on revealed preferences for a sanctioning norm. We hypothesize that social choice will depend on how sanctions affect earnings, and that the propensity to vote for sanctions will be increasing in the earnings that they generate. We employ an incentive compatible voting mechanism such that we can directly elicit the revealed preferences for the individuals in the group as well as the

² Apart from parsimony and consistency with economic theory, this measure has the decided advantage of leading to hypotheses that are easy to make operationally meaningful. One could posit that agents also care about relative payoffs, the distribution of payoffs, or the payoffs to the group as a whole, but these hypotheses require elaborate experimental designs to make them operational (e.g., Rutström and Williams [2000] and Cox [2004]). We believe that such concerns are valid but deserve more careful study in designs that allow one to test them meaningfully.

³ Or else there may be some other latent metric of fitness we cannot (currently) observe.

social choice outcome for the group as a whole.

Many studies report *detrimental* effects on earnings from unrestricted sanctions (Carpenter and Matthews [2004], Egas and Riedl [2005], Page, Putterman and Unel [2005], Sefton, Shupp, and Walker [2005], Casari and Luini [2005], Nikiforakis [2006] and Anderson and Putterman [2006]). Additional support for the argument that earnings have a strong influence on the survival of sanctioning institutions is provided by Gürer, Irlenbusch, and Rockenbach [2006], although in their experiment the earnings criterion ends up *favoring* the sanctioning institution.

We report an experiment that extends the Fehr and Gächter (FG) [2000] design with an explicit public choice phase where preferences over the institutions can be observed directly, and the factors driving institutional choice examined. We find that in the absence of reputation effects *none of our laboratory societies chose to live in a world with sanctions*. This result is robust to the order in which participants experienced the alternatives 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 relax our re-encounter conditions to allow for some reputation effects, we find one case in which the majority votes for the world with sanctions: this case requires that the rewards to contributing be relatively high, and that the different regimes be experienced in a certain order. Our findings suggest that the conditions under which a group or a society would choose a constitution that is based on voluntary costly sanctions are very special.

In our experiments the driving force behind this reluctance to adopt the sanctions scheme appears to have been the reduction in profits that it caused. 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 profits are a key candidate for what motivated actions.⁴

⁴ Consider a world of sanctions in which a majority of subjects made more profit on average than they would in a non-sanctions world, but a minority of subjects earned virtually nothing. Average profit 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.

Beyond the question of social norms and social choice, our experiments identify an important methodological point about the use of inferences from experiments in the design of policy. If subjects in the field have mechanisms by which they can avoid, lobby or self-select into or out of institutions, we must consider the effects of those margins of choice before drawing conclusions about which institutions are best. Another way to express this is to consider if the laboratory environment that takes a particular institution as fixed is correctly modeling the naturally-occurring environment in its salient features, if that environment includes ways in which subjects can endogenously opt out of that institution.

1. The Value of a Social Norm Enforced by Punishment

Relying on an established literature that shows 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. More precisely, what is the net value of allowing the punishment technology that endogenously generates the cooperative norm?

In the experiment in 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 they 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 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 profitability of allowing sanctions, based on the data generated by FG [2000].⁵ 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 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.⁶ Thus, the value of the norm depends on the extent to which one is willing to extrapolate beyond the life of the experiment, and more importantly the extent to which the subjects are willing to extrapolate into the future.

Figure 2 reports comparable calculations for the experiments in FG [2002]. 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 the norm. In each Series the aggregate loss in value from the norm 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 would generate a positive net value.

These experiments show that sanctions can have a strong effect on cooperation. What is less

⁵ 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.

⁶ For example, average profits in the Second Series were 22.73 currency units and 18.85 currency units, respectively without and with the punishment norm, for a difference of 3.88 currency units or 17.1%.

clear, however, is the extent to which the earnings effect of the sanctions are perceived as favorable by the participants. More to the point, would the participants in any of these experiments want to have this technology 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 rules⁷ 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 social technology as a “constitutional” matter?

One particularly nice feature of the FG [2000][2002] design is that it allows in-sample comparisons of the value of the norm 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 norm. 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 social norm would not be adopted under either of these social choice criteria*.⁸ Furthermore, it is much harder to argue *a priori* that simple extrapolation beyond the life of the experiment provides any basis for predicting that the norm would be socially acceptable under these criteria. Our experiment was designed to investigate the question of social choice directly, by allowing participants to vote over social norms 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 or not.

⁷ 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 norm 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).

⁸ 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.

2. Voting for a Social Norm Enforced by Punishment

A. Basic Experimental Design

We design a simple experiment to test whether subjects would choose to live in a world with 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 norm and also in games in which there are no such norms. 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.⁹ Although this does reduce the chance that the subject will 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.¹⁰ 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

⁹ 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 of regimes (Punishment followed by Non-punishment), and did not maintain the Perfect Strangers treatment after the first regime of 6 periods. Rather than debate if such comparisons are conclusive, we prefer to ensure the control against any reputational effects afforded by a Perfect Strangers design.

¹⁰ Most public goods experiments use four subjects per group, although the effect of larger group sizes has been studied by Isaac and Walker [1988] and others. Harrison and Hirshleifer [1989] and Goeree, Holt and Laury [2002] employed groups of 2 in their public goods experiments.

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.¹¹

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

We are now ready for your final task. This will consist of only one period. The task will be a repetition of one of the two tasks you have just completed. Which task this will be will be determined by a common vote in a moment. In this one period the stakes will be increased so that each token is now worth 50 cents, not just 5 cents. This is therefore 10 times the value that a token has had in each of the earlier periods.

Before you play out this one period, you will be asked which environment you would like to participate in. You may choose either the one where you can reduce other participants' earnings and they can reduce yours (**environment B**) or you may choose the environment in which there is no such opportunity (**environment A**). Everyone will be asked to vote for the environment that they prefer, and **we will implement the environment that a majority of the participants in this room vote for.** Thus, we will implement the same environment on all matched pairs.

In the event of a tied vote we will roll a ten-sided die for you all to see. If the die comes up 0-4 we will implement environment A, where earnings reductions are not available, if it comes up 5-9 we will implement environment B, where earnings reductions are available.

Before you are asked to vote you will be shown a screen with a review of your earnings across the periods in both of the environments.

One variation of these instructions simply reverses the references to environment B and

¹¹ 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.

¹² 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.

environment A since the order of the two in the first part of the experiment was reversed. 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 themselves. 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 anything, such as a punishment norm, 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.¹³

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).

C. Procedures

We recruited 180 subjects from the University of Central Florida (UCF) in 2005.¹⁴ 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 [1999].¹⁵ The same experimenter

¹³ 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].

¹⁴ 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 diverse: 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.

¹⁵ 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>.

(Rutström) delivered the oral instructions for all sessions, to ensure comparability.¹⁶ 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 punishment regime. Overall only 18% of participants in the Perfect Strangers treatment voted for the punishment regime. 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 to implement the no-punishment regime. 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

¹⁶ A digital recording of the oral instructions in one typical session is available at the ExLab archive.

positive effect on contributions, much less on profits.

As one of the candidates for explaining the variation in the effect of punishment we find a significant difference in public goods contributions under Perfect Strangers and Random Strangers conditions, and this difference is also reflected in differences in voting behavior. One shot games are ideally modeled using a Perfect Strangers matching protocol since there is no repeated interaction, not even a probabilistic one. Since behavior in Random Strangers is significantly different, we conclude that it would be inappropriate to assume that subjects treat Random Strangers designs as if they were one-shot experiments. Botelho, Harrison, Pinto and Rutström [2005] discuss this methodological issue further.

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.

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 regime 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 regime might be of value, since nobody needs a punishment regime if everyone is contributing heavily. And, of course, we test for such order effects from the sequencing of the two regimes. One could also argue that the vote was stacked against the punishment regime 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 appendix displays similarly detailed outcomes for each of the other sessions.

Figures 4 through 9 show the average “take home profits” in each period and session, along with the percentage vote for the punishment regime.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 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 punishment regime. 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

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.

