ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Egas, Martijn; Riedl, Arno

Working Paper

The Economics of Altruistic Punishment and the Demise of Cooperation

Tinbergen Institute Discussion Paper, No. 05-065/1

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Egas, Martijn; Riedl, Arno (2005) : The Economics of Altruistic Punishment and the Demise of Cooperation, Tinbergen Institute Discussion Paper, No. 05-065/1, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/86431

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

TI 2005-065/1
Tinbergen Institute Discussion Paper

The Economics of Altruistic Punishment and the Demise of Cooperation

Martijn Egas[1]
Arno Riedl[2]

[1] *Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam,*
[2] *Center for Research in Experimental Economics and Political Decision-Making, University of Amsterdam, and Tinbergen Institute.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

The Economics of Altruistic Punishment and the Demise of Cooperation

Martijn Egas[1] and Arno Riedl[2]

[1] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, The Netherlands, e-mail: egas@science.uva.nl
[2] Tinbergen Institute and Center for Research in Experimental Economics and political Decision-making, University of Amsterdam, The Netherlands, e-mail: a.m.riedl@uva.nl

Version: 1.0 (June 16, 2005)
Word count: 2520 (excluding title, affiliations, references, figure legends, and supporting material)

Abstract:
Explaining the evolution and maintenance of cooperation among unrelated individuals is one of the fundamental problems in biology and the social sciences. Recent experimental evidence suggests that altruistic punishment is an important mechanism to maintain cooperation among humans. In this paper we explore the boundary conditions for altruistic punishment to maintain cooperation by systematically varying the cost and impact of punishment, using a subject pool which extends beyond the standard student population. We find that the economics of altruistic punishment lead to the demise of cooperation when punishment is relatively expensive and/or has low impact. Our results indicate that the 'decision to punish' comes from an amalgam of emotional response and cognitive cost-benefit analysis. Additionally, earnings are lowest when punishment promotes cooperation, suggesting that the scope for altruistic punishment as a means to maintain cooperation is limited.

Recently, the puzzle of human cooperation experiences a revival of interest (1-9). In particular, altruistic punishment has been shown to effectively enforce cooperation, even among unrelated and anonymous humans (3,5). People punish uncooperative individuals at a cost to themselves, inducing cooperation of the sanctioned individuals. These results challenge our view of human behavior in social dilemma situations: cooperation is maintained even in the absence of traditional mechanisms such as reciprocity and reputation. Importantly, in these experiments the conditions for altruistic punishment were favorable in terms of effectiveness: low cost and high impact. However, it is highly unlikely that such favorable conditions prevail, both during human evolution and in common interactions in daily life.

Therefore, it is important to explore boundary conditions for altruistic punishment to maintain cooperation.

Preliminary experimental results suggest that the costs an individual incurs when punishing affects the frequency of such punishment negatively (10). Also, intuitively, only if punishment has high impact may individuals be effectively coerced to avoid such sanctioning in the future. Indeed, theory shows that free-riding and punishment are alternative stable states in simplified versions of the altruistic punishment game (11,12). The argument is that rare punishers in a group of free-riders would have to incur high costs to punish all free-riders. Conversely, rare free-riders in a group of punishers would experience lots of punishment. Hence, a critical mass of punishers is needed to guarantee effective sanctioning of free-riding. Furthermore, it is hypothesized that the threshold frequency of altruistic punishers increases with higher costs and with lower impact of punishment, as these determine the payoff of the punishers and the punished. The success of altruistic punishment should therefore depend on the initial composition of the subject group, as well as on the costs and impact of punishment. Varying costs and impact of punishment may therefore have dramatic effects on cooperation.

We conducted an altruistic punishment experiment with real money at stake where we systematically varied both the cost and the impact of punishment. This design allows us to assess the role of economic cost-benefit analysis in altruistic punishment. We implemented four different treatments with punishment and a control treatment without punishment. In each treatment, subjects engaged in a public good game in groups of three. Each subject was endowed with 20 experimental money units (EMUs) and could contribute between 0 and 20 EMU of this endowment to a group project. Each EMU invested returned 0.5 EMU to each of the three group members. From an individual perspective, each EMU kept paid off one EMU whereas each invested EMU only paid off 0.5 EMU. Hence, material self-interest dictated to contribute nothing to the group project. Collectively it is optimal for everybody to contribute the entire endowment. If all group members kept their endowment, everybody earned 20 EMU. If all contributed their entire endowment to the group project each would earn 60*0.5=30 EMU. Decisions were made simultaneously and anonymously.

After decisions in the public good game were made, each group member was informed about the other group members' contributions and the resulting earnings. Nobody was ever informed about the identity of the other group members. In the treatment without punishment this ended the interaction between the group members. In the punishment treatments each member had the possibility to punish other members by assigning between 0 and 10 punishment points (PPs) to each of the two other members. Importantly, the punishment treatments differed in the cost and impact of punishment. In treatment T13, which is akin to the standard punishment experiment (3), each assigned PP costs the punisher 1 EMU and reduces the payoff of the punished with 3 EMU. In T31 the costs per assigned PP were 3 EMU for the punisher and only 1 EMU for the punished. In treatments T11 and T33 this relation was 1:1 and 3:3, respectively. For convenience, we call the cost to impact ratio *effectiveness* of punishment. The control treatment without punishment is indicated by T00. All punishment decisions were made simultaneously and anonymously.

Because the act of punishing comes at a cost, purely selfish subjects will not punish in any treatment. Given that nobody punishes and because not contributing is a dominant strategy in the public good game, selfish individuals will also not contribute in any treatment.

To allow for learning the experiment was repeated for six rounds in each replicated group of subjects. One replicated group consisted of 18 subjects, exposed to exactly one treatment. To exclude potential effects of reciprocation (13,14) and reputation building (1,15-17) we ensured that no subject ever met another subject more than once. Hence, also in the repeated interaction purely selfish individuals do not punish or contribute to the public good. Previous work (3,18), however, indicates that at least in environments where punishment is relatively cheap and has strong impact the frequency of altruistic punishment is surprisingly high, stimulating high contribution rates in the public good game. Below we show that these results do not survive our variations in the effectiveness of punishment.

Because of the large-scale set-up (in total 846 participants) we used the internet to facilitate this experiment. Any Dutch-speaking person could participate, which extends our subject pool beyond the typically used undergraduate students and allows us to asses the external validity of such laboratory experiments. The socio-economic characteristics of the participants confirm that our subject pool differs from a students pool. The average gross income of our subjects was close to the actual average gross income in the Netherlands. The average age was 35 (range: 12-80). The educational level of our participants was relatively high for Dutch standards, but it covered all types of education from secondary school (3%) up to university degrees (33%). Female participants (28%) were underrepresented. (See "Supporting on-line material (SOM)" for details.)

