Punishment in public-goods games in Japan

Yukihiko Funaki[a], Simon Gächter[d], Masao Ogaki[b], Fumio Ohtake[c], Róbert F. Veszteg[a,*]

[a] *Waseda University, School of Political Science and Economics, 1-6-1 Nishiwaseda, Shinjuku-ku, 169-8050 Tokyo, Japan*
[b] *Keio University, Department of Economics, 612 Mita-Kenkyu-shitsu, 2-15-45 Mita Minato-ku, 108-8345 Tokyo, Japan*
[c] *Osaka University, Institute of Social and Economic Research, 6-1 Mihogaoka, Ibaraki, 567-0047 Osaka, Japan*
[d] *The University of Nottingham, School of Economics, Room B54 Sir Clive Granger Building, University Park, NG7 2RD Nottingham, UK*

Abstract

Although decentralised punishment has been shown to induce high contribution levels in public-goods games in numerous countries (Gächter et al, 2010), it fails to do so in Japan - both in Osaka and Tokyo. These results are puzzling given the general image of Japanese society as being cooperative, having strong social norms of cooperation, and inflicting harsh punishment on deviators. Based on previous insights into cooperation in Japan by Yamagishi (1988), we investigate in the experimental laboratory how Japanese use punishment opportunities in a public-goods setting. We find that, although antisocial punishment (Herrmann et al., 2008) is rather infrequent in Japan, the lowest contributors manage to hijack the provided (unconstrained) decentralised-punishment institution to fight the highest contributors. This perverse punishment activity (Erhan et al., 2008) is what keeps cooperation low in Japan. We show that this coordination failure can be solved when punishment is executed in a constrained way: collectively and/or by pre-specifying a target for punishment (e.g., the lowest contributor).

Keywords: antisocial punishment, perverse punishment, free-riding, Japan,

*Corresponding author
Email addresses: funaki@waseda.jp (Yukihiko Funaki), simon.gaechter@nottingham.ac.uk (Simon Gächter), mogaki@econ.keio.ac.jp (Masao Ogaki), ohtake@iser.osaka-u.ac.jp (Fumio Ohtake), rveszteg@waseda.jp (Róbert F. Veszteg)

1. Introduction

Economists and biologist call *strong reciprocity* the willingness to punish wrong-doers and to reward right-doers even when such actions are costly and do not bring direct benefits to those who carry them out (Guala, 2012). Although the standard game-theoretical models of cooperation are unable to rationalise strong reciprocity, experimental economists have gathered a large amount of data during the past decade which suggest that strong reciprocators do show up in the experimental laboratory. The workhorse of this research is the linear public-goods game complemented with a punishment stage in which participants are allowed to assign *deduction points* to each other in a decentralised and completely anonymous way. This game is the straightforward generalisation of the prisoners' dilemma to accommodate more than two actors and to allow punishment among them. In its unique Nash equilibrium in dominant strategies no one cooperates (i.e., no one contributes to the public good) and no one punishes even if that leads to a suboptimal (inefficient) outcome.

In spite of the above prediction, early research (e.g., Fehr and Gächter, 2000, 2002) has shown that decentralised punishment can sustain a high level of cooperation in the experimental laboratory and that it is precisely strong reciprocity that accounts for any punishment activity that targets free riders. Follow-up research, on the one hand, has replicated those findings and delivered further supporting evidence from different subject pools and experimental setups. On the other hand, it has also refined previous results and explored the limits of *strong reciprocity*. For example, Casari (2005) argues that the available punishment technology - in particular its effectiveness measured by the fine-to-fee ratio - affects the frequency with which punishment is carried out. Also, Nikiforakis and Normann (2007) find that "high effectiveness leads to near complete cooperation and welfare improvement".

Understandably, the amount of information that participants receive after the contribution stage is important, too. Without detailed feedback on individual contribution levels it would be simply impossible to decide whom to punish. Too much information however, for example on individual earnings, could be harmful, because participants might use it as a coordination device to establish a new contribution standard which typically is lower than the initial contribution level (Nikiforakis, 2010). In other words, while the possibility of decentralised punishment can be used by strong reciprocators to reduce the attractiveness of free-riding and to increase welfare for the group, it can

also be hijacked by low contributors who wish to fight back (or even carry out preemptive attacks). The literature refers to the latter phenomenon as *antisocial punishment* which shows up frequently in the experimental data and can even dominate (Hermann et al., 2008). Similarly, information on the punisher's identity can lead to costly feuds and lower contribution levels (Nikiforakis and Engelmann, 2011).

From a different, evolutionary and cultural perspective, punishment triggered by strong reciprocity can be weak or might even fail to exist. Hermann et al. (2008) and Gächter et al. (2010) report important cross-country differences in the success of decentralised punishment in sustaining cooperation from 16 subject pools around the world. They not only leave it clear that experimental subject pools are heterogeneous on individual, group, and also cultural level, but also try to explore what makes one different from the other. Hermann et al. (2008) find that it is precisely antisocial punishment that explains variations in cooperation across locations, and that in turn it is significantly correlated with the weakness of the *rule of law* and the weakness of *norm of civic cooperation* which characterise the country where the experimental data was collected.[1] In conclusion, decentralised punishment is unable to solve coordination failures (or social dilemmas) in the absence of a strong social norm of cooperation.

In this paper, we report results from a series of experimental sessions implemented in Japan on the linear public-goods game which appears in Fehr and Gächter (2000) and many follow-up papers.