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 and the evolution of a social norm. 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 option. 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 punishment regime 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 regime 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 punishment regime. 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 regime in session 2 compared to the no-punishment regime, maybe they were higher for those 21% that voted for the punishment regime.

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 punishment regime.

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

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.

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 kept 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 destroy any evolving social norm. 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 punishment regime, 60% of the subjects presumably viewed that as an outlier from the promise of things to come if they had another, final period of the punishment regime. 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 social norm of cooperation under the threat of punishment to break down, but it did not.

Sessions 12 and 13 exhibited roughly the same average profit in each regime, but the variation in profitability in the punishment regime 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 punishment regime 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 punishment regime in the voting stage.

In summary, we only find one session in which there is majority support for the punishment regime. 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 punishment regimes.

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 during periods 1-10, whether the subject received a high 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 3¹/₄, cumulative GP above 3³/₄, 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.

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 were selected.

We are concerned *a priori* that two of the explanatory variables of particular interest might be endogenous to the vote. These variables 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 [2005; p.523ff.]. 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,²⁰ and we report estimates under the assumption that it is endogenous. Since we 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 earnings in the NP versus P treatments that determines the vote. Subjects that experienced higher profit in NP compared to P were 63 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 one comes to the same qualitative conclusion.

²⁰ In an independent test of exogeneity of own profits, we *can* reject the null hypothesis of exogeneity (p -value of 0.031). Similarly, in an independent test of other-player contributions, we cannot reject the null hypothesis of exogeneity (p -value of 0.32).

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 high rewards encourages outcomes in which participants are significantly less likely to vote for the NP environment, since the return to encouraging cooperation by the *efficient use of punishment* is higher. The effect of this treatment is to lower 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 earnings that subjects have experienced in them, and not by the extent to which a cooperative norm is upheld. 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 to constitutional choices.

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. Overall, average contributions to the public 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 Gächter [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 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 are very encouraging for our hypothesis that choices over constitutions with or without punishment mechanisms will depend on earnings. They would also suggest that subjects should vote in favor of punishment constitutions, which is not confirmed in our data. Our data instead suggest that the conditions under which such constitutions are preferred are very special, depending on the exact experiences that participants have.

Ertan, Page and Putterman [2005] and Sutter, Haigner and Kocher [2006] 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 [2005] 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.

Sutter, Haigner and Kocher [2006] study if 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.

4. Conclusions

Our experiments address the question of fitness, as measured by earnings, in determining the attractiveness of alternative constitutional choices and the potential for survival in evolutionary processes. In addition, our focus on endogenous institutions identifies an important methodological point about the value of inferences from experiments with exogenous institutions. In naturally occurring environments participants have mechanisms by which they can avoid, lobby or self-select into or out of institutions. Although related to the sample selection and sample attrition problems, the issue can be better framed by asking if the experimental control of imposing an exogenous institution removes the very margin of choice that it is supposed to help explain. We illustrate the natural methodological extension to experiments in which subjects can explicitly choose the environment in which they are to participate. Our case study is a setting in which costly informal sanctions may or may not be allowed to solve social dilemmas.

In addition to these methodological implications, our findings also have substantive import.

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 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*²¹ and the principles that apply to actions *guided by institutions*. While 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.

²¹ That is, the set of rules, including the norms of social groups, combined with their enforcement mechanisms, that constrain the choices of individuals.

