G-COE GLOPE II Working Paper Series

Normative sentiment and "Altruistic" Punishment

Kazumi SHIMIZU (Waseda University)

Daisuke UDAGAWA (Waseda University)

Motoki WATABE (Waseda University)

Working Paper No. 2

Normative sentiment and "Altruistic" Punishment[1]

Kazumi SHIMIZU (Waseda University)[2]

Daisuke UDAGAWA (Waseda University)[3]

Motoki WATABE (Waseda University)[4]

[1] Please don't cite or circulate this paper without author's permission.

[2] Kazumi SHIMIZU is the corresponding author and Associate Professor of Economics, Waseda University. Corresponding address: 1-6-1, Nishi-waseda, Shinjuku-ku,Tokyo 169-8050, Japan; E-mail: skazumi1961@gmail.com.

[3] Daisuke Udagawa is Assistant Researcher of Institute for Research in Contemporary Political and Economic Affairs, Waseda University, Corresponding address: Institute for Research in Contemporary Political and Economic Affairs, 1-6-1 Nishiwaseda, Shinjuku-ku, 159-8050,Tokyo, JAPAN;E-mail: udagawa_daisuke@suou.waseda.jp.

[4] Motoki WATABE is Associate Professor of Social Psychology, Waseda University. Corresponding address: Waseda Institute for Advanced Study, 1-6-1 Nishiwaseda, Shinjuku-ku, 159-8050,Tokyo, JAPAN;E-mail: mwatabe@aoni.waseda.jp.

Abstract

Recent studies on social dilemmas suggest that the punishment of noncooperative behavior leads to higher rates of contribution or toward full cooperation. However, the provision of punishment creates a second-order "free-riding" problem: group members free-ride on the costly punishment given by others. In spite of this problem, on the assumption of the group conflicts or competitions, multilevel selection may allow the proliferation of an otherwise unviable trait: "altruistic" punishment (Bowles 2004; Gintis 2000). In this case, a necessary condition of "altruistic" punishment is that the group member can expect some reciprocal benefit. On the other hand, we observe the "altruistic" behavior even in a situation where the reciprocal benefits can not exist objectively; e.g., cooperation in a one-shot prisoner's dilemma. To explain this kind of behavior, we rely on the evolutionary psychology theory. That is, we suppose that the "altruistic" behavior must have an evolutionarily adaptive function. Following this account, we intend to show, by an experimental approach, that it is not group affiliation, but a specific sentiment, "fairness," that drives "altruistic" behavior. The results of the experiment (N=47) support two hypotheses: 1) Group affiliation does not drive altruistic punishment in the absence of expectation of reciprocal benefits; and 2) punishment against "free-riders" is driven by a sentiment of fairness. More specifically, 2-a) the

higher the endorsement of fairness, more severe the punishment and 2-b) “fair-minded” participants will punish a “free rider” independent of their group affiliation.

Acknowledgements

We thank Urs Fischbacher who provided us with his z-Tree software. We gratefully acknowledge financial support from Waseda University 21COE program.

1. Background

Over the past few years, a considerable number of experimental studies on human behavior in social dilemmas suggest that punishment of noncooperative behavior leads to a higher rate of contribution or toward full cooperation (Fehr & Gächter 2000a). However, the provision of punishment creates a second-order "free-riding" problem: group members' free-ride on the costly punishment given by others (we call this kind of punishment "altruistic" in the sense in which biologists use the term "altruistic". That is, the altruism refers to behavior by an individual that increases the benefit of another individual while decreasing the benefit of the actor). Considering resource exchange among related actors like kin or non-kin members in repeated interactions, the theoretical constructs of "kin selection" (Hamilton 1964) or "reciprocal altruism" (Axelrod & Hamilton 1981; Trivers 1971) are able to give us a persuasive explanation about the resolution of this problem. If we assume that the exchange occurs among unrelated actors where the dyadic interactions are not repeated, it becomes more complex to explain the resolution of the problem in evolutionary terms. However, several models intend to answer this question. Among them, "indirect reciprocity" (Alexander 1987; Nowak & Sigmund 1998) is worth paying attention. In the same context, considering the group conflicts or competitions explicitly, multilevel selection

which supposes the selection above the individual level may allow proliferation of an otherwise unviable trait: altruistic behavior. “Group selection” (Gintis 2000; Bowles 2004) or “cultural group selection” (Henrich & Boyd 2001) are notable examples. All of these hypotheses are consistent with theories of evolutionary biology in the sense that they suppose that the altruistic behavior contributes to altruistic persons’ adaptive fitness.