Stereotypically, Japan is a highly civilised country with strong social norms that promote cooperation. In terms of the above-mentioned variables of rule of law and of norm of civic cooperation, Japan closest neighbours are Germany and Korea in the former, and Australia, Denmark, Germany, the UK and the USA in the latter.[2] Decentralised punishment seems to succeed in all of these neighbour countries in sustaining cooperation. They constitute six of the eight best-performing locations in terms of average contribution levels (in the decentralised-punishment treatment reported by Hermann et al.,2008). It is noteworthy that Germany appears among Japan's five closest neighbours according to any of the groups of variables considered by Hermann et al. (2008), i.e., social capital, economic prosperity, law enforcement and democracy, cultural dimension, and value orientation. A quick look at the Inglehart-Welzel cul-

[1]The proxy variable for *rule of law* is based on data from the World Bank, and the one for *norm-of-civic-cooperation* was created by averaging responses to a number of questions in the World Value Survey. For more details refer to the online supporting material for Hermann et al. (2008).

[2]We used observations from the same dataset to locate the closest neighbours, which are at most half a standard deviation away from Japan. Japan scores 8.30 in norms of civil cooperation, and 1.39 in rule of law. For the other country scores refer to the online supporting material for Hermann et al. (2008).

tural map (Inglehart and Welzel, 2014) suffices to see that Japan is an outlier of its Confucian group of countries, and again Germany is its closest neighbour. If Japanese subjects behaved as the Germans, according to Hermann et al. (2008) we should observe a stable and substantial contribution level in the no-punishment treatment (with an average of 9.2 tokens out of 20, and without the usual negative time trend), and a considerably higher contribution level (with an average of 14.5) in the decentralised-punishment treatment with a positive time trend sustained by punishment targeting free-riders (with the mean punishment expenditure being around 3.5 tokens). Also, we should not expect to observe much of antisocial punishment in Japan (the mean expenditure should be well below 1 token).

The data that we collected in Osaka and Tokyo on decisions made by 264 participants show a completely different picture, far from the above-detailed expectations. Contributions levels are not only low and decrease quickly in the no-punishment treatment, but decentralised punishment is unable to sustain cooperation in Japan. Participants spend considerably less (than expected) on punishing free-riders and also considerably less on antisocial punishment.[3] At the same time, our observations depict an intense fight between the highest and lowest contributors, from which the latter come out victorious. It seems, that it is not only antisocial, but *perverse punishment* that shows up in our records from Japan, and causes decentralised punishment to fail. The term *perverse punishment* was introduced by Erhan et al. (2008) to describe punishment of high contributors.[4] They report results from a similar public-goods game from the USA (Brown University) and show that, when allowed to vote, no groups of participants allowed the punishment of high contributors, and groups which only allowed punishment of low contributors reached levels of cooperation unmatched in the literature.

While participants could not choose or even fine-tune the punishment technology in our experiments, our constrained-punishment treatment is similar to the one by Erhan et al. (2008) as it only allows the punishment of the lowest contributors. It was directly inspired by Yamagishi (1988) who used centralised punishment with the help of a *pun-*

[3]The above statement is based on predictions on punishment of free-riders and antisocial punishment from the regression analysis presented by Hermann et al. (2008) that simultaneously includes the strength of norms of civic cooperation and the strength of the rule of law as regressors among other behavioural and demographic variables. We used the reported coefficient estimates to compute out-of-sample predictions for Japan. According to the two variables, Japan should appear between Saudi Arabia and the USA in the ranking of punishing free-riders, and appear between China and Turkey when it comes to antisocial punishment. However, our observations rank it below the predicted position for both punishment categories.

[4]*Antisocial punishment* refers to the punishment of any non-negative deviation with respect to one's own contribution level.

ishment fund to which participants could contribute voluntarily and which would only target the lowest contributors. We show that by preventing perverse punishment, participants in our experiments not only contribute substantially more, but the constrained-punishment treatments induce a remarkably high efficiency level of almost 83%.

2. Experimental design

We gathered data from 264 participants across 15 experimental sessions that took place between January 2012 and May 2013 at Waseda University (Tokyo, Japan) and Osaka University (Osaka, Japan). Each session was composed of two treatments programmed in z-Tree (Fischbacher, 2007), and lasted about 90 minutes with an average pay (including a ¥500 show-up fee) of ¥1960. The experimental procedures were identical to those in Fehr and Gächter (2002): participants would read the instructions individually and answer a list of test questions to show their understanding of the rules of the interaction.[5] In fixed groups of four, they would play 10 rounds of a linear public-goods game followed by 10 rounds (in newly created but fixed groups) of the same game complemented by a punishment stage. Participants were not allowed to communicate with each other, and their interaction was completely anonymous. After the last round they would respond to a questionnaire covering demographics, culture, and other personal matters including attitudes towards various social issues (e.g., trust, power, social norms). Table 1 offers a brief summary of our sessions.