Figure 1: Value of the Social Norm in AER Experiments

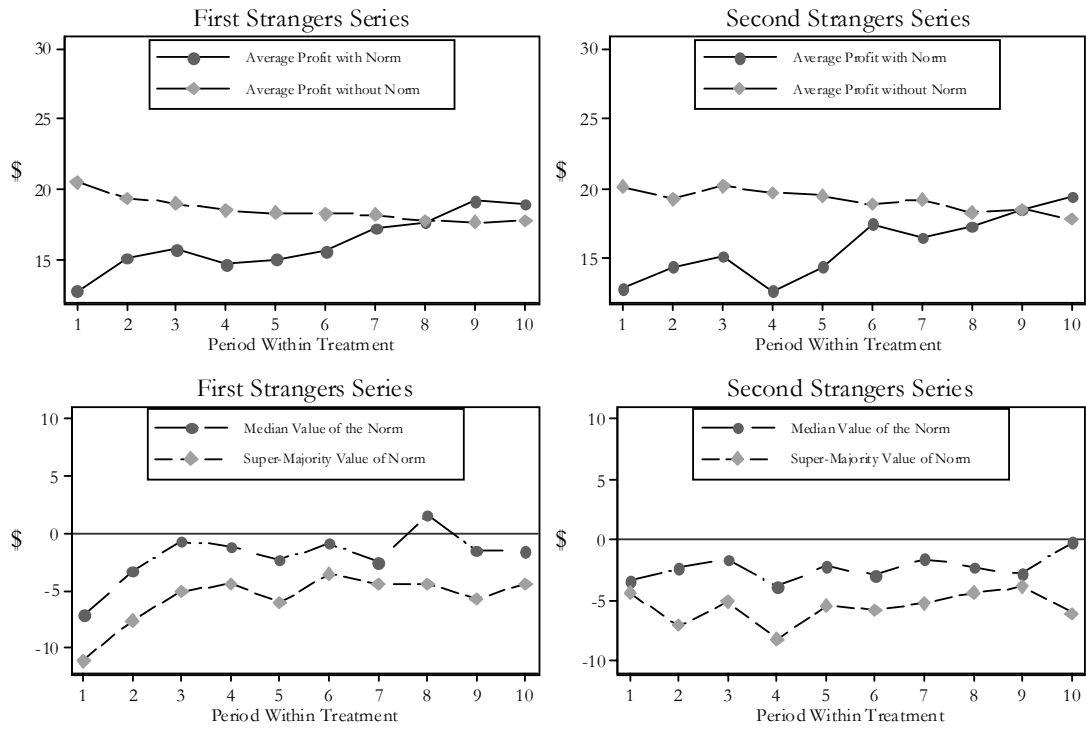


Figure 2: Value of the Social Norm in Nature Experiments

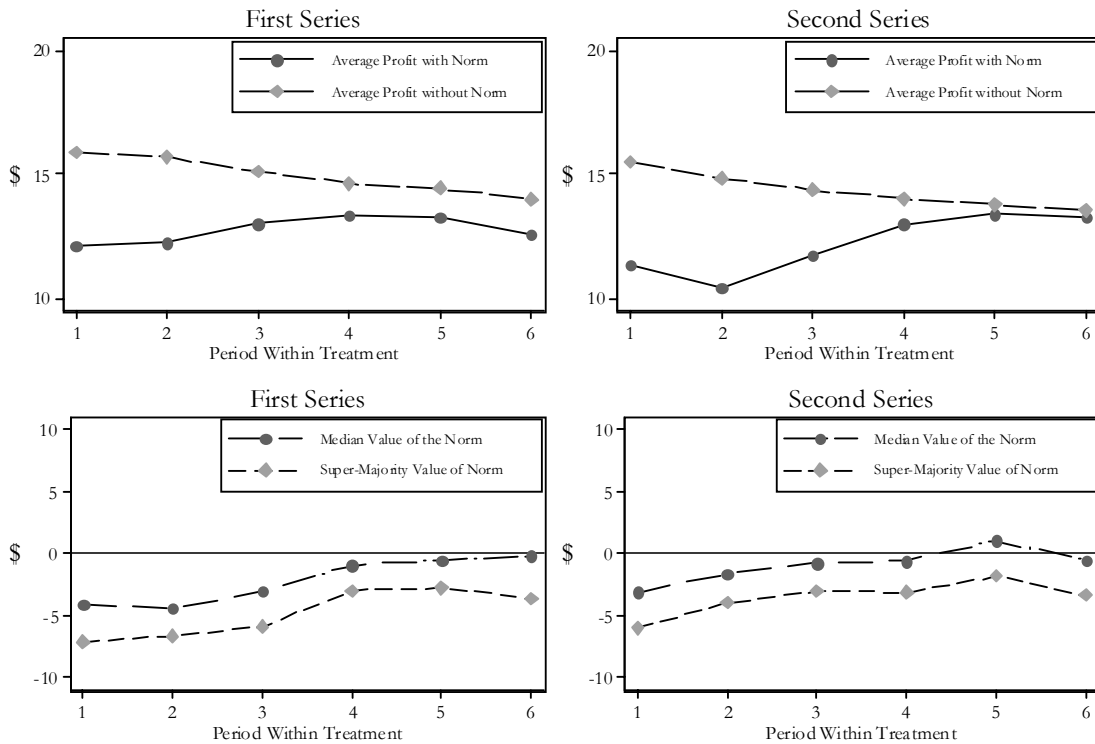


Table 1: Experimental Design

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

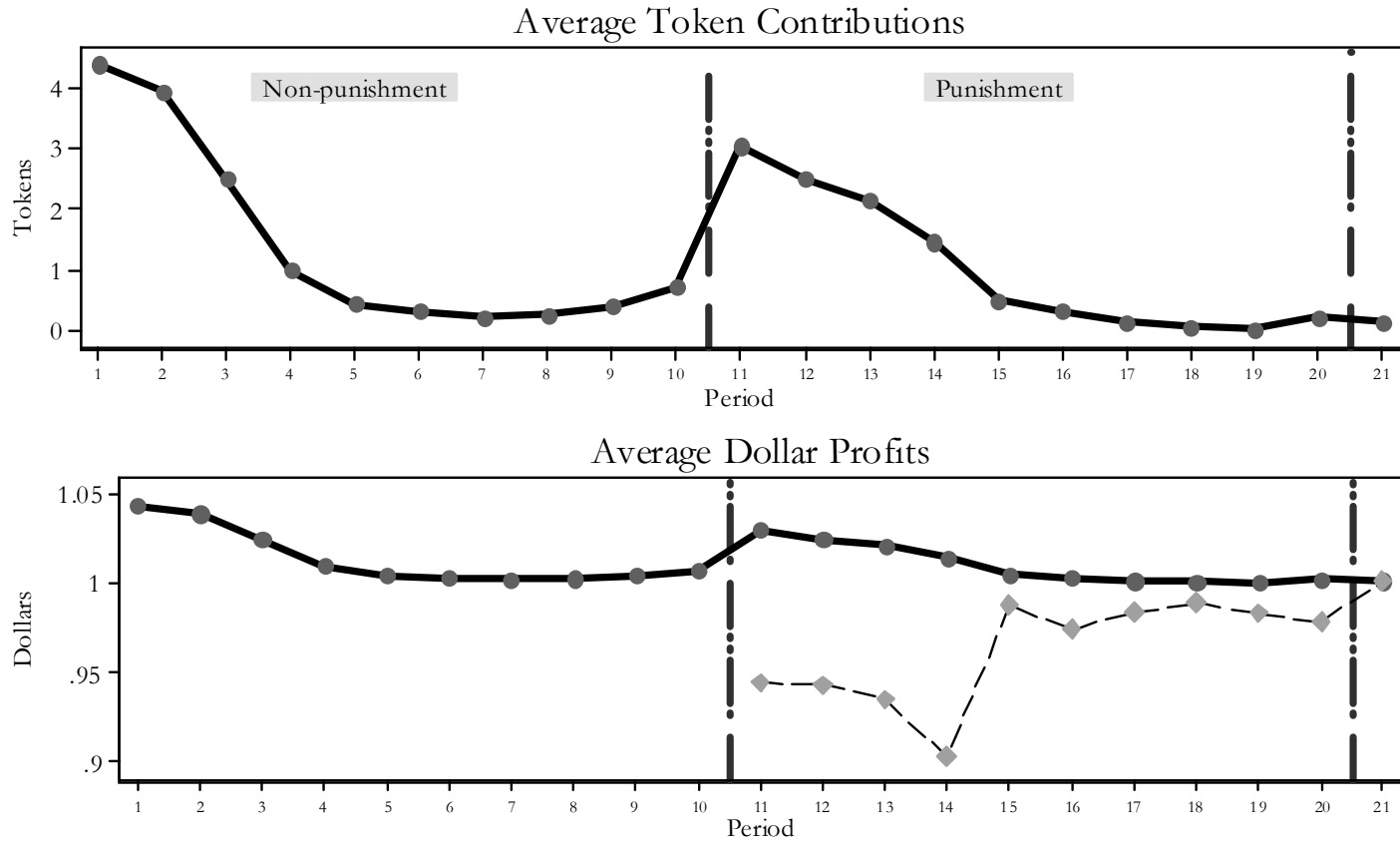


Figure 4: Average Profits With Perfect Strangers and NP-P History

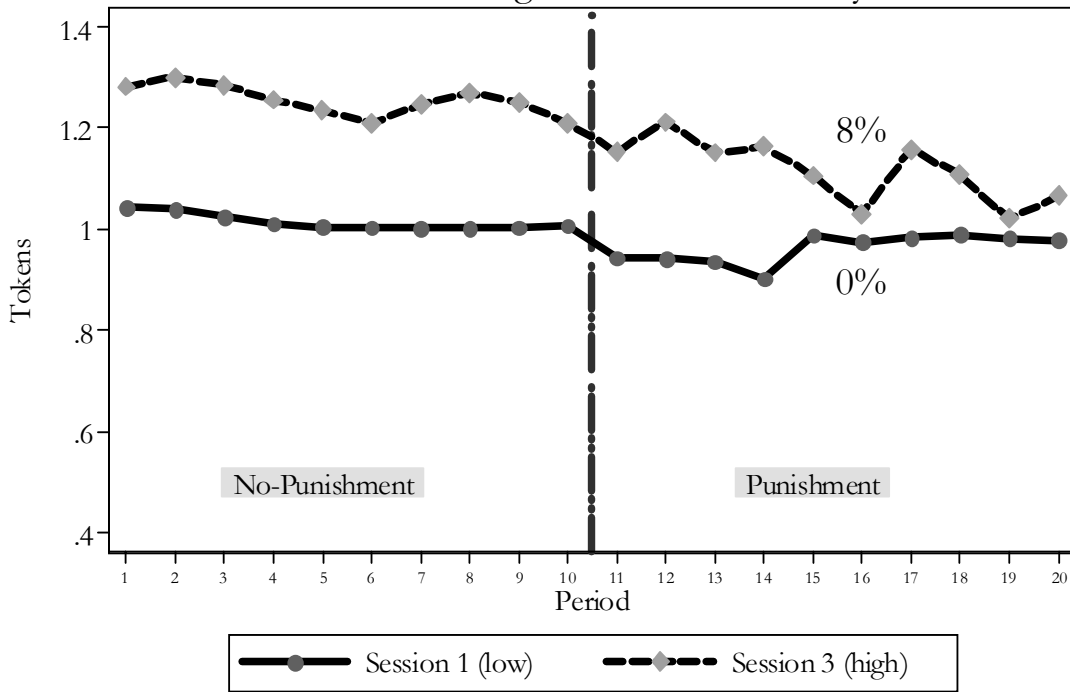


Figure 5: Average Profits With Perfect Strangers and P-NP History

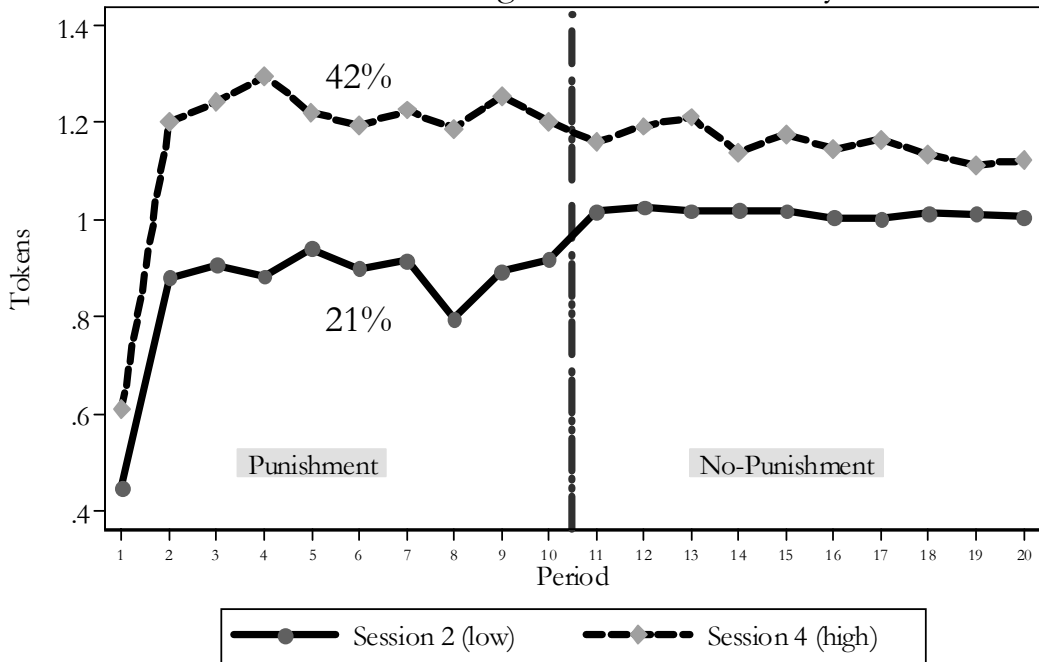


Figure 6: Average Profits With Random Strangers and P-NP History

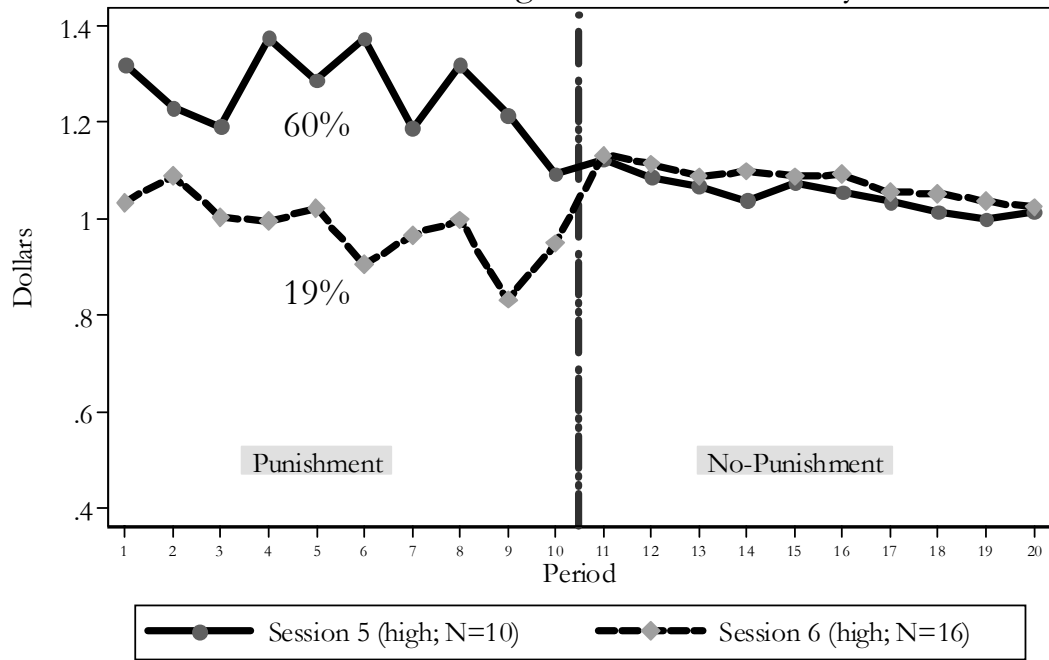


Figure 7: Average Profits With Random Strangers and P-NP History

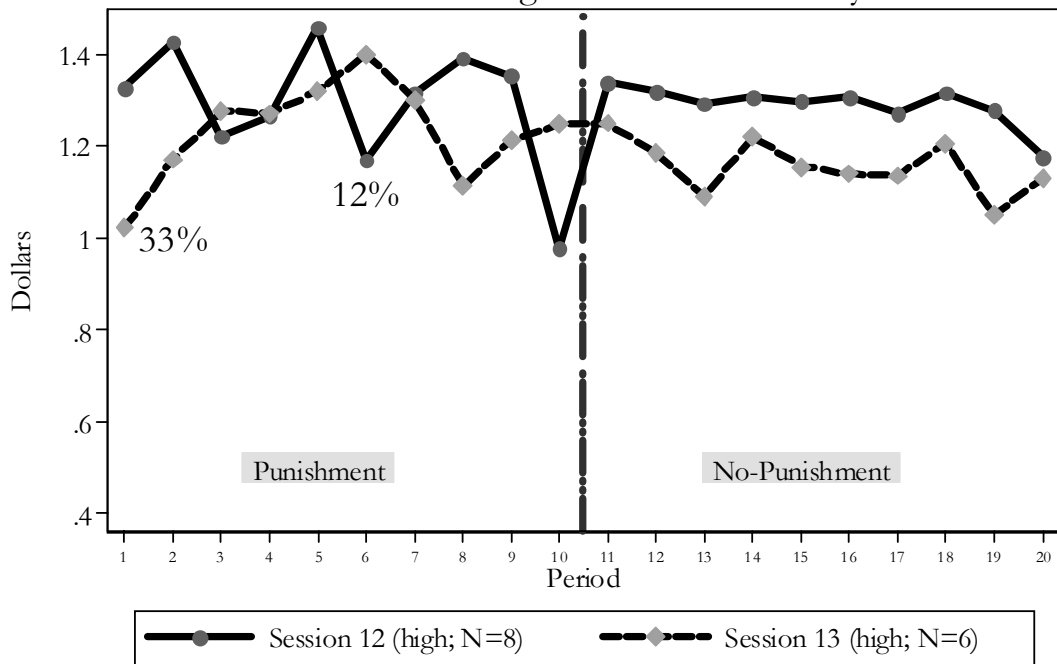


Figure 8: Average Profits With Random Strangers and NP-P History

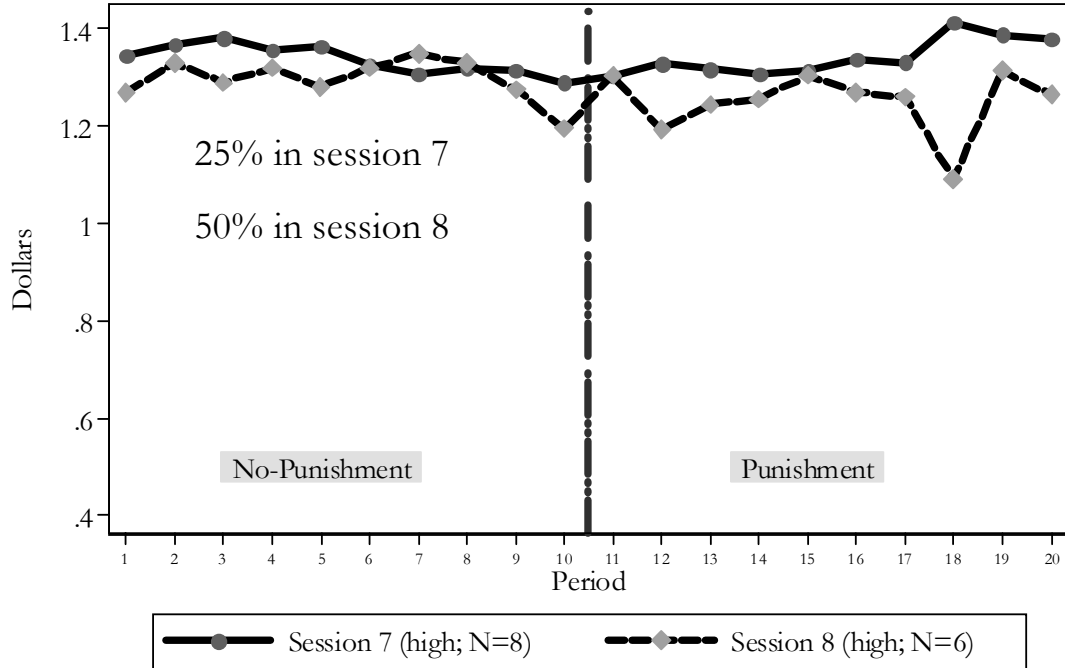


Figure 9: Average Profits With Random Strangers and NP-P History

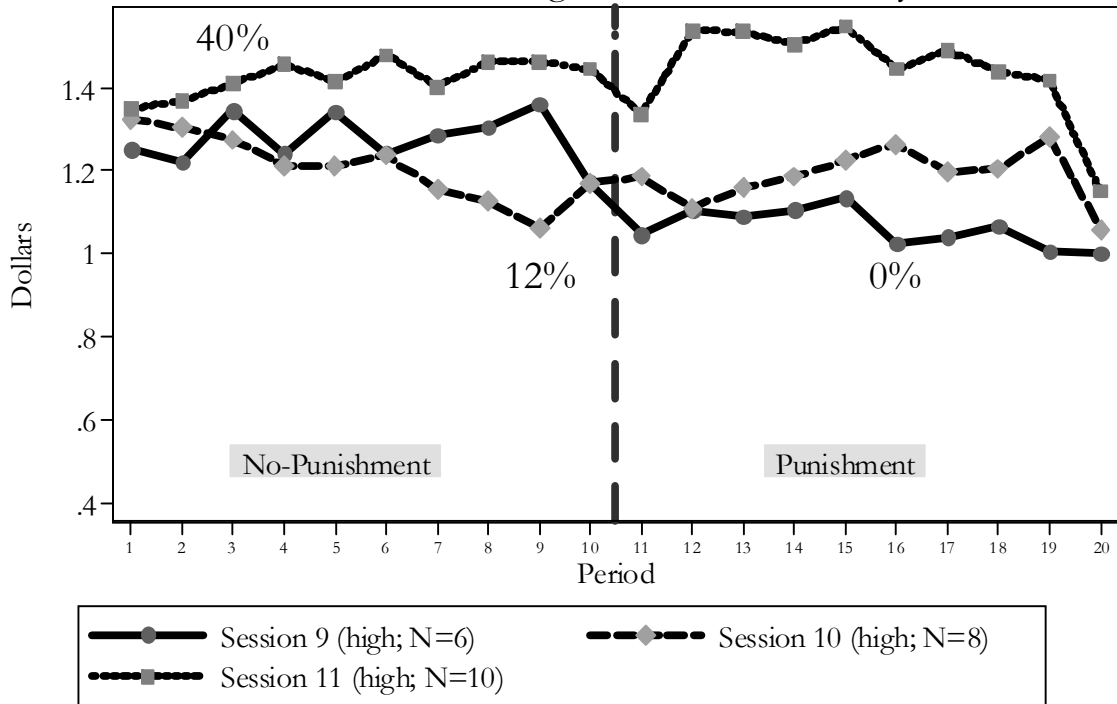


Table 3: Descriptive Statistics for Variables in Voting Model

Variable	Mean		Description
	Full Sample	High Returns Sample	
	(Standard Deviation)		
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
high	-0.153	0.059	0.010	-0.270	-0.037
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).
- Botelho, Anabela; Harrison, Glenn W.; Pinto, Lıgia M. Costa and Rutstrom, Elisabet E., "Testing Static Game Theory with Dynamic Experiments: A Case Study of Public Goods," *Working Paper 05-25*, Department of Economics, College of Business Administration, University of Central Florida, 2005.