A major premise of these evolutionary models is that the group member can expect some reciprocal benefits in return . On the other hand, we observe the altruistic behavior even in a situation where the reciprocal benefit does not *actually* exist; for example, cooperation in a one-shot prisoners’ dilemma. (we call this behavior “purely altruistic” to distinguish it from “altruistic” in the biological/evolutionary sense). Several models have been proposed to explain this “purely altruistic” behavior: “Altruistic motives” theory (Baston 1987) emphasizes individual psychological forces; “minimal group paradigm” (Tajfel & Turner 1979) or “group affiliation” (Bernahard et al. 2006; Goette et al. 2006) insist that the group identity can drive the “purely altruistic” behavior. However, these explanations, especially group affiliation paradigm, do not necessarily support the evolutionary theories, because “altruistic” behavior does not necessarily increase the fitness of individuals who provide it. How can we reconcile them?

We have two hypotheses. The first one follows some studies which emphasize the effect of subjective heuristics on the given payoff structure as being an important condition which enables "purely altruistic" behavior (Kiyonari et al. 2000). The second is original to our work and it emphasizes the role of sentiment as a booster for "purely altruistic" behavior. In other words,

1) It is not the group affiliation but a possibility of reciprocity that ensures "altruistic" behavior,

2) It is a specific sentiment which drives "purely altruistic" behavior *even* in the absence of reciprocal opportunity[5].

About the first hypothesis, we have several pieces supporting evidence. Of particular importance is the "bounded generalized reciprocity" theory (Yamagishi et al. 1999). It shows the importance of the *subjective* expectation of reciprocity (even if there is not such possibility *objectively*) to evolve the cooperation in the sizeable group[6]. Following this account, our second hypothesis intends to show that a specific sentiment,

---

[5] Though it seems to be paradoxical, our first hypothesis is not contradictory with experimental results on basis of "group affiliation" paradigm (Bernahard et al [2006], Goette et al [2006]). As Bernahard et al [2006] used two tribes of Papua New Guinea and Goette et al [2006] used soldiers of Swiss army as subjects, it is very probable that they expect the reciprocity subjectively though the reciprocity does not exit objectively.

[6] If we add to it, many others experimental findings which show gradual decrease of cooperation rate in a partner treatment but without reciprocal benefit (e.g., Fehr & Gächter 2000) may support our first hypothesis.

"fairness," drives purely altruistic behavior, that is, altruistic behavior occurring in a situation where the actor is *subjectively* aware of the absence of reciprocity[7]. Note that in accordance with evolutionary psychology theory, we presuppose the role of "sentiment" as an unconscious "adaptive" strategy which has been fostered in the human EEA (Environment of Evolutionary Adaptedness) (Cosmides & Tooby 1989, 2000; de Waal 1996). We focus on sentiment because by examining the psychological process in play when purely altruistic behavior occurs, we can recognize that sentiment plays the most important role in this process. When we display purely altruistic behavior in ordinary life (e.g., anonymous fund-raising, tip paid in a restaurant that will not be visited twice), this decision does not usually depend on long-term speculation, rather, it is provided quite automatically. It seems to rely on the sentiment rather than reason, because sentiment has only a short-term scope but enables rapid decision making[8] (Buller 2005). In other words, it is very difficult to explain this kind of behavior rationally, because, rationally speaking, purely altruistic behavior only costs, without any attending benefits.

---

[7] According to our previous researches (Morimoto & Watabe 2005; Shimizu 2005; both in Japanese), the "fair-minded" person has a tendency to provide a punishment as warning, while the "not-so-fair-minded person" punishes with the motive of retaliation. These findings are coherent with the results of our experiment.

[8] We note that sentiment and reason are not conflicting, but both of them are differently *adaptive*.

The purpose of our study is to verify experimentally the two hypotheses stated previously. For the first hypothesis, we show that the effect of "group affiliation" on cooperation and punishment does not occur in a situation where no reciprocal benefit can be expected by participants. For the second, it will be shown that purely altruistic punishment is provided even in this situation, but only to the degree that it is based on a specific "normative sentiment": fairness.