Table 1: Session summary

TREATMENT	PARTICIPANTS	DATE	UNIVERSITY
N+P	68	January 2012	Waseda University
N+P	20	February 2013	Osaka University
N+L	80	January, February, May 2013	Waseda University
N+L	96	February, March 2013	Osaka University

Treatments N (no-punishment treatment) and P (decentralised-punishment treatment) follow Fehr and Gächter (2002) in implementing a linear public-goods game in which each participant has to decide privately how to allocate an initial endowment of 20 tokens between a personal and a public account. The personal account pays no interest in the game, but it is safe as the monetary value of the tokens deposited there does not depend on other participants' decisions. The public account pays 0.4 times

[5]The instructions were translated to Japanese. The *original* English version is available for download as part of the supporting online material for Herrmann et al. (2008).

the total contribution by all four members of the group. While this rule is fixed, the income generated by the public account is uncertain as it depends on other participants' decisions. Thus, any participant i's total income can be written mathematically as

$$\pi_i(c_i, c_{-i}) = (20 - c_i) + 0.4 \sum_{j=1}^{4} c_j,$$

where $c_i \in [0; 20]$ denotes participant i's contribution to the public account, and c_{-i} stands for everybody else's choice. Given the above specification, the public-goods game has a unique Nash equilibrium in dominant strategies which is inefficient as no one contributes to the public account. Note that the above payoff function is strictly decreasing in participant i's contribution to the public account: $\frac{\partial \pi_i}{\partial c_i} = -0.6$. In other words, individually each participant can maximise her income by not contributing anything to the public project (i.e., the equilibrium prediction is $c_i = 0$ for all i) independently of the others' decision. Nevertheless, the group would do best if each member contributed her entire endowment of 20 tokens. To see this, consider the aggregated income for the entire group $\sum_i \pi_i = 4 \cdot 20 - \sum_{i=1}^{4} c_i + 4 \cdot 0.4 \sum_{i=1}^{4} c_i = 120 + 0.6 \sum_{i=1}^{4} c_i$ which is a strictly increasing function of the aggregate contribution, i.e. $\sum_{i=1}^{4} c_i$.

After each round in the experiment, participants would receive information about personal gains and detailed, yet anonymous, information about others' contribution to the public account.

In treatment P, participants are allowed to assign so-called *deduction points* to each other in an anonymous, unconstrained and fully decentralised way. Each deduction point reduces the sender's income by 1 token and the receiver's income by 3 tokens. After the punishment stage, participants are informed about the total number of reduction points they received, but do not find disaggregated information about their origin. Also, while the receiver's income is bounded by 0 from below, the sender's income is allowed to turn negative after accounting for the costs of assigning deduction points. Note that the introduction of the second, punishment stage does not alter the theoretical properties and the equilibrium outcome of the game. Any participant i's total income can now be written mathematically as

$$\pi_i(c_i, d_{ij}, c_{-i}, d_{-ii}) = \max \left\{ 0; (20 - c_i) + 0.4 \sum_{j=1}^{4} c_j - 3 \sum_{j=1}^{4} d_{ji} \right\} - \sum_{j=1}^{4} d_{ij},$$

where d_{ij} denotes the number of deduction points assigned by participant i to partici-

pant j, and d_{-ii} denotes the number of deduction points assigned by other participant to participant i. Note that the inclusion of the punishment stage does not alter the game-theoretical incentives. It essentially introduces a secondary public-goods game, therefore its Nash equilibrium in dominant strategies is such that participants contribute nothing to the public account and refrain from assigning deduction points to anyone ($c_i = 0$ and $d_{ij} = 0$ for all i and j).

Treatment L (lowest-contributor-punishment treatment) was inspired by Yamagishi (1988). It includes a punishment stage identical to the one in treatment P except that deduction points can only be assigned to the participant(s) who contributed the least to the public account. The following rules impose equal treatment in case more than one participant fall in the category of *lowest contributor*.[6]

1. The lowest contributor(s) are not allowed to assign deduction points.
2. If a participant decides to assign deduction points, she must assign the same number of deduction points to each of the lowest contributors.

Note that the restrictions imposed in the punishment stage of treatment L (as compared to treatment P) do not change the game-theoretical equilibrium. Participants have strong incentives to free-ride and not to assign deduction points.

3. Experimental results

In spite of its game-theoretical properties, the punishment stage in treatment P has been shown to be able to sustain cooperation in the underlying public-goods game in the experimental laboratory across various cities (Gächter et al., 2010). Its success relies on strong reciprocity as discussed in the introduction.

As for our subject pool, independently whether one considers the culturally hard-to-categorise metropolises of Japan as members of the *Confucian group* - as it is done on the Inglehart-Welzel cultural map of the world (Inglehart and Welzel, 2014) - or virtually part of *Protestant Europe* - based on their cultural distance from that group's members -, one would expect that punishment is able to solve the public-goods problem both in Osaka and in Tokyo (Herrmann et al., 2008; Gächter et al., 2010).

However, our experimental data summarised in figure 1 reveal that it is not the case.

[6] We implemented two slightly different versions of treatment L. In treatment L_C deduction points were assigned in a decentralised way (like in treatment P), while punishment in treatment L_D was centralised. In treatment L_D, participants could contribute to a *deduction fund* which then took care of assigning the deduction points according to the above rules to the lowest contributors (to the public account). Theoretically, logically and also experimentally the two methods are identical. For this reason we have pooled the data for the sake of analysis in the paper.

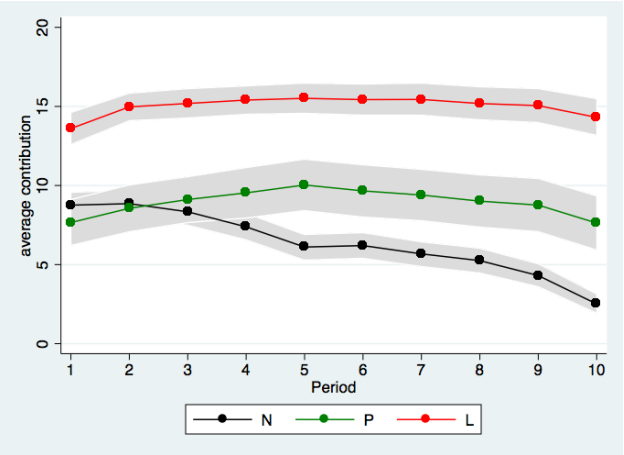

Figure 1: Evolution of individual contribution levels. N: no-punishment treatment; P: decentralised-punishment treatment; L: constrained, lowest-contributor-punishment treatment. The grey area shows the 95% confidence interval around the point estimates.