- 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.
- Cox, James C., "How To Identify Trust and Reciprocity," *Games and Economic Behavior*, 46(2), 2004, 260-281.
- Dawkins, Richard, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design* (New York: Norton, 1986).
- Egas, Martijn, and Riedl, Arno, "The Economics of Altruistic Punishment and the Demise of Cooperation," *Discussion Papers 05-065/1*, Tinbergen Institute, the Netherlands, 2005.
- 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, "Can Endogenously Chosen Institutions Mitigate the Free-Rider Problem and Reduce Perverse Punishment?" *Working Paper 2005-13*, Department of Economics, Brown University, 2005.
- Falk, Armin; Fehr, Ernst, and Fischbacher, Urs, "Driving Forces Behind Informal Sanctions," *Econometrica*, 73(6), 2005, 2017-2030.
- Fehr, Ernst, and Gachter, Simon, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), September 2000, 980-994.
- Fehr, Ernst, and Gachter, Simon, "Altruistic Punishment in Humans," *Nature*, 415, 10 January 2002, 137-140.

- Fischbacher, Urs, “z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter’s Manual,” *Working Paper Nr. 21*, Institute for Empirical Research in Economics, University of Zurich, 1999.
- 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, “The Competitive Advantage of Sanctioning Institutions,” *Science*, 312, April 7, 2006, 108-111.