2. Experiment

2.1. Sentiment measures

As we are focusing on the relation between sentiment and human behavior, we have begun to collect "sentiment measures" data from undergraduate students at Waseda University since 2003(Shimizu 2005). The data enable us to classify, subjects by their recent scores on their sentiment measures, prior to recruiting them into experiments. These indices consist of a general trust scale, a trust/care scale (Yamagishi & Yamagishi, 1994), a self-fairness scale, an emotional sympathy scale, and a perspective-taking scale. Note that these sentiments are considered to have been elaborated along the human evolutionary process (Cosmides & Tooby 1989; de Waal 1996). The general trust scale, trust care scale, self-fairness scale, and perspective-taking scale consist of 6 items each,

and the emotional sympathy scale of 10 items.[9] Each item is a Likert-type rating scale of 7 levels. The α value of each aggregated scale has been consistently higher than 0.8 since 2003; thus, these indices can be considered to be stable. As of May 2006, we have constructed a subject pool at Waseda University (N=320), and we have sorted them for the experiment according to their scores on these scales.

2.2. Procedure

The experiment was conducted at July 2006 in an experimental laboratory at Waseda University in Japan. This laboratory consists of 20 computers on small desks. Each desk is divided from the others by a partition so that each participant is aware of other participants' presence but can see neither their faces nor their behaviors. Upon arrival to the laboratory, each participant was immediately led a desk to minimize the chances that they might meet one another. This was intended to ensure the anonymity of subjects as perfectly as possible. Participants did not meet each other either before or after the experiment. Each experimental session lasted less than about 45 minutes.

At the beginning of the experiment, participants were led to believe that there were eight participants in each experiment session, divided into two 4-person groups.

[9] See the Appendix 1 for the scale items.

However, only one of the eight was a real participant. The other seven "participants" were computer programs running on the computer used by the human participant. Again, because the desks were separated only by thin partitions, participants were aware that others were present. This was intended to reduce any suspicions among them that they were interacting with computer programs. Actually, only three participants revealed suspicions about the others' presence in the post-experimental questionnaire.

2.3. Design of experiment

2.3.1 Grouping condition

The experiment began with verbal instructions to the participants about how the ostensible group of eight was to be divided. Participants were assigned to two conditions sequentially, although the order in which they experienced the conditions was randomly assigned to control for order effects. For the random matching condition, participants were instructed that they had been divided into two 4-person groups by the computer randomly. For the "*gakubu"* condition[10], they were instructed that one of the groups consisted of students who attended the same *gakubu* and that the

[10] The Japanese word "gakubu" is an equivalent of school or college in American universities.

other consisted of students from various other *gakubu*[11]. After each instruction, participants played a gift-giving game, faced a punishment opportunity, and answered the questionnaires immediately after each opportunity for punishment. The following figure is a simple illustration of the basic structure of this experiment.

[Figure 1 about here]

2.3.2. Gift giving game and punishment opportunity

After the verbal instructions, participants read instructions on the computer monitor telling them that they would play a gift-giving game repeatedly in each group and that this game would be followed by an opportunity to penalize (punish) other players. In each round, every participant was endowed with 120 yen (≈ 1 dollar or 1 euro) and decided whether to contribute all or nothing. If he/she contributed all (the "cooperator" strategy), the amount was multiplied by 1.5, to 180 yen, and distributed equally among the other 3 participants, 60 yen to each. If all the participants were to contribute all, each one would obtain 180 yen as a result; on the contrary, if nobody contributed (the "free-rider" strategy), each one would keep only his/her original endowment.[12] If only one did not contribute, the free-rider would gain 300 yen and each cooperator would

[11] Actually, all participants in the *gakubu* condition were led to believe that they belonged to the same college group.

[12] Actually, the experimenters set the endowment and the multiplier such that the profit of each participant might be between 800 and 1500 yen. 900 yen is approximately equivalent to a student's hourly wage. Thus, the stake was substantial.

gain only 120 yen. Free riding was the dominant strategy; on the contrary, the Paretian efficient strategy is full cooperation(Luce & Raiffa 1989). Therefore, the cooperator strategy could be seen as a public good.

The payoff structure was as follows. Subject *i*'s payoff in a group is given by π and subject *i*'s contribution by g_i :

$$\pi_i^1 = \omega - g_i + a \sum_{j=1}^{n} g_j \qquad 0 < a < 1 < na$$

$$\partial \pi_i^1 / \partial g_i = -1 + a < 0 \therefore g_i = 0$$

ω = 120 yen
n =4
g = 0 or 120
a = 0.5

In the gift-giving game, participants were told that there would be several rounds of trading, but they did not know exactly how many rounds the game would last. It was only after all the rounds of trades was complete that the subjects were informed about the individual contributions and the monetary profit of the other group members on their computer monitor. A one-shot situation was ensured, so there was no possibility of reciprocal benefit. Note that the information about contributions and profits was artificially made *identical* for all of the subjects: every participant was led to believe that there was one perfect free-rider and one perfect cooperator in each group.