Our first result concerns the degree of cooperation in the public good game. If sanctioning is anticipated, one may expect differences in initial contribution rates between the punishment treatments and the control treatment. However, we find no evidence for this. In all treatments, average contributions start off at about the same level (Fig. 1; Kruskal-Wallis test: χ^2=3.042, d.f.=4, p=0.551). However, the dynamics of cooperative behavior over the six rounds are strikingly different across treatments. Only when punishment is very effective (T13) contributions increase over rounds, in accordance with previous studies (3,18). In all other treatments, however, contributions are quickly declining (Fig. 1). This clearly indicates that the scope for punishment to maintain cooperation in the long run is limited by its effectiveness.

In all punishment treatments, inflicted punishment depends on the difference in contributions between the punisher and the punished. Generally, punishers allocate increasingly more PPs the more the other's contribution falls short of their own contribution (Fig. 2a). For convenience, in the following we use 'deviation in contribution' as the difference between the contribution of the focal participant and the contribution of her co-participant. (Some low constant level of punishment also occurs when co-participants invested more.)

The systematic variation of cost and impact of punishment allows us to investigate how participants change behavior in response to changes in these critical parameters. Comparing

punishment behavior across treatments clearly shows that significantly more PPs are dealt out when it is cheap and has high impact (T13) than when it is expensive and has low impact (T31). For intermediate effectiveness (T11 and T33) the allocated PPs lie between the other two treatments. To examine the differences between treatments statistically we performed Tobit regressions with PPs as dependent and deviations in contribution as independent variable (Fig. 2a, SOM). Interestingly, the regression analysis indicates that once participants punish no difference across treatments can be found. That is, the marginal propensity to increase punishment with increasing deviation in contribution is the same for all four treatments. This is manifested by the equality of the regression coefficients of deviations in contribution in all treatments (for details: Fig. 2a, SOM). However, the effectiveness of punishment has a significant effect on the threshold of deviation in contribution at which participants start to punish free-riders. The higher the costs and the lower the impact of punishment the larger the deviation has to be before participants start to punish at all. The estimated deviation threshold is significantly increasing from 2.41 (T13) to 5.34 (T11) to 8.33 (T33) to 11.3 (T31) (two-sided non-linear Wald-tests, $p<0.03$ in all cases). The significant difference in thresholds between T33 and T11 indicates that coarser smallest units in terms of costs and impact make participants more reluctant to punish, even if the effectiveness of punishment is the same. These results show that the amount of altruistic punishment follows an economic logic. The allocated PPs decrease significantly with decreasing effectiveness of punishment. This difference is solely due to an increasing threshold in the deviation of contribution at which participants start punishing.

Similar results are found for the frequency of punishing free-riders, which decreases monotonically with decreasing effectiveness of punishment. Examination based on logit regressions (SOM) shows that the differences in punishment frequencies are solely due to a shift in the deviation threshold at which participants start to punish free riders. The marginal propensity to punish is the same in all treatments. Furthermore, the frequency of punishment is monotonically increasing with the deviation from the punishers' contributions (Fig. 2b, SOM).

The overall effect of punishment on the punished differs rather dramatically between the most effective treatment T13 and the other three punishment treatments. In all categories of deviation in contribution punished participants suffer much more in T13 than in the other treatments. Tobit regressions corroborate this finding (Fig. 3, SOM). Particularly, in T13 already very small deviations in contribution lead to punishment. In the other treatments, the punishment effect sets in only at rather large deviations (6.83 in T11, 7.35 in T33, and 12.0 in T31). Also, in T13 the punishment effect increases faster with deviation in contribution than in the other treatments. Consequently, for large deviations in contribution the average effect on a punished participant is several times larger in T13 than in the other treatments. Note that the effect of punishment in EMUs is remarkably similar in treatments T11 and T33, as may be expected when punishment effectiveness governs this variable, rather than impact or cost of punishment alone. The threshold difference in the effect of punishment suffices to explain the dynamics of cooperation over the six rounds across the punishment treatments (Fig. 1).

In our experiment, cooperation is only maintained in case punishment is cheap and has high impact (T13). Cooperative behavior diminishes rapidly in all other treatments. Less effective sanctioning regimes cause a dramatic drop in the effect of punishment on defectors. This results from differences in thresholds across treatments at which participants start to deal out punishment points to free-riders. Importantly, our results are in line with the theoretical prediction that cooperation induced by (the threat of) punishment and defection are alternative attracting states, as mentioned above.

In T13 a proximate fairness theory (20,21) could explain observed behavior. This theory states that inequity-averse individuals will punish free-riders because punishing reduces payoff differences between punishers and punished. In treatments T11, T33, and T31, this model cannot explain the still existing non-negligible amount of punishment. In these treatments punishment either leaves the relative payoffs the same or even increases the inequality to the disadvantage of the punisher. Under such conditions fairness theory predicts behavior indistinguishable from purely selfish behavior. This indicates that other factors, like emotions (19,22-24), may play an important role in punishing. However, as our study clearly shows altruistic punishment is also guided by economic incentives. Altruistically punishing humans also take the effectiveness of punishment into account and adjust behavior accordingly. It seems that the 'decision to punish' comes from an amalgam of emotional response and cognitive cost-benefit analysis. The precise relationship of these different factors is still to be disclosed.

Considering the average payoff per group uncovers a sobering picture: in the one treatment (T13) where punishment successfully increases cooperation, groups earn significantly less than groups in any of the other treatments (Fig. 4). Furthermore, compared to the control treatment without punishment, earnings in T13 (as well as all other punishment treatments) are clearly smaller and show no statistically significant tendency to catch up. On top of that, the average actual earnings in T13 are even lower than the potential earnings of full free-riding without punishment. The reason for this result is that the opportunity to punish is used most frequently in T13 but the increase in contributions is not sufficient to compensate for the cost of punishment. Altruistic punishment destroys resources. It might be expected that the earnings in T13 would rise and eventually exceed that in the other treatments as a result of increasing cooperation and decreasing incidence of punishment. However, it is evident from our results that it will take many more rounds before the cumulative income in T13 will exceed that in the other treatments. This casts into doubt group-selection models for the evolution of cooperation by altruistic punishment (e.g. 12), where sanctioning within a group is supposed to increase cooperation and thereby the competitive strength of that group. These models assume that the cooperation rate *per se* determines the competitive advantage of a group. When taking the costs of punishment into account, however, the competitive advantage of groups with altruistic punishers appears severely mitigated. In experiments where participants are facing the same co-participants in every round (so-called “partner” setup (3,24)), cooperation levels rise much faster, potentially allowing for higher earnings under effective altruistic punishment within a smaller number of rounds.