While the black series show the *usual* pattern in the benchmark treatment N (i.e., participants contribute roughly 40% of their endowments in the first round, and contribution levels decrease significantly over time), the green series representing average contribution levels from treatment P are surprisingly flat. The two curves are statistically identical in the first four rounds after which contributions levels stay essentially constant in treatment P and decrease significantly in treatment N.[7] The failure of decentralised punishment in sustaining cooperation is even more apparent when income - instead of contribution - levels are considered. Arguably, income is a better measure for the performance of decentralised punishment as it takes into consideration the cost of punishment, and when compared to the ideal income of 32 tokens (achieved when everybody contributes her entire endowment to the public account) it reflects efficiency.

Figure 2 shows that, *in the end*, decentralised punishment is unable to solve the coordination failure caused by free-riding in the laboratory (with Japanese participants). The curve that represents average individual income in treatment P starts at a significantly lower level than the one from treatment N due to relatively intense punishment activity during the initial rounds. Although individual income (and with it, overall efficiency) is slowly increasing from round to round in punishment treatment P, it is unable to exceed - in a statistically significant way - income in the corresponding pe-

[7] All statements in this sections are based on results of parametric hypothesis tests and are statistically significant at (at least) 5% significance level, unless stated otherwise.

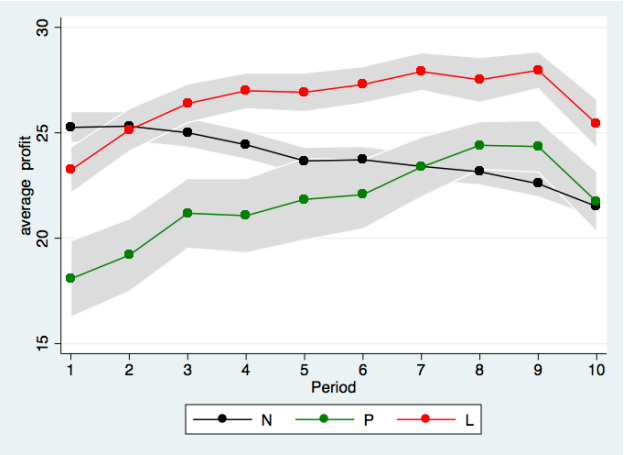

Figure 2: Evolution of individual income levels. N: no-punishment treatment; P: decentralised-punishment treatment; L: lowest-contributor-punishment treatment. The grey area shows the 95% confidence interval around the point estimates.

riods from treatment N. The overall efficiency levels are 74.4% in treatment N and 67.9% in treatment P (*p-value* $= 0.0000$).

However, when punishment is only allowed to target the lowest contributors (treatment L), both contribution and income levels grow and stay above the ones from the other two treatments. The difference is not only statistically significant, but it is also large. In period 9, the average contribution is 15.1 in treatment P and 8.8 in treatment L, while the average incomes are 28.0 and 24.3 in treatments L and P, respectively.[8] Overall efficiency in treatment L is of 82.7%. Results 1 summarises the above findings.

Result 1. *(Unconstrained) decentralised punishment is unable to sustain cooperation and to solve the free-riding problem in the laboratory in Japan, while constrained decentralised punishment which only allows for punishing the lowest contributors is able to do so.*

In what follows, we argue that the failure of (unconstrained) decentralised punishment is the result of a battle between the lowest and highest contributors in which the former win. It seems that constraints on punishment (as introduced in treatment L) do not significantly alter the average - and for that matter, the total - amount of tokens that participants dedicate to punishment, but they do weaken the position of the lowest

[8] Although the time series in figures 1 and 2 show important end-game effects, the differences remain between the two treatments.

contributors enough so that the group is able to overcome the coordination failure.

Overall, participants do not seem to be afraid of using punishment in treatments P or L. On average they spend significantly more on punishment (0.95 tokens) in treatment P than in treatment L (0.65 tokens). However, if we look at the differences period by period, those turn out to be statistically insignificant and essentially non-existent, especially in the last four rounds (figure 3).

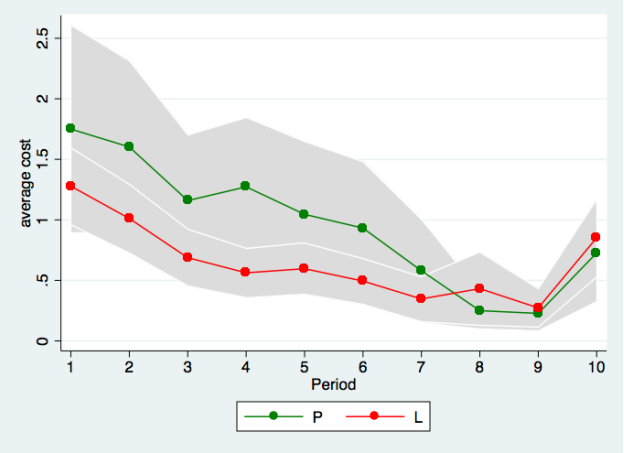

Figure 3: Evolution of the number of assigned deduction points. P: decentralised-punishment treatment; L: lowest-contributor-punishment treatment. The grey area shows the 95% confidence interval around the point estimates.