- 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.
- Isaac, R. Mark and Walker, James M., “Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism,” *Quarterly Journal of Economics*, 53, 1988, 179-200.
- Nikiforakis, Nikos, “Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?” *Unpublished Manuscript*, Department of Economics, Royal Holloway, University of London, February 2006.
- 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.
- Rutström, E. Elisabet, and Williams, Melonie B., “Entitlements and Fairness: An Experimental Study of Distributive Preferences,” *Journal of Economic Behavior and Organization*, 43, 2000, 75-80.
- Sefton, Martin; Shupp, Robert S., and Walker, James, “The Effect of Rewards and Sanctions in Provision of Public Goods,” *Working Paper 05-04*, Department of Economics, Ball State University, 2005.
- Simonsohn, Uri, “Review of *Moral Sentiments and Material Interests*,” *Journal of Economic Literature*, XLIV, September 2006, 745-747.

StataCorp, *Stata Base Reference Manual, Release 9* (College Station, TX: Stata Press, 2005, volume 1: A-J).

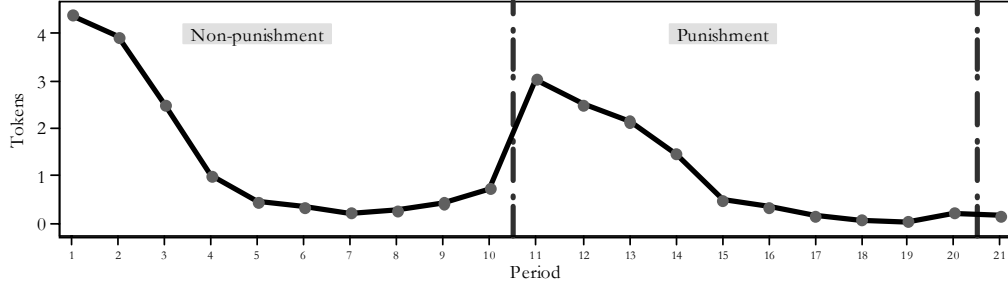
Sutter, Matthias; Haigner, Stefan, and Kocher, Martin, "Choosing the stick or the carrot?-Endogenous institutional choice in social dilemma situations," *Discussion Paper No. 5497*, Centre for Economic Policy Research, London, February 2006.