In line with theoretical models (11,12), our results indicate that defection and cooperation under the threat of punishment are alternative stable states, where the effectiveness of punishment determines whether an anonymous group of people gravitate to cooperation or defection. Our study clearly suggests that altruistic punishment is important in the evolution of cooperation only in combination with other cooperation-enhancing mechanisms such as reputation and reciprocity or the possibility of opting-out (26). The reason is that altruistically punishing humans also take the costs and effects of their actions into account.

References

1. Nowak, M.A., Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. Nature 393:573-577.
2. Wedekind, C., Milinski, M. (2000). Cooperation through image scoring in humans. Science 288:850-852.
3. Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. Nature 415:137-140.
4. Milinski, M., Semmann, D., Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons". Nature 415:424-426.
5. Fehr, E., Fischbacher, U. (2003). The nature of human altruism. Nature
6. Fehr, E. & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. Nature 422:137-140.
7. Doebeli, M., Hauert, C., Killingback, T. (2004). The evolutionary origin of cooperators and defectors. Science 306: 859-862.
8. Hauert, C., Doebeli, M. (2004). Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature 428: 643-646.
9. Nowak, M.A., Sasaki, A., Taylor, C., Fudenberg, D. (2004). Emergence of cooperation and evolutionary stability in finite populations. Nature 428: 646-650.
10. Kosfeld, M. & Riedl, A. (2004). The design of (de)centralized punishment institutions for sustaining cooperation. Forthcoming in Raith M. (ed.) Procedural approaches to conflict resolution (Springer: Berlin and New York).
11. Sigmund, K., Hauert, C., Nowak, M.A. (2001). Reward and punishment. Proceedings of the National Academy of Sciences USA 98: 10757-10762.
12. Boyd, R., Gintis, H., Bowles, S., Richerson, P.J. (2003). The evolution of altruistic punishment. Proceedings of the National Academy of Sciences USA 100: 3531-3535.
13. Trivers, R. (1971). The evolution of reciprocal altruism. Quarterly Review of Biology 46: 35-57.
14. Axelrod, R., Hamilton, W.D. (1981). The evolution of cooperation. Science 211: 1390-1396.
15. Sugden, R. (1986). The Economics of Rights, Cooperation and Welfare. Blackwell, Oxford.
16. Alexander, R.D. (1987). The Biology of Moral Systems. Aldine de Gruyter, New York.

17. Leimar, O., Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. Proc. R Soc. Lond. B 268:745-753.
18. Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments. American Economic Review 90: 980-994.
19. Gächter, S. (personal communication) also reports similar results for experiments on altruistic punishment with subjects pools consisting of students but also of rural and urban adults in in Germany, Switzerland, Russia and Belarus. Gächter also reports that emotions seem to be an important proximate mechanism sustaining altruistic punishment.
20. Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114: 817-868
21. Fowler, J.H., Johnson, T., Smirnov, O. (2005). Egalitarian motive and altruistic punishment. Nature 433: E1-E2.
22. Bosman, R., Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. The Economic Journal 112: 147-169.
23. Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. Science 300: 1755-1758.
24. de Quervain, D.J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E. (2004). The neural basis of altruistic punishment. Science 305:1254-1258.
25. Masclet, D., Noussair, C., Tucker, S., Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review 93: 366-380.
26. Fowler, J.H. (2005). Altruistic punishment and the origin of cooperation. Proceedings of the National Academy of Sciences USA 102: 7047-7049.

FIGURES AND LEGENDS:

Fig. 1 (Dynamics of cooperation over the rounds)

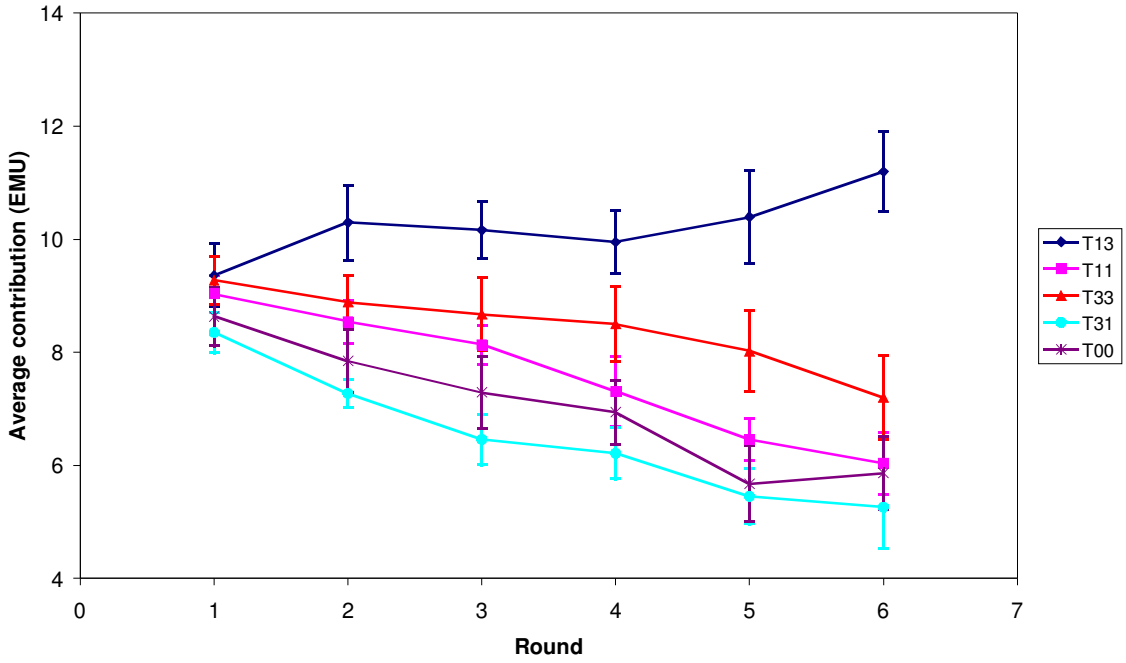
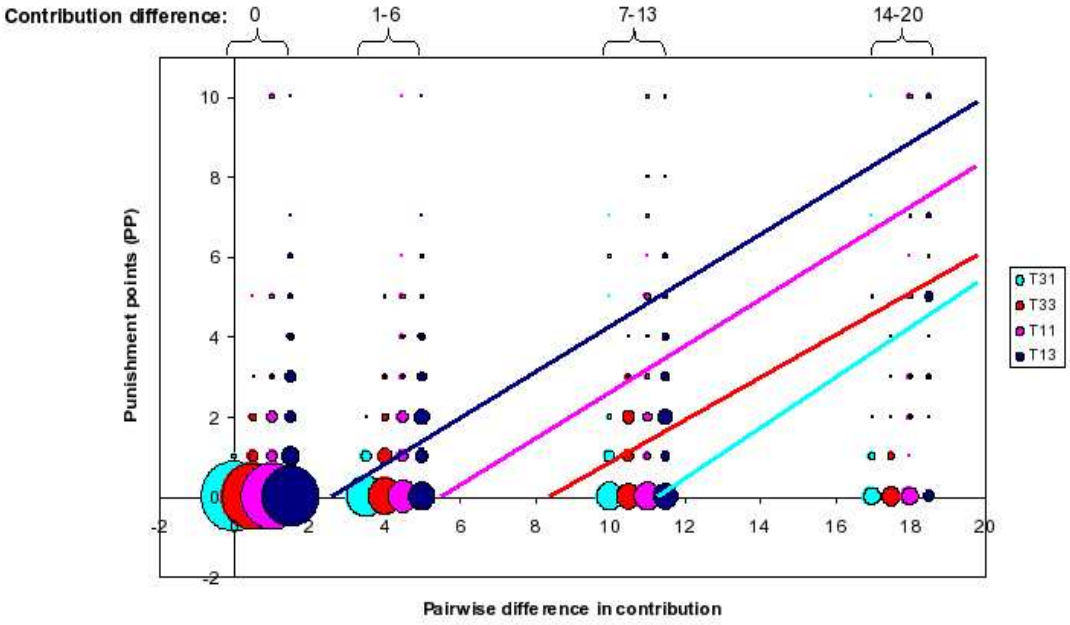