In light of the data collected by Herrmann et al. (2008), Japan's closest neighbour in terms of punishment would be Switzerland (St. Gallen, Zurich), Germany (Bonn) and China (Chengdu) where decentralised punishment has been show to succeed in sustaining cooperation at best. Herrmann et al. (2008) report the average punishment expenditure for five different categories according to how much the punished participant deviated from the punisher in terms of contribution. They include non-punishers in the averages which therefore are to be interpreted as expected punishment reflecting both the probability and the intensity of punishment.[9]

Using the data from Herrmann et al. (2008), we have estimated a Heckman selection model to explain punishment behavior and create out-of-sample predictions for

[9]In Japan, the average expected punishment is of 1.88 points for deviations between -20 and -11, 0.55 points for deviations between -10 and -1, 0.05 points for 0 deviation, 0.18 points for deviations between 1 and 10, and 0.28 points for deviations between 11 and 20.

Japan.[10] The results, shown in table A.3, offer a more nuanced view of punishment activity separating the decision on participation (whether to punish at all) from the decision on intensity (how many deduction points to assign). They suggest that the strength of norms of civic cooperation is negatively related to the intensity of punishment that participants inflict on each other, independently of whether the punishment is targeting free riders or is antisocial. The rule of law has a similar negative impact on punishment intensity, but it is smaller in absolute value and only matters in punishing free riders. It is in the participation decision where the two variables play remarkably different roles. They both have positive and significant (although the rule of law a much smaller) effect on participants' willingness to punish free riders, while that effect is negative and significant (and equally important) when it comes to antisocial punishment. These finding do not only disentangle the effect of social norms on punishment, but also deliver a means to create a prediction for Japan.[11] Our model typically overestimates the antisocial punishment frequencies, but does a fairly accurate job for Japan: all the specifications predict that Japanese participants would use 11% of all antisocial punishment opportunities and we observed that 5% did so. In terms of punishment frequencies of free riders Japan turns out to be a clear outlier: the model predicts 7%, but we observed 34%. In this respect Japan is not remarkably different from the other locations that Herrmann et al. (2008) analysed, even is based on the strength of norms and the rule of law we would have expected otherwise.

When it comes to punishment intensity, Japan is a true outlier again. Based on our model, Japanese participants should have excelled both in terms of antisocial punishment (above all other locations) and the punishment of free riders (in third position among all locations), but in reality Japan is least harsh location in terms of any kind of punishment among all locations.

In Japan, when punishment happens, the average punishment targeting free-riders (i.e., participants who contributed less than the punisher) cost 2.4 tokens and the antisocial punishment (i.e., punishment targeting participants who contributed at least as much as the punisher) cost 2.3 tokens.

49.6% of the all punishment that we recorded in treatment P in Japan happens be-

[10] Although Herrmann et al. (2008) present coefficient estimates from tobit regressions to support their main findings, we find their approach inadequate for our purposes. Given that they code the frequent *no punishment* as assigning 0 deduction points, and at the same time use 0 as a lower censoring limit in the estimation process, their regressions deliver overwhelmingly negative estimates which have no practical meaning in the underlying game.

[11] We have estimated predictions for each decision our participants faced in the Japanese experimental sessions. In other words, we took into consideration their real demographic characteristics, contribution levels, etc.

tween the lowest and the highest contributors. 40.1% of all the punishment (without controlling for its intensity) is inflicted by the highest contributors on the lowest contributors in the group. 9.5% goes in the opposite direction. We find similar proportions when restricting our attention to the first five periods in which 68.1% of the punishment take place.[12]

The first graph on the left in figure 4 depicts the evolutions of punishment frequency between the two prominent groups. We find that the highest contributors (blue triangles) face increasing resistance from the lowest contributors (red squares) in the first five periods. The lowest contributors do not only increase the frequency of punishment, but they assign large (in periods 1 and 3 the largest) number of deduction points to the highest contributors (second graph in figure 4). As a consequence, the highest contributors' income - when compared to the lowest contributors' - is relatively low due to free-riding and also to antisocial (and also perverse) punishment. Perverse punishment occurs even in the first period and on average it happens with a very large intensity. Note that this observation suggests that the lowest contributors expect to be punished by the highest contributors and decide to launch a preemptive strike, given that they must have assigned the deduction points to others before leaning how many (if at all) they had received.[13] Period five seems to be the turning point. That is the moment in which the lowest contributors win their battle and start reducing their punishment activity against the highest contributors. Period five also marks the maximum of the - otherwise flat - average-contribution series in figure 1. The third graph in figure 4 shows that the highest contributors reduce their contribution level sharply over the 10 rounds, while the lowest contributors seem to increase it a little during the first six rounds and reduce it to zero at the end. The latter time series are rather flat if the first and the last rounds are ignored.

The highest contributors reduce both the frequency and the intensity of their punishment activity (towards the lowest contributors) after period five. They only increase them again in the final round in which the lowest contributors do not assign any deduction points to the highest contributors (figure 4). Our interpretation of this pattern, i.e. of punishment that by definition can not have any impact on the target's future behavior, is that the highest contributors are strong reciprocators who are willing to sacrifice part of their income in order to discipline the lowest contributors - who seem

[12]The reported results have been computed by excluding groups in which everybody contributed the exact same amount to the public account, because otherwise the categories of lowest and highest contributors would overlap. Our conclusions do not change qualitatively when we include all observations.

[13]It could also be that the lowest contributors punish the highest contributors for the latter showing a mirror in which the lowest contributors appear in a bad light.

to be deviating from the social norm of cooperation. However, they are discouraged by the strategic punishment that lowest contributors inflict on them in the battle that the highest contributors end up losing (by period five).