Appendix: Detailed Results (NOT FOR PUBLICATION)

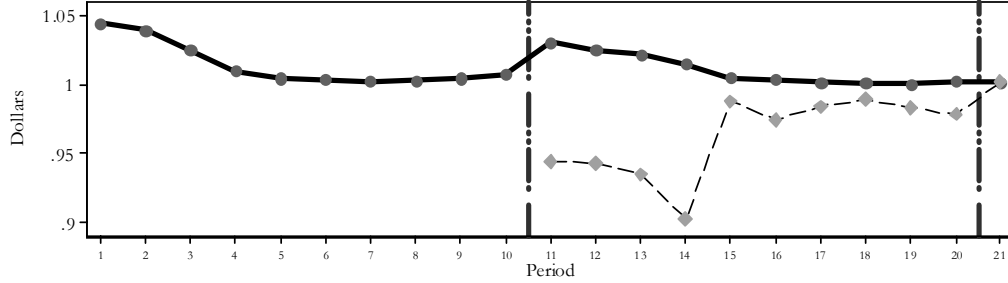
Results in Session 1 (N=26 Perfect Strangers)

Low Return to Public Good

Average Token Contributions



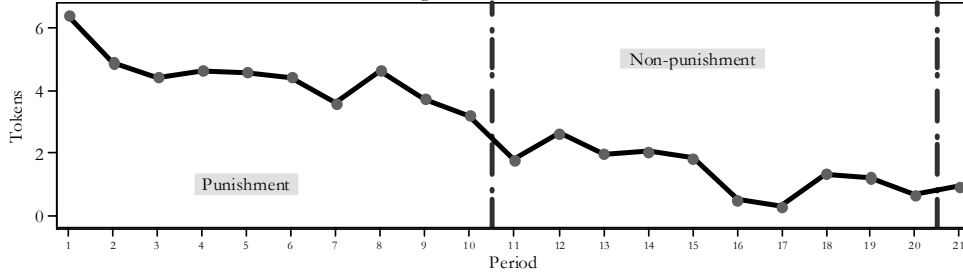
Average Dollar Profits



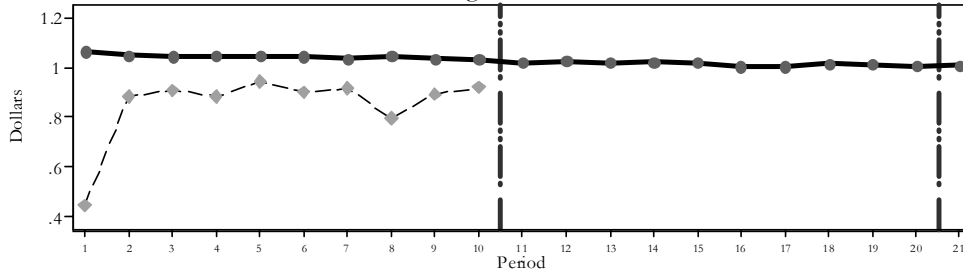
Results in Session 2 (N=24 Perfect Strangers)

Low Return to Public Good

Average Token Contributions

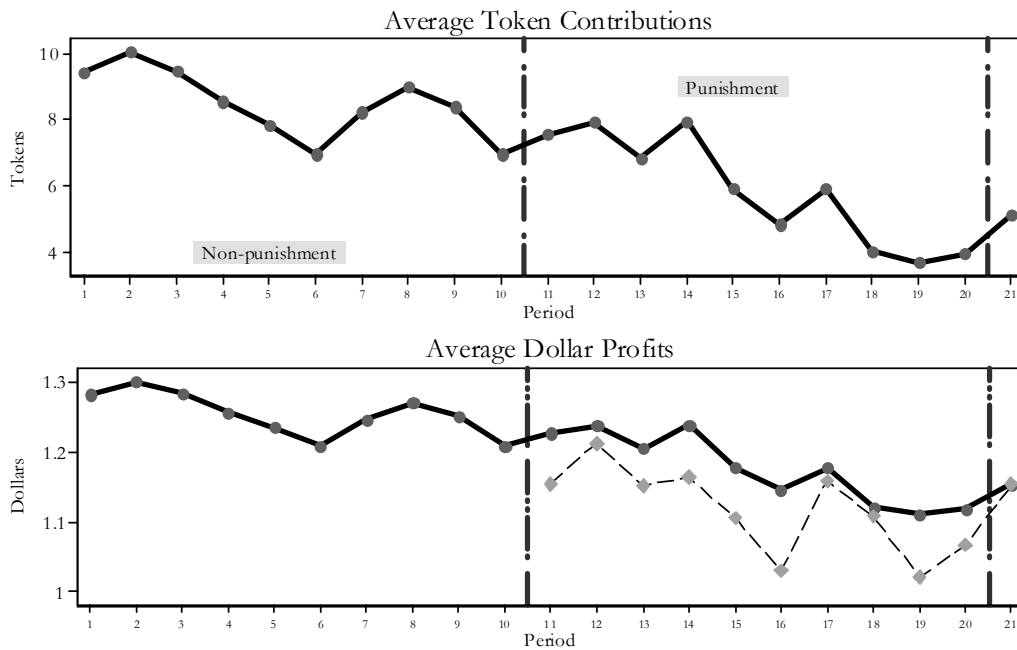


Average Dollar Profits



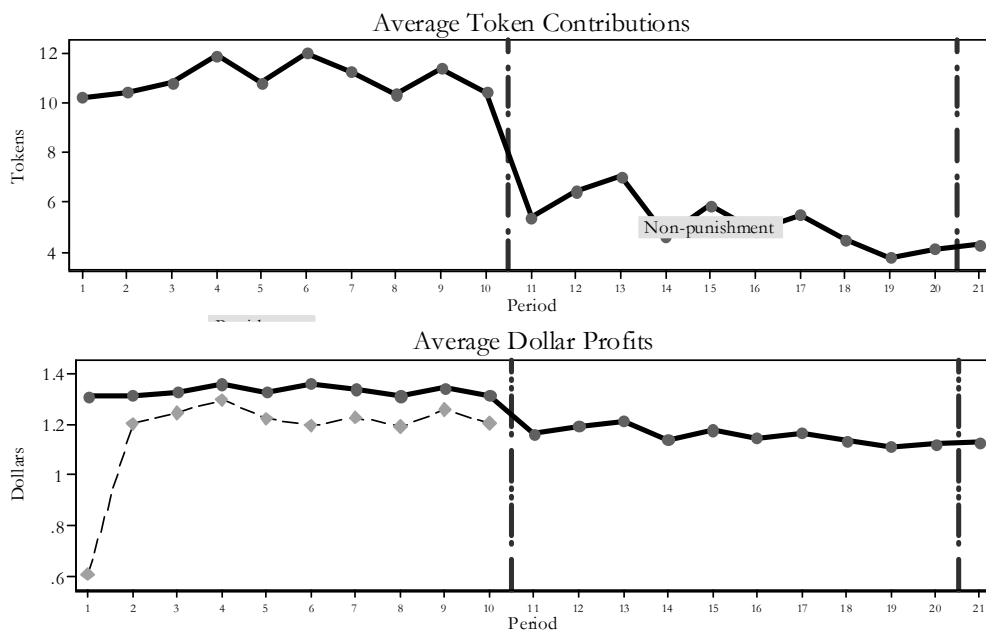
Results in Session 3 (N=26 Perfect Strangers)

High Return to Public Good



Results in Session 4 (N=26 Perfect Strangers)

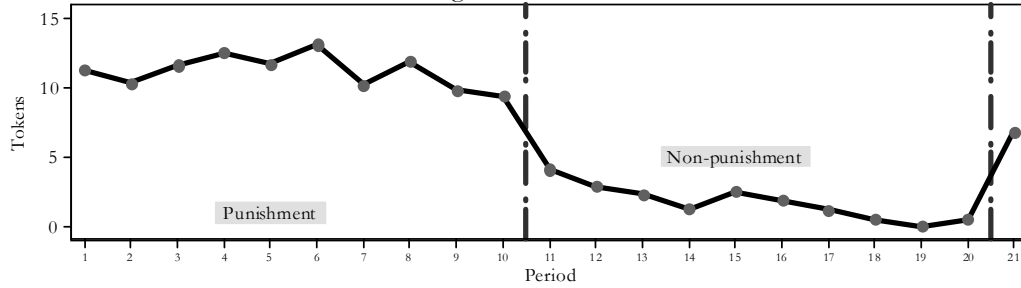
High Return to Public Good



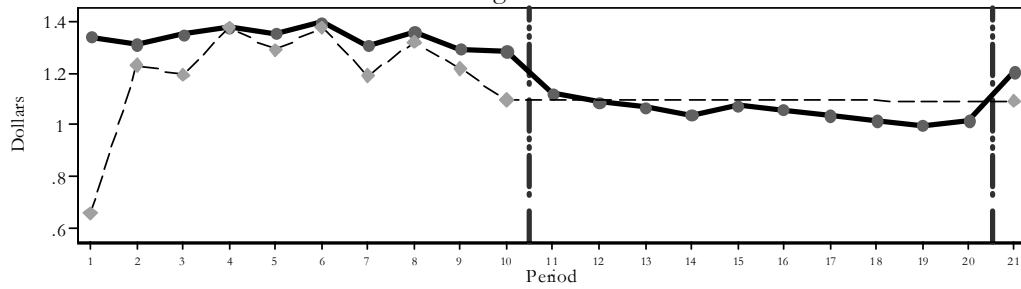
Results in Session 5 (N=10 Random Strangers)