Figure 1 The average contribution (± 1 s.e.) to the public good in all five treatments over the six rounds of the experiment. In all treatments, the initial contribution rate is approximately 9 EMU (45% of the endowment). Only in T13 cooperation increases over rounds; in all other cases cooperation is in clear decline (repeated measures ANOVA, treatment: $F_{4,42}=11.522$, $p<<0.001$; round: $F_{5,210}=18.146$, $p<<0.001$; treatment*round: $F_{20,42}=4.056$, $p<<0.001$). Tukey posthoc tests showed that contributions in T13 differed from all other treatments, and that T33 differed from T31.

Fig. 2 (Punishment behavior; a) PPs, and b) punishment frequency)

a)



b)

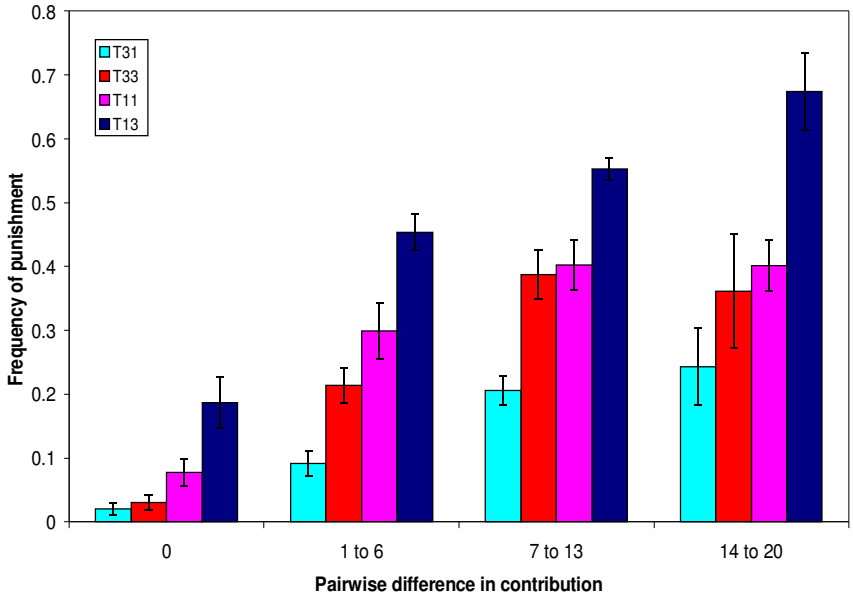


Figure 2 Punishment characteristics in the four punishment treatments as a function of the deviation in contribution of the punished participant: a) punishment points, b) punishment frequency.

Presented punishment data are across rounds. (Punishment did not vary significantly over the six rounds.)

(a) Punishment points (PPs) dealt out to "defecting" participants. Using four categories (0, 1-6, 7-13, 14-20) of deviation in contribution, the size of each circle indicates the relative frequency of PPs allocated to participants with such deviations in contribution. Straight lines represent linear Tobit regressions (SOM) of PPs on deviation in contributions separately for all four punishment treatments. PPs dealt out are significantly increasing with deviation in contribution ($p<0.01$ in all treatments). Slopes of the regression lines are not significantly different from each other (T13: 0.5561, T11: 0.5556, T33: 0.5107, T31: 0.5989; $X^2=1.84$, $p=0.6064$, joint test); intercepts are significantly different and have the order T13>T11=T33>T31; (T13 vs. T11, $X^2=7.81$, $p=0.0311$; T11 vs T33, $X^2=4.44$, $p=0.2111$; T33 vs T31, $X^2=19.12$, $p=0.0001$; all p-values Bonferroni adjusted for multiple comparisons); estimated deviation thresholds (*TH*) where participants start punishing are significantly different and have the order *TH*(T13=2.41)<*TH* (T11=5.34)<*TH*(T33=8.33)<*TH*(31=11.29); (T13 vs. T11, $X^2=13.29$, $p=0.0016$; T11 vs T33, $X^2=9.49$, $p=0.0124$; T33 vs T31, $X^2=8.08$, $p=0.0268$; all p-values Bonferroni adjusted for multiple comparisons).

(b) Frequency of punishment of defecting participants. Logit regression analysis (SOM) shows that the marginal likelihood that deviating participants are punished when increasing deviation in contribution is significantly increasing in all treatments ($p<0.001$) but not significantly different across treatments (T13: 0.2114, T11: 0.1982, T33: 0.2088, T31: 0.2444; $X^2=4.29$, $p=0.2319$, joint test). The estimated intercepts follow the order T13>T11=T33>T31, where inequalities indicate statistically significant differences (T13 vs. T11, $X^2=10.84$, $p=0.0059$; T11 vs T33, $X^2=2.89$, $p=0.5335$; T33 vs T31, $X^2=17.96$, $p=0.0001$; all p-values Bonferroni adjusted for multiple comparisons).