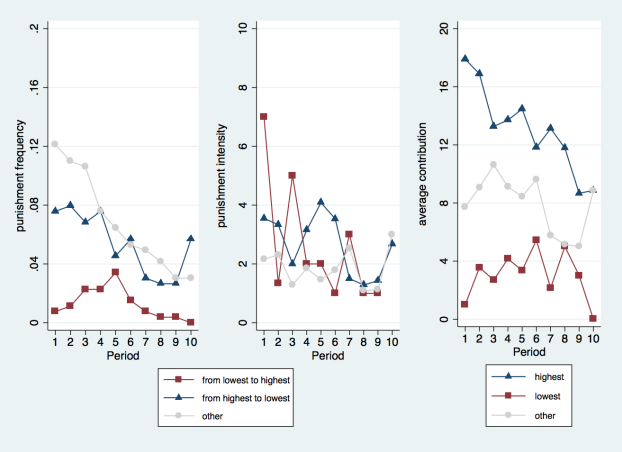

Figure 4: Evolution of punishment frequency and intensity, and the average contribution per punisher category in treatment P. "from lowest to highest": punishment directed from the lowest contributors toward the highest contributors; 'from highest to lowest": punishment directed from the highest contributors toward the lowest contributors; "highest": contribution by the highest contributors; "lowest": contribution by the lowest contributors; "punishment intensity" is measured with the number of deduction points assigned.

The results of the regression analysis that we report in table 2 deliver additional support to our *theory* summarised in result 2.

Result 2. *(Unconstrained) decentralised punishment fails to sustain cooperation in Japan, because the lowest contributors use it to fight the highest contributors who in turn are unable to enforce the social norm of high contribution.*

In order to provide further statistical support, we have estimated regressions of punishment and contribution levels for the groups of lowest and highest contributors separately. Note that the composition of these group changes from round to round, depending on the participants' contribution level as compared to their opponents'. The upper half of the table shows logit results for punishment, while the lower half shows censored tobit results for contribution.[14] Participants seem to have a predisposition

[14]The regressions on punishment concentrate on punishment frequency and ignore punishment intensity. We have performed the same statistical analysis with the number-of-assigned-deduction-points as a dependent variable and found that the regressions lose substantial explanatory power (as measured by the pseudo R^2 statistics. Also, we believe that the presented results are in line with our interpretation of the observed pattern detailed above (refer to figure 4).

both in terms of contributing and punishing behavior, given that the lagged value of the dependent variables turn out to be a highly significant and important explanatory variable in all four regressions. More importantly, the punishment received in the previous round (highlighted in blue in the table) has an important and significant positive impact on the lowest contributors' punishing behavior without significantly affecting their contribution levels.

The regressions include two explanatory variables related to received punishment. PUNISHED (LAG1) is a binary variable that takes value 1 when punishment was received in the previous round. PUNISHMENT RECEIVED (LAG1) measure the intensity of that punishment. The coefficients of the two variables separate the constant base impact of punishment from the graduate impact of its intensity. The odds of punishing, i.e. assigning deduction points to, highest contributors are on average 15% higher for each unit of punishment that a lowest contributor received in the previous period (*p-value* $= 0.099$)

As for the highest contributors, they seem to reduce their punishment activity gradually over time (there is a significant time trend in their behavior) without conditioning punishment on whether they received punishment in the previous period. The punishment that they receive does have however a significant and important negative impact on their contribution level. After punishment, it decreases by roughly 3 tokens on average independently on the intensity of the punishment.

As an attempt to open the black-box of cultural differences, we ran a logit regression to explain whether the participant belong to the group of lowest or highest contributors. Among the regressors we have variables decoding to answers to an extensive list of questions in the post-experimental questionnaire. Tables A.4 and A.5 report the estimation results that we are briefly going to discuss. Note that the groups of lowest and highest contributors are not fixed as participants might drop out from, switch between, or newly appear in them in each period. Our objective here is simply to check which personal characteristics (demographic characteristics, attitudes in the game, attitudes in the society, etc.) correlate with the *choice* of one's contribution and the category of one's contribution as compared to others'.

The estimation results suggest that those who try to maximise group earnings, do not expect others to act in the same way, and who believe that others have a right to punish, but that high contributors are not going to be punished are significantly and substantially more likely to find themselves among the highest contributors. Given the structure of our logit regression analyse, the opposite statement hold for the group of lowest contributors.

Table 2: The effect of punishment on contribution and future punishment by lowest any highest contributors (treatment P).

	RECEIVER'S CATEGORY IN PREVIOUS ROUND	
	LOWEST CONTRIBUTOR	HIGHEST CONTRIBUTOR
	PUNISH ...	
	... HIGHEST CONTRIBUTOR	... LOWEST CONTRIBUTOR
PERIOD	-0.0538	-0.2330***
CONTRIBUTION	-0.2083**	0.1712***
CONTRIBUTION (LAG1)	0.0937	-0.1195**
CONTRIBUTION (TOTAL)	0.0416	-0.0395***
PUNISH (LAG1)	2.2924***	2.1115***
PUNISHED (LAG1)	0.5914	-0.4020
PUNISHMENT RECEIVED (LAG1)	0.1422*	0.2745
CONSTANT	-4.4690***	0.9970
PSEUDO R^2	0.3259	0.2542
	CONTRIBUTION	
PERIOD	-0.4550***	-0.2814
CONTRIBUTION (LAG1)	1.2625***	0.7648***
CONTRIBUTION (TOTAL, LAG1)	-0.0018	0.2142***
PUNISHED (LAG1)	0.3298	-2.9904**
PUNISHMENT RECEIVED (LAG1)	0.0646	0.2758
CONSTANT	1.7580	-2.3693
PSEUDO R^2	0.1995	0.1904
OBSERVATIONS	211	213

NOTE: Logit regression results for punishment and tobit results (censored by 0 from below and 20 from above) for contribution. Coefficient significantly different from zero at *10%, **5%, ***1% significance level. PUNISH: 1 if the participant assigned any deduction points, 0 otherwise. PUNISHED: 1 if the participant received any deduction points, 0 otherwise. PUNISHMENT RECEIVED: intensity of the punishment received (tokens lost).