High Return to Public Good

Average Token Contributions



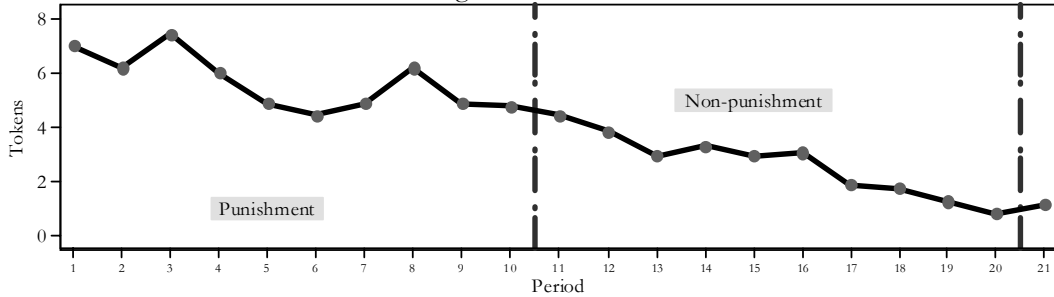
Average Dollar Profits



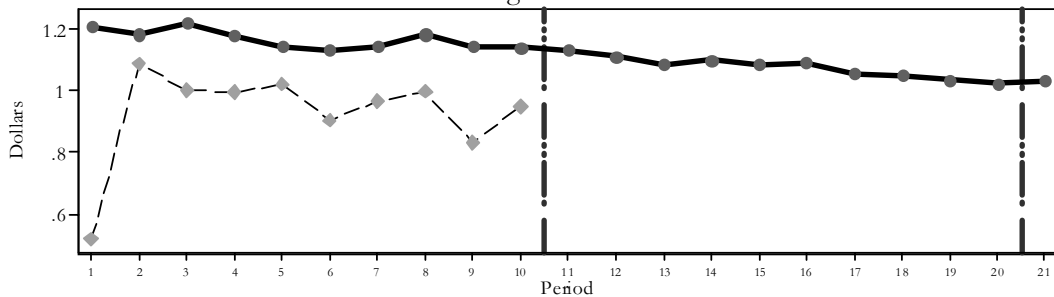
Results in Session 6 (N=16 Random Strangers)

High Return to Public Good

Average Token Contributions



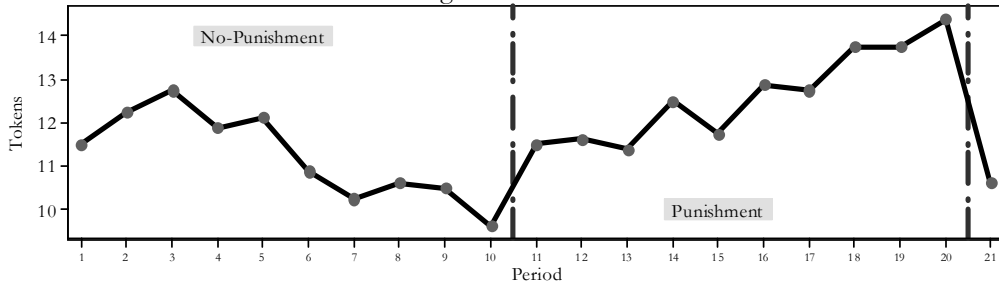
Average Dollar Profits



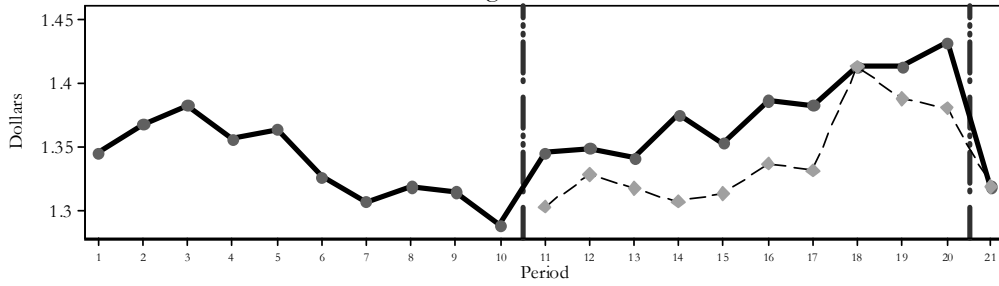
Results in Session 7 (N=8 Random Strangers)

High Return to Public Good

Average Token Contributions



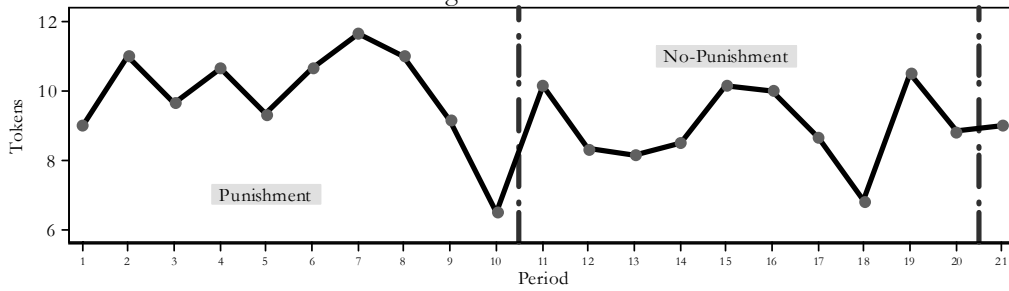
Average Dollar Profits



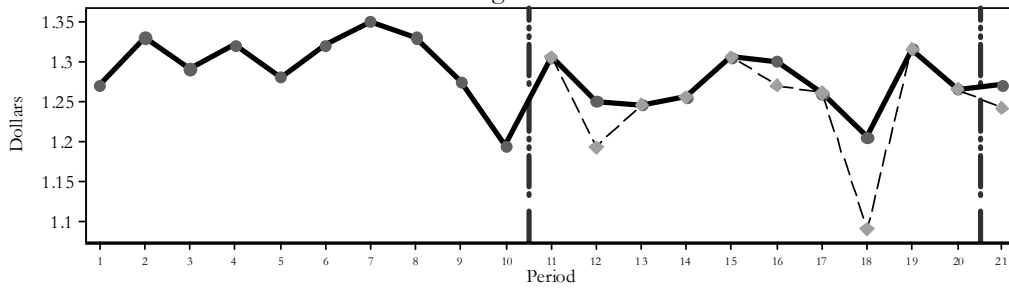
Results in Session 8 (N=6 Random Strangers)

High Return to Public Good

Average Token Contributions



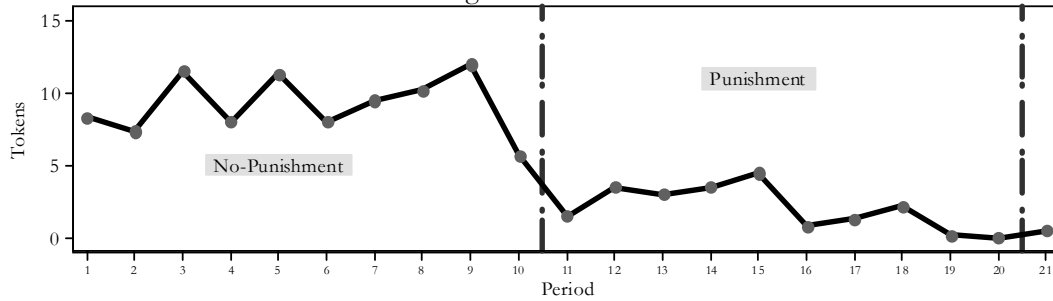
Average Dollar Profits



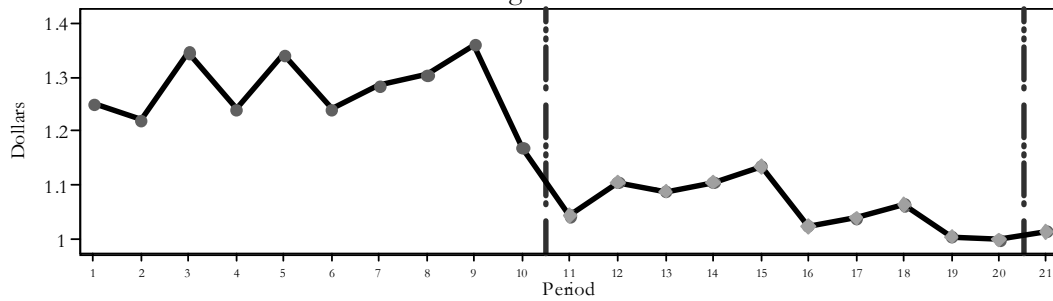
Results in Session 9 (N=6 Random Strangers)