Fig. 3 (Effect of punishment)

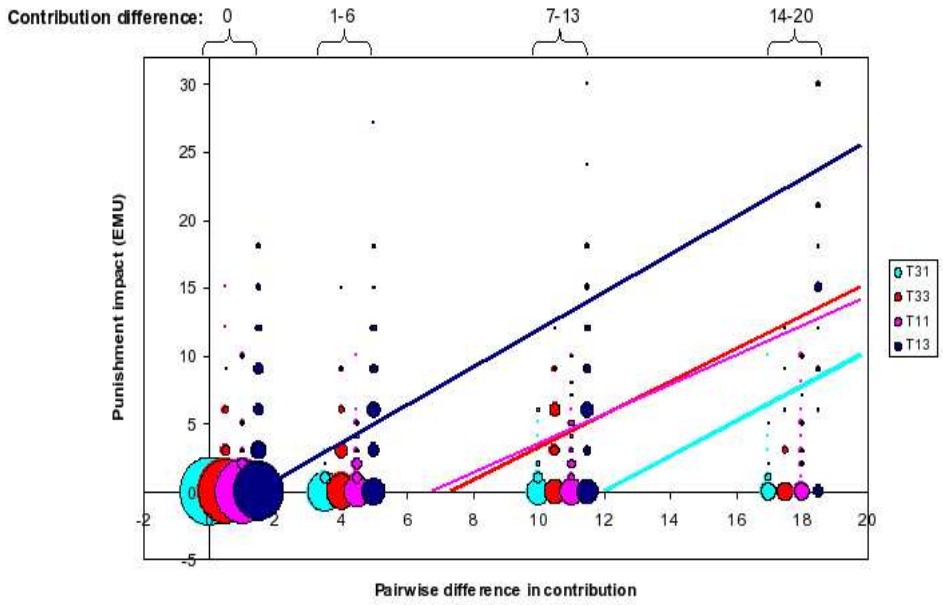


Figure 3 Effect of punishment on punished participant in the four punishment treatments as a function of the deviation in contribution of the punished participant.

Presented data are across rounds. (Effect of punishment did not vary significantly over the six rounds.)

Using four categories (0, 1-6, 7-13, 14-20) of the deviation in contribution, the size of each circle indicates the relative frequency of the effect of punishment (in EMU) on participants with such deviations in contribution. Straight lines represent linear Tobit regressions of PPs on deviation in contribution separately for all four punishment treatments. Slopes of the regression lines are not significantly different from each other except for the comparison of T13 with T11. (T13 vs. T11, X^2=10.13, p=0.0088; for all other comparisons p>0.6295; all p-values Bonferroni adjusted for multiple comparisons); intercepts are significantly different and have the order T13>T11=T33>T31; (T13 vs. T11, X^2=16.39, p=0.0003; T11 vs T33, X^2=1.41, p=1.0000; T33 vs T31, X^2=23.14, p=0.0000; all p-values Bonferroni adjusted for multiple comparisons); estimated deviation thresholds (*TH*) at which participants start punishing are significantly different and have the order *TH*(T13=1.43)<*TH*(T11=6.83)=*TH*(T33=7.35)<*TH*(31=11.95); (T13 vs. T11, X^2=59.27, p=0.0000; T11 vs T33, X^2=0.37, p=1.0000; T33 vs T31, X^2=23.97, p=0.0000; all p-values Bonferroni adjusted for multiple comparisons).

Fig. 4 (Relative gains over rounds)

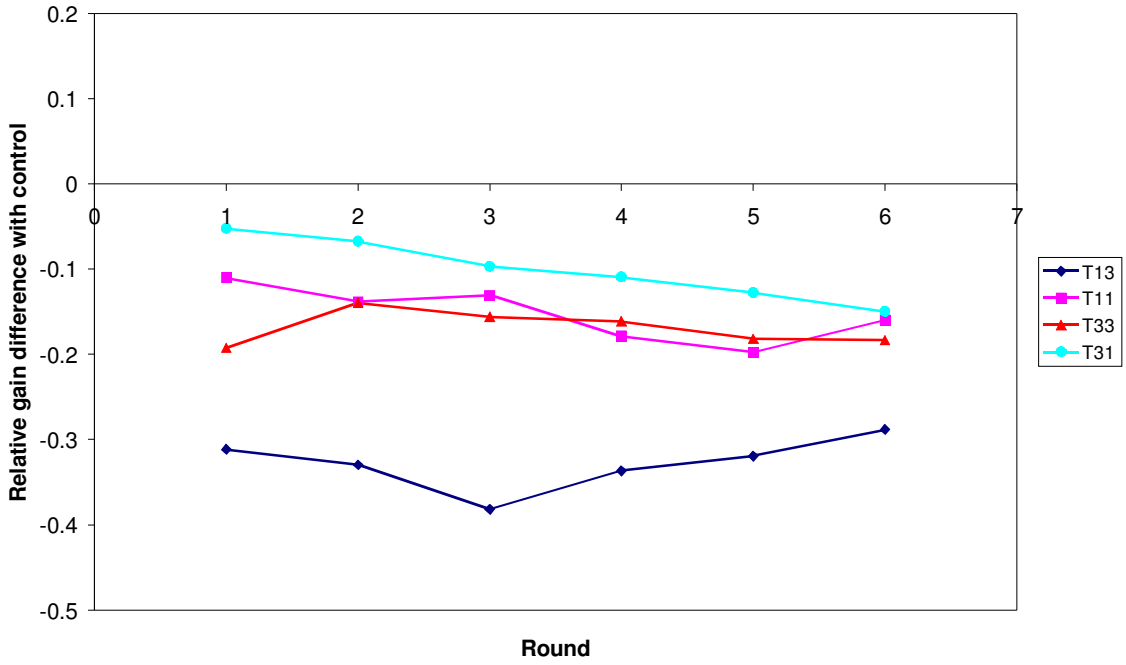


Figure 4 The average gains in earnings of groups in the four punishment treatments relative to the earnings in the control treatment without punishment over the six rounds of the experiment ([earnings in punishment treatment – earnings in control treatment]/earnings in control treatment). The average gains neither significantly increase nor decrease over rounds. (Spearman rank order correlation; T13: ρ =-0.7143, n=6, p=0.1108; T11: ρ =-0.0857, n=6, p=0.8717; T33: ρ =0.0857, n=6, p=0.8717; T31: ρ =-0.7714, n=6, P=0.0724). Relative gains in round 6 are all significantly smaller than zero ($p<0.02$, two-sided t-tests).

SUPPORTING ONLINE MATERIAL

Methods

A total of 846 people (28 percent female) participated in the experiment. The experiment was conducted via the internet on a secure server using client-based software. The subject pool consisted of volunteers from the Dutch-speaking (world) population with internet connection. The recruitment took place via mass media in the Netherlands (ads in newspapers and on radio) and the science web-site of the Dutch public broadcasting station VPRO. In all recruitment announcements we made sure that the actual content of the experiment was not revealed. The only information about the experiment given during the recruitment period was that a scientific experiment will take place with the possibility to earn money. (The title of the experiment was "Speel je rijk", which loosely translates as "Play to get rich"). No further information about the content was revealed until the last experimental session was finished. The announcements further communicated that the experiment was going to take place from May 24 to May 28, 2004, with two sessions per day (one at 4 pm and one at 8:30 pm), that one is only allowed to participate in one of the sessions, and that a session will take approximately 45-60 minutes. A person interested in participating was asked to send an email and to indicate two preferred sessions (dates and times). They were then sent an acknowledgement-email. This email contained the following information: (i) that a random lottery will decide whether he or she is chosen as an actual participant; (ii) that if he or she is chosen this information will be transmitted shortly (usually 24 hours) before the chosen session takes place. All together more than 4000 people subscribed for the experiment. From these around 1000 were selected as participants, 846 of which actually participated. (Not all selected people "showed up" at the experiment.)