4. Discussion

"The nail that sticks out gets hammered down" goes the saying in Japan. It is one of the main pillars of the still remarkably homogeneous Japanese society by making deviation from the norm very costly. On the positive side, it keeps the megalopolises in the Kanto and Kansai area clean, safe and operational, as citizens are forced to follow the uncountable social norms that characterise the Japanese society. On the negative side, it introduces a massive *status-quo* bias as it proscribes any spontaneous change independently from whether it would benefit the society or not. In other words, it does not matter whether one would want to move society out of an inefficient equilibrium or move it into one, it is hard to be a leader in Japan in any case.

The relatively low cooperation levels observed in experimental laboratories around Japan are not a new phenomenon. Yamagishi (1988) found similar evidence in a comparison to American subjects and argue that it is precisely the fact that Japanese society relies on strong mutual monitoring and sanctioning systems that explains the results. It seems that the availability of those systems make the population short-sighted in that it never questions the purpose of the system. It simply makes sure that the system is sustained and its rules are enforced and followed. When the system suddenly disappears, or it is removed just like in our treatment N, people stop following the rules, act in a more self-regarding manner and contribute less to the public good. One could also say that social institutions - in this case, outside the laboratory - influence preferences that in turn motivate behavior inside the laboratory.[15]

The failure of unconstrained decentralised punishment (treatment P) in Japan could be explained in a similar fashion. Because its decentralised structure and rules are alien to Japanese, it can not remediate the coordination failure among self-regarding participants who manage to hijack it through perverse punishment. Constrained decentralised punishment, on the other hand, is successful. Not necessarily because it resembles the real-life sanctioning system, but because it does not allow perverse (or antisocial) punishment to happen.

[1] Casari, Marco (2005) "On the design of peer punishment experiments", *Experimental Economics* 8: 107-115.

[2] Ertan, Arhan, Talbot Page, Louis Putterman (2009) "Who to punish? Individ-

[15] Rodriguez-Sickert et al. (2008) show experimental evidence from a common-pool resource game that shows how the choice of a sanctioning institution is able to rewire participants' preferences and change the proportion of selfish, conditionally and unconditionally cooperative people in the subject pool.

ual decisions and majority rule in mitigating the free rider problem", *European Economic Review* 53: 495-511.

[3] Fehr, Ernst, Simon Gächter (2000) "Cooperation and punishment in public goods experiments", *American Economic Review* 90(4): 980-994.

[4] Fehr, Ernst, Simon Gächter (2002) "Altruistic punishment in humans", *Nature* 415: 137-140.

[5] Fischbacher, Urs (2007) "z-Tree: Zurich Toolbox for Ready-made Economic Experiment", *Experimental Economics* 10(2): 171-178.

[6] Gächter, Simon, Benedikt Herrmann, Christian Thöni (2010) "Culture and cooperation", *Philosophical Transactions of The Royal Society B* 365: 2651-2661.

[7] Guala, Franceso (2012) "Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate", *Behavioral and Brain Sciences* 35: 1-59.

[8] Henrich, Joseph, Steven J. Heine, Ara Norenzayan (2010) "The weirdest people in the world?", *Behavioral and Brain Sciences* 33: 61-135.

[9] Herrmann, Benedikt, Christian Thöni, Simon Gächter (2008) "Antisocial punishment across societies", *Science* 319: 1362-1367.

[10] Inglehart, Ronald, Christian Welzel (2014) "Inglehart–Welzel Cultural Map", World Values Survey, Retrieved 16-05-2014

[11] Nikiforakis, Nikos (2010) "Feedback, punishment and cooperation in public good experiments", *Games and Economic Behavior* 68: 689-702.

[12] Nikiforakis, Nikos, Dirk Engelmann (2011) "Altruistic punishment and the threat of feuds", *Hournal of Economic Behavior & Organization* 78: 319-332.

[13] Rodriguez-Sickert, Carlos, Ricardo Andrés Guzmán, Juan Camilo Cárdenas (2008) "Institutions influence preferences: Evidence from a common pool resource experiment", *Journal of Economic Behavior & Organization* 67: 215–227.

[14] Yamagishi, Toshio (1988) "Seriousness of social dilemmas and the provision of a sanctioning system", *Social Psychology Quarterly* 51(1): 32-42.