High Return to Public Good

Average Token Contributions



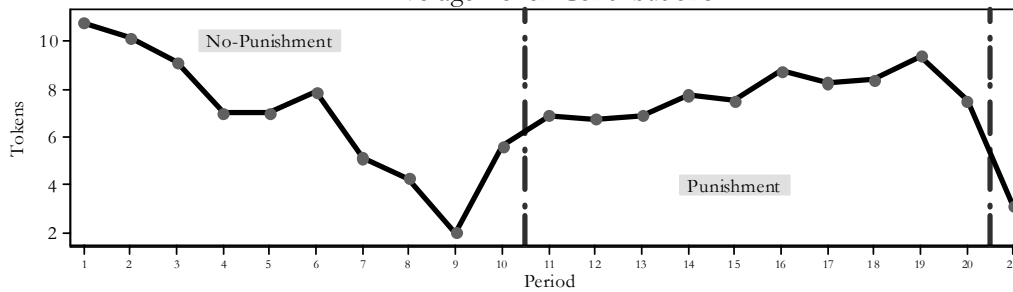
Average Dollar Profits



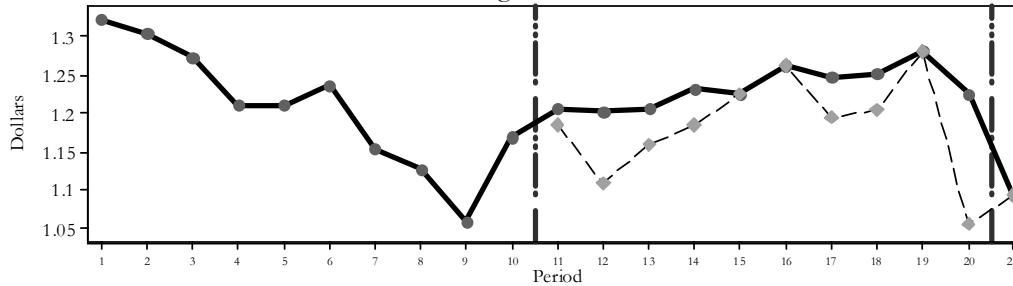
Results in Session 10 (N=8 Random Strangers)

High Return to Public Good

Average Token Contributions



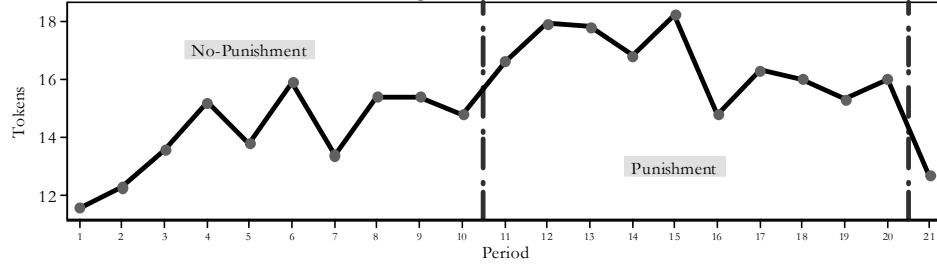
Average Dollar Profits



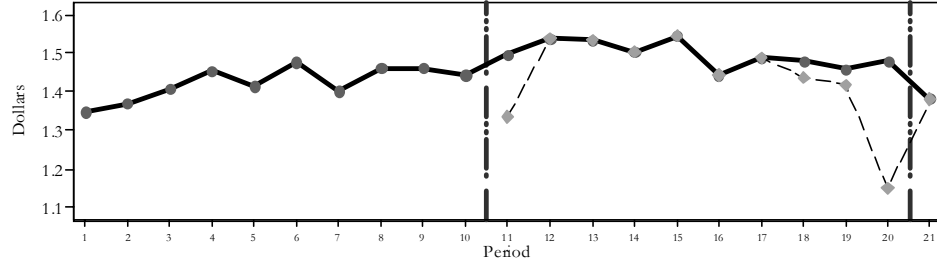
Results in Session 11 (N=10 Random Strangers)

High Return to Public Good

Average Token Contributions



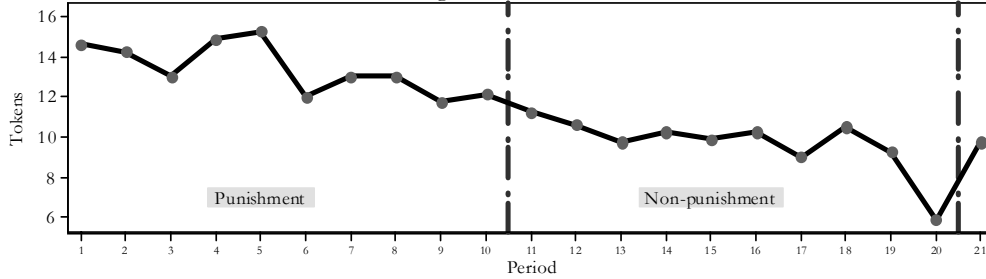
Average Dollar Profits



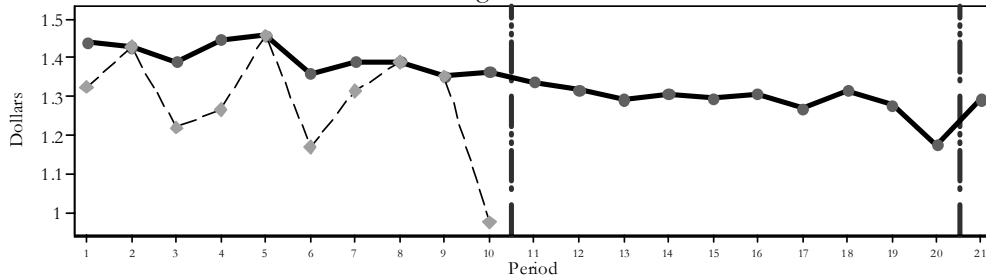
Results in Session 12 (N=8 Random Strangers)

High Return to Public Good

Average Token Contributions



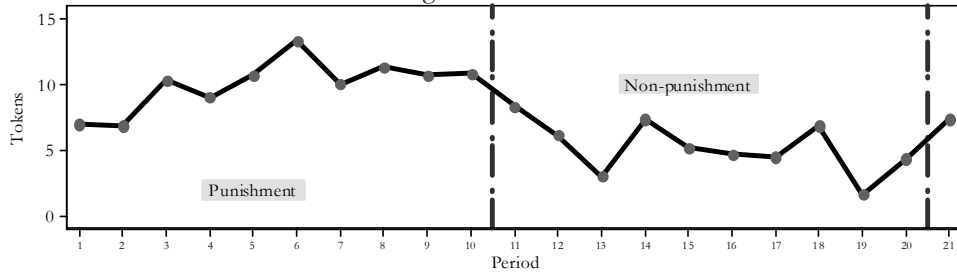
Average Dollar Profits



Results in Session 13 (N=6 Random Strangers)

High Return to Public Good

Average Token Contributions



Average Dollar Profits