A session was organized as follows. All selected participants received an email with a password and the www-address from where the experiment was going to start. With the password and his or her email address a participant could log into the experiment itself. There each participant received on-line instructions that explained the structure of the experiment in detail. That is, participants were explained how to make decisions, how to calculate earnings, how their own earnings and the earnings of others depend on their own decisions and the decisions of others, that they were going to play six rounds in groups of three, that these groups are going to be recomposed after each round to guarantee that nobody meets anyone twice, that all interaction will be anonymous, and that the history of behavior of participants was not going to be disclosed to anybody. After having read the instructions each participant had to answer a number of control questions which allowed us to check that (s)he understood the instructions. (Experimental instructions are available upon request from the authors.) These questions concerned the reshuffling of the groups after each round, the consequences of (not) contributing and (not) punishing, and the calculation of one's own earnings and the earnings of other group members in a number of hypothetical situations. Only if a subject answered all questions correctly he or she was allowed to participate in the experiment itself. (Only a few subjects dropped out in this phase.) During the instructions and the control

questions subjects had the possibility to ask questions to the experimenters (Martijn Egas and Arno Riedl) using a chat-window that was built in the software.

When a subject had answered all questions correctly he or she entered a waiting queue until a group of 18 participants was formed. Each of the 18 participants then played six rounds of the public good game with or without punishment, depending on the treatment (for a description of the treatments see the main text). The timing of our five treatments (T00, T13, T11, T31, T33) was determined beforehand and guaranteed a balanced distribution of afternoon and evening sessions across treatments. In each experimental session we implemented only one treatment. Since subjects could only participate in one session each participant was only participating in one treatment. Subjects were not aware of the fact that there were different treatments. The number of subjects participating in a session varied between 54 (three groups of 18) and 126 (seven groups of 18). In total, ten replicate groups of 18 participated in T13, T11 and T31, nine in T33 and eight in T00.

In the first round of a session the 18 participants of a group were randomly allocated to six subgroups of three. In the five subsequent rounds the groups of three were recomposed such that nobody met anybody else twice. No group member knew anything about the past behavior (contribution to the public good and punishment decisions) of the other group members. In each round everybody made a contribution decision simultaneously, which were subsequently disclosed to the other two subgroup members. Also, the earnings of all three subgroup members were shown to each of the subgroup members. Hence, everybody knew – in principle – the earnings of the other members in the subgroup. In T00 this ended the round, but in the treatments with punishment this was the stage where participants could assign between 0 and 10 punishment points (called "deduction points" in the experiment) to each of the other two subgroup members. Finally, each participant was informed how many punishment points in total (s)he received and what the net earnings over this round amounted to. Total earnings were accumulated across rounds.

After the sixth round participants were asked to fill out a questionnaire with questions on their socio-economic background. The participants were informed that they had to answer all questions in order to be able to receive the money they earned in the experiment. Since some of the questions could be regarded as sensitive and/or private information, the option "no answer" was provided for each question (except age). Subjects were informed about this before they entered the questionnaire. Complete anonymity was guaranteed as well as privacy protection: the answers will only be used by the experimenters for scientific ends, and not coupled to the names of the participants. All participants except three filled in the questionnaire. Furthermore, an overwhelming majority did not use the "no answer" option for any question, although use of this option varied with the content of the questions (see **Socio-economic characteristics of the subject pool**).

Finally, each participant was asked to give his or her name, place of residence and bank account number to be able to transfer their earnings. A group of 18 participants was typically finished after 40 minutes. The average earnings were 12.2 euros per participant.

Socio-economic characteristics of the subject pool

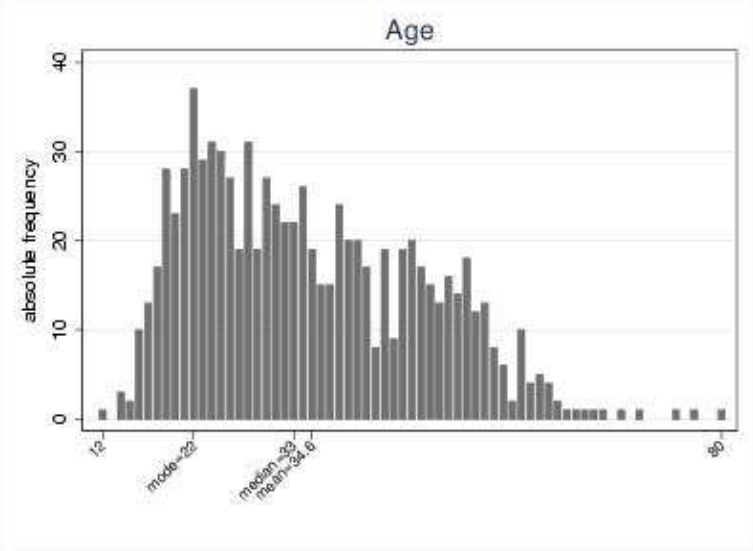
The socio-economic characteristics of the participants differ clearly from a student population, where participants of controlled laboratory experiments usually belong to. The subject pool is a reasonable reflection of Dutch society, although it also reflects the fact that the experiment was conducted via the Internet.

There are relatively few female participants (28%). The age distribution of the subject pool lies between 12 and 80, with an average of 34.6 and a median of 33 (Fig. A1a). A clear majority (58%) of our participants is either employed or self-employed, whereas only 29% is still in training (pupils, college and university students). The remaining subjects (13%) are either not employed, retired or are not covered by any of these categories (Fig A1b). The (gross) income distribution reflects the occupation in that a sizeable fraction (those still in training) has a low income. The modal income is in reasonable agreement with that of The Netherlands (roughly 2000 euro per month; Fig. A1c).