AppendixA. Additional statistical results

Table A.3: Punishment behavior (separating punishment frequency and intensity)

	PUNISHMENT OF FREE RIDING (NEGATIVE DEVIATIONS)			ANTISOCIAL PUNISHMENT (NONNEGATIVE DEVIATIONS)		
PUNISHMENT INTENSITY						
NORMS OF CIVIC COOP.	-0.2337***	-	-0.1860***	-0.4180***	-	-0.4379***
RULE OF LAW	-	-0.1133***	-0.0861**	-	-0.0333	0.0365
CONSTANT	4.9295***	3.0619***	4.4398***	4.9302***	1.7952***	5.1181***
PUNISHMENT FREQUENCY						
NORMS OF CIVIC COOP.	0.1635***	-	0.1466***	-0.1703***	-	-0.1038***
RULE OF LAW	-	0.0662***	0.0298*	-	-0.1096***	-0.1191***
CONSTANT	-1.9211***	-0.5606***	-1.7444***	0.9263***	-0.4069***	0.2171
CONTROLS	YES	YES	YES	YES	YES	YES
P-VALUE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
OBSERVATIONS	8347	8947	8347	19770	20580	19770

NOTE: Heckman selection model results. Coefficient significantly different from zero at *10%, **5%, ***1% significance level. Data from Hermann et al. (2008). Controls identical to those in Hermann et al. (2008) in both equations. The selection equation also controls for the participant's own contribution and for others' contribution.

Table A.4: What makes one to be among the lowest or the highest contributors? (Part 1)

CONTRIBUTION CATEGORY (0 - LOWEST; 1 - HIGHEST)	
PERIOD	1.0291
FEMALE	5.4000**
AGE	0.9267
SIBLINGS	2.5702**
STUDY (BASE: HUMANITIES)	
NATURAL SC.	0.0140**
ENGENEERING	0.0063*
MEDICINE	0.0283**
ECONOMICS	0.2173
BUSINESS	1554.3710***
POLITICAL SC.	0.3361
LAW	21.1194***
OTHER SOCIAL SC.	3.6973
CITYSIZE (BASE: $< 2{,}000$)	
100,000 - 1 MILLION	8.9496
1 MILLION +	15.2790
LIVE WITH OTHERS	1.1837***
MONTHLY EXPENDITURE	1.0000
CLUB MEMBER	3.1084
ATTITUDE IN THE GAME (1/AGREE - 5/DON'T AGREE)	
"MAXIMISE GROUP EARNINGS"	0.5715*
"MAXIMISE PERSONAL EARNINGS"	1.6785
"FEEL EXPLOITED IF OTHERS CONTRIBUTE LESS"	1.9653
"MATCH OTHERS' CONTRIBUTION LEVEL IN TRT. N"	0.4982**
"MATCH OTHERS' CONTRIBUTION LEVEL IN TRT. P"	1.2933
"OTHERS MAXIMISE PERSONAL EARNINGS"	0.5517
"OTHERS MAXIMISE GROUP EARNINGS"	2.7131***
"CONTRIBUTE IF OTHERS DO"	0.8511
"GOOD TO HAVE A PUNISHMENT OPPORTUNITY"	0.7547
"HAVE NO RIGHT TO PUNISH OTHERS"	0.9262
"AVOID PUNISHMENT"	0.8749
"IGNORED PUNISHMENT STAGE WHEN CONTRIBUTING"	0.5371*
"LOW CONTRIBUTORS WILL GET PUNISHED"	0.7225
"PEOPLE SHOULD ALWAYS CONTRIBUTE"	0.8939
"OTHERS HAVE NO RIGHT TO PUNISH ME"	0.3846**
"YOUR FAULT IF EXPLOITED"	0.7899
"HIGH CONTRIBUTORS GET PUNISHED"	6.9815***
RELIGIOUS	0.5525**
POLITICALLY ON THE RIGHT	0.7685
...	...

NOTE: Logit regression results (odds ratios). Coefficient significantly different from zero at *10%, **5%, ***1% significance level.

Table A.5: What makes one to be among the lowest or the highest contributors? (Part 2)

CONTRIBUTION CATEGORY (0 - LOWEST; 1 - HIGHEST)	
HAPPINESS (1 - 10)	
NOW	1.4135
IN 5 YEARS TIME	0.6558
ATTITUDE IN THE SOCIETY (0/DON'T AGREE - 1/AGREE)	
"CAN NOT TRUST MOST PEOPLE"	2.0032
"PEOPLE TEND TO BE FAIR"	6.3366*
"PEOPLE DO NOT HELP EACH OTHER"	0.6158
"DO NOT TRUST STRANGERS"	1.3290
GENERAL TRUST (1/VERY OFTEN - 5/RARELY)	
"LEAVE DOOR OPEN"	0.9368
"LEND MONEY"	0.5501
"LEND POSSESSIONS"	2.0844**
TRUSTWORTHINESS	0.8769
CIVIC NORMS (0 - NOT JUSTIFIABLE; 1 - JUSTIFIABLE)	
"CHEAT ON SOCIAL SECURITY"	0.8473
"CHEAT ON PUBLIC TRANSPORTATION"	0.7388
"CHEAT ON TAXES"	2.0987**
"KEEP SOMEBODY'S LOST MONEY"	1.5570**
"LEAVE AFTER CAUSING CAR DAMAGE"	1.0954
ATTITUDE TOWARD POWER (1/DISAGREE - 7/AGREE)	
"CHILDREN SHOULD LEARN TO RESPECT AUTHORITY"	0.9205
"DAMAGING HONOUR SHOULD BE PUNISHED"	1.0636
"PEOPLE ARE EITHER WEAK OR STRONG"	0.7956
"PEOPLE SHOULD LOVE THEIR PARENTS"	0.7931
"ADMIT MAKING MISTAKES"	0.1574*
"TRY TO RETALIATE"	0.1199**
"POLITE TO UNPLEASANT PEOPLE"	0.3053*
CONSTANT	142.2561
PSEUDO R^2	0.4606
OBSERVATIONS	470

NOTE: Logit regression results (odds ratios). Coefficient significantly different from zero at *10%, **5%, ***1% significance level.