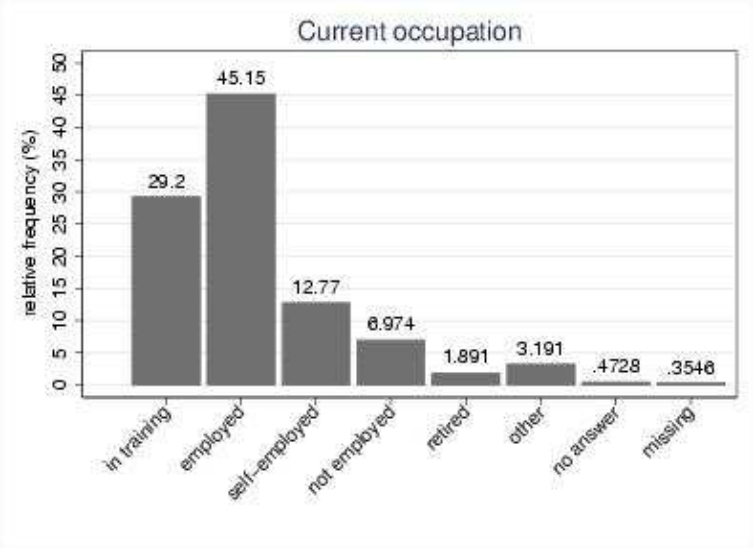
The majority of participants (65%) does not have any children (likely due to the over representation of younger adults; Fig. A1d). Nevertheless, 83% does share a household with other people (Fig. A1e) and virtually everybody has at least one sibling (Fig. A1f). Also, only 10% of the participants did not vote in the last national elections (88% answered "yes" to this question). The distribution of participants over the political parties shows a left-liberal bias compared to the outcome of the most recent election (Fig. A1g).

Figure A.1: Socio-economic background of participants

a)



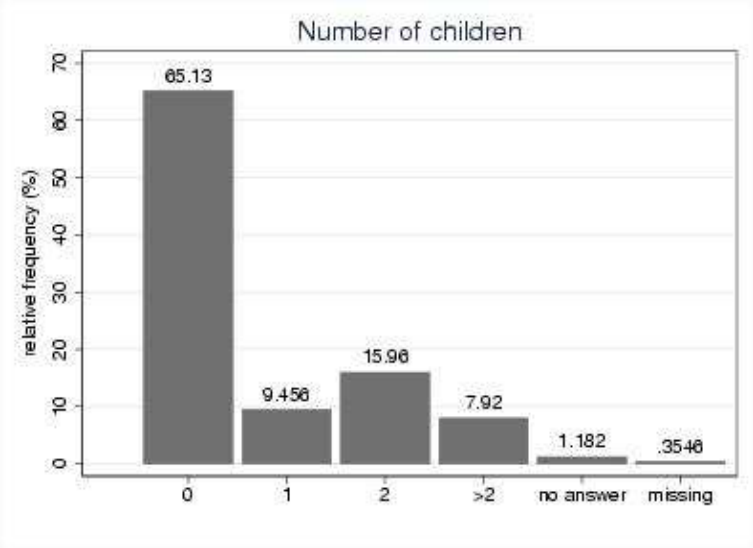
b)


Current occupation
relative frequency (%)
0
5
10
15
20
25
30
35
40
45
50
29.2
45.15
12.77
6.974
1.891
3.191
.4728
.3546
in training
employed
self-employed
not employed
retired
other
no answer
missing

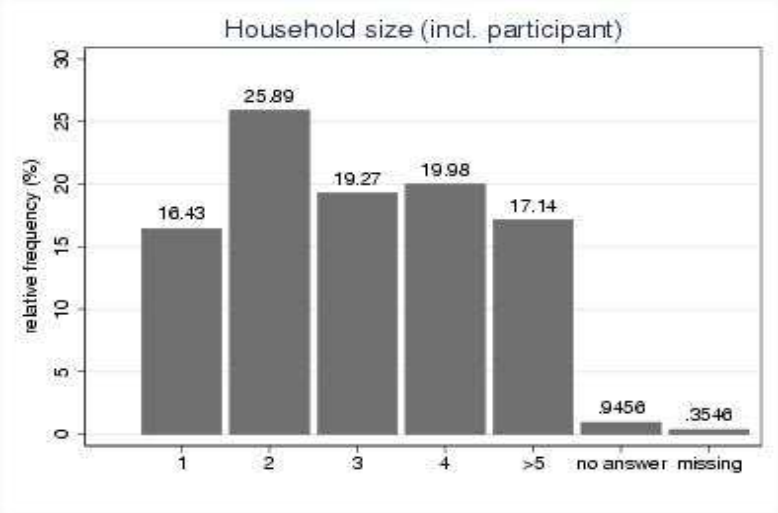
c)


Monthly gross salary (in euro)
relative frequency (%)
0
5
10
15
20
25
30
23.4
25.65
21.04
13.59
7.683
4.255
4.019
.3546
up to 600
601-1800
1801-3000
3001-4200
4201-5400
more than 5400
no answer
missing

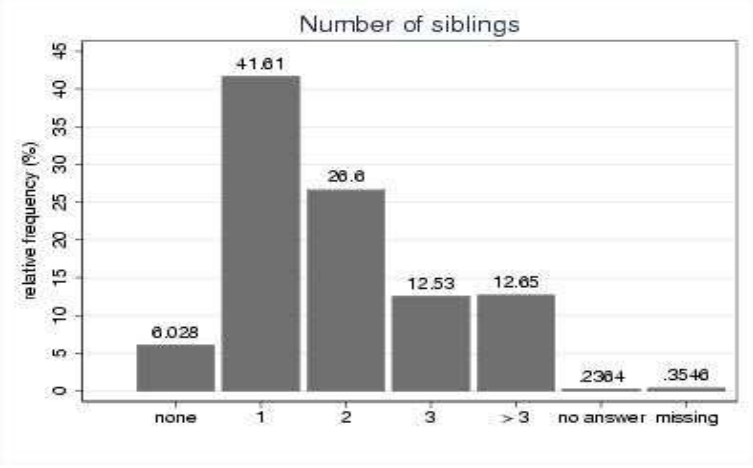
d)


Number of children
relative frequency (%)
0
10
20
30
40
50
60
70
65.13
9.456
15.96
7.92
1.182
.3546
0
1
2
>2
no answer
missing

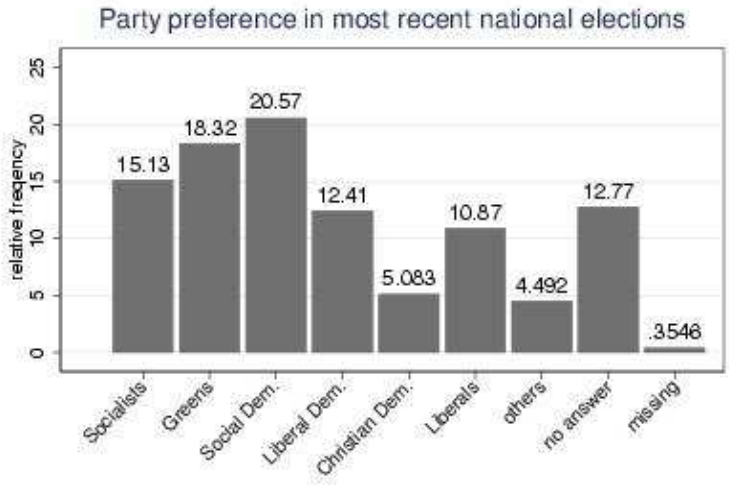
e)


Household size (incl. participant)
relative frequency (%)
0
5
10
15
20
25
30
16.43
25.89
19.27
19.98
17.14
.9456
.3546
1
2
3
4
>5
no answer
missing

f)


Number of siblings
relative frequency (%)
0
5
10
15
20
25
30
35
40
45
6.028
41.61
26.6
12.53
12.65
.2364
.3546
none
1
2
3
>3
no answer
missing

g)


Party preference in most recent national elections
relative freqency
0
5
10
15
20
25
15.13
18.32
20.57
12.41
5.083
10.87
4.492
12.77
.3546
Socialists
Greens
Social Dem.
Liberal Dem.
Christian Dem.
Liberals
others
no answer
missing

Regression analysis:

Legend to the following regression tables:
"T...dummy" represents the intercept of treatment "..."; "neg.dev." ("pos.dev.") is the negative (positive) deviation in contribution of punished participant; "neg.dev.*...dummy" ("pos.dev.*...dummy") is an interaction term of "neg.dev." ("pos.dev.") with the treatment "..." and reflects the marginal change of the dependent variable with "neg.dev." ("pos.dev.") in this treatment; "own contributions" is the contribution of the punishing participant, "tot.contrib.others" is the total contribution of both other group members, and "round...dummy" is a dummy variable for round "..."; robust standard errors are corrected for possible dependencies of observations within each independent group of 18 participants.

Tobit regression – Allocated punishment points:

		# of obs. = 8424		
		Wald $X^2(19)$ = 1708.25		
Log pseudo-likelihood = -6966.3362		Prob > X^2 = 0.0000		
(standard errors adjusted for clustering on replicated groups)				
dependent. variable:		Robust		
allocated punishment points by punisher	Coef.	Std. Err.	z	P>|z|
T31dummy	-6.7610	0.6720	-10.0600	0.0000
T33dummy	-4.2537	0.5462	-7.7900	0.0000
T11dummy	-2.9684	0.5233	-5.6700	0.0000
T13dummy	-1.3385	0.5216	-2.5700	0.0100
neg.dev*T31dummy	0.5989	0.0631	9.4900	0.0000
neg.dev*T33dummy	0.5107	0.0405	12.6000	0.0000
neg.dev*T11dummy	0.5556	0.0325	17.1000	0.0000
neg.dev*T13dummy	0.5561	0.0303	18.3600	0.0000
pos.dev*T31dummy	-0.1967	0.0622	-3.1600	0.0020
pos.dev*T33dummy	-0.2131	0.0572	-3.7300	0.0000
pos.dev*T11dummy	-0.2405	0.0268	-8.9900	0.0000
pos.dev*T13dummy	-0.1877	0.0360	-5.2100	0.0000
own contribution	-0.2843	0.0226	-12.5900	0.0000
tot.contrib.others	0.0728	0.0111	6.5400	0.0000
round2dummy	-0.1728	0.1840	-0.9400	0.3480
round3dummy	-0.4697	0.2277	-2.0600	0.0390
round4dummy	-0.6206	0.2259	-2.7500	0.0060
round5dummy	-0.5837	0.2697	-2.1600	0.0300
round6dummy	-1.1008	0.2254	-4.8800	0.0000

Logit regression – Likelihood to punish

			# of obs. = 8424	
			Wald $X^2(19)$ = 2895.03	
Log pseudo-likelihood = -3529.5275			Prob > X^2 = 0.0000	
(standard errors adjusted for clustering on replicated groups)				
dependent. variable:		Robust		
punishment (yes = 1)	Coef.	Std. Err.	z	P>|z|
T31dummy	-2.5569	0.2459	-10.4000	0.0000
T33dummy	-1.4091	0.2007	-7.0200	0.0000
T11dummy	-0.9861	0.2041	-4.8300	0.0000
T13dummy	-0.1774	0.2166	-0.8200	0.4130
neg.dev*T31dummy	0.2444	0.0226	10.8000	0.0000
neg.dev*T33dummy	0.2088	0.0171	12.2500	0.0000
neg.dev*T11dummy	0.1982	0.0153	12.9600	0.0000
neg.dev*T13dummy	0.2114	0.0200	10.5400	0.0000
pos.dev*T31dummy	-0.1193	0.0413	-2.8900	0.0040
pos.dev*T33dummy	-0.1125	0.0333	-3.3700	0.0010
pos.dev*T11dummy	-0.1207	0.0188	-6.4100	0.0000
pos.dev*T13dummy	-0.0974	0.0159	-6.1100	0.0000
own contribution	-0.1241	0.0122	-10.1700	0.0000
tot.contrib.others	0.0296	0.0052	5.7000	0.0000
round2dummy	-0.0977	0.0878	-1.1100	0.2660
round3dummy	-0.2621	0.0948	-2.7600	0.0060
round4dummy	-0.3137	0.1010	-3.1100	0.0020
round5dummy	-0.3625	0.1168	-3.1000	0.0020
round6dummy	-0.5949	0.0997	-5.9700	0.0000

Tobit regression – Effect of punishment on punished

		# of obs = 8424		
		Wald $X^2(19)$ = 1135.85		
Log pseudo-likelihood = -8260.3915		Prob > X^2 = 0.0000		
(standard errors adjusted for clustering on replicated groups)				
dependent. variable:		Robust		
received punishment points by punished	Coef.	Std. Err.	z	P>|z|
T31dummy	-14.7099	1.4231	-10.3400	0.0000
T33dummy	-8.6045	1.1331	-7.5900	0.0000
T11dummy	-7.1948	1.1163	-6.4500	0.0000
T13dummy	-1.9414	1.1255	-1.7200	0.0850
neg.dev*T31dummy	1.2315	0.1387	8.8800	0.0000
neg.dev*T33dummy	1.1706	0.1259	9.3000	0.0000
neg.dev*T11dummy	1.0532	0.1022	10.3100	0.0000
neg.dev*T13dummy	1.3543	0.0938	14.4400	0.0000
pos.dev*T31dummy	-0.4347	0.1385	-3.1400	0.0020
pos.dev*T33dummy	-0.4656	0.1319	-3.5300	0.0000
pos.dev*T11dummy	-0.5264	0.0762	-6.9100	0.0000
pos.dev*T13dummy	-0.3944	0.0943	-4.1800	0.0000
own contribution	-0.6372	0.0844	-7.5500	0.0000
tot.contrib.others	0.1570	0.0270	5.8100	0.0000
round2dummy	-0.3646	0.4059	-0.9000	0.3690
round3dummy	-1.0002	0.4922	-2.0300	0.0420
round4dummy	-1.4835	0.5151	-2.8800	0.0040
round5dummy	-1.4087	0.6121	-2.3000	0.0210
round6dummy	-2.4772	0.4705	-5.2600	0.0000