At this point the subjects had an opportunity to penalize other players. Participants were told that the punishment could be provided only once to both in-group players and out-group players and that anonymity was ensured. In other words, the punished person would not know by whom he/she had been punished. When they decided to provide punishment, they paid from their own earnings exactly the amount by which they wished to penalize other participants. The same penalty amount was ostensibly subtracted from the earnings of the punished participants by the experimenter (therefore the cost/benefit ratio was 1). Rationally speaking, punishers could not expect any benefit from their punishment behavior. Thus, no punishment is the dominant strategy.

Before beginning the game, every participant was asked to try the paper test about the payoff structure of the game. They were required to think until arriving at the correct answer. This was intended to make participants understand that they could not expect reciprocity.

2.4. Participants

Fifty-two undergraduate students (34 males, 18 females) from a subject pool that we have constructed at Waseda University in Japan participated in this experiment. The potential participants in the subject pool were first-year undergraduate students from

various colleges who registered in response to solicitation for participation, and for whom we had measured "sentiment indices." Monetary rewards were emphasized while recruiting potential participants. Examining the answers to the post-experimental questionnaire, five participants were eliminated from the analysis, three who expressed suspicions about the experimental procedure and another two who failed to understand the procedure.

3. Results

3.1. Group effect

According to group affiliation paradigms like the "minimal group paradigm", there should be some kind of group effect on the punishment costs and the cooperation rates. To the contrary, we hypothesized that the absence of reciprocal benefits would prevent the participants from cooperating and punishing *altruistically*. If in-group favoritism is supported, participants should punish the false friend more severely than the outside betrayer: that is, they should punish in-group free-riders more severely than out-group free-riders. Looking at Table 1, the cell C should be larger than the cell D and the largest. However our hypothesis predicted that there would not be any difference.

[Table 1 about here]

A T-test reveals that there is no significant difference in the average penalty assessed against free-riders across conditions or free-rider group membership[13].

As Table 1 shows high variance in the mean penalty amount, we used a logit model to compare the difference of punisher ratio between the two conditions. We can see from Table 2 that there is no significant difference in the percentage of participants who chose to assess penalties on free riders between the two experimental conditions[14].

[Table 2 about here]

If in-group favoritism were to occur, it should increase the in-group cooperation rate: participants should cooperate more eagerly with the in-group than with out-group members (that is, in the "grouping" condition, where group membership is made salient). However, as Table 3 shows, we did not observe any significant difference in this matter.

[Table 3 about here]

In conclusion, these findings contest the group affiliation paradigm in the absence of the possibility of reciprocal benefit. Yet it follows the concept of "bounded generalized reciprocity" (Yamagishi et al. 1999) and is consonant with evolutionary

---

[13] There is a weak negative correlation between each one's profit and sum of punishment (*correlation coefficient*$= -0.215, p < 0.05$). About the mean and the standard deviation of profit, see Appendix.

[14] Note that 47 samples are not large enough to use adequately logit model. However we can safely state that this result is consistent with the T-test result.

biological knowledge.

3.2. Fairness and altruistic punishment

We have predicted that normative sentiment could drive altruistic punishment behavior. This prediction is verified, since altruistic punishment is correlated significantly with participants' endorsement of "fairness": the higher the endorsement of fairness, the more severe the punishment Fairness correlates weakly with cooperation rate (r = 209, p < .05). In addition, if we compare the costs of punishment incurred against in-group and out-group free-riders, the results indicate that their punishment is independent of grouping conditions. That is, participants who choose to punish do so *impartially*.

[Table 4 about here]

3.3. Supportiveness and altruistic punishment

We conducted a questionnaire after each punishment opportunity. In this questionnaire, there was a question about supportive motif: "Suppose there is a person punishing free-riders. How strongly would you support him/her financially?" The question is answered by a rating on a Likert scale of 7 levels. The rate of cooperation correlates significantly with rated supportiveness (p <.01): highly cooperative persons are more willing to support altruistic punishment. Fairness also correlates significantly with

supportiveness ($p < .05$), but less strongly.

[Table 5 about here]

4. Conclusions

We have reported the results of an experiment which examines the impact of group affiliation and of "fairness" sentiment on the altruistic behavior. We found that group affiliation does not play a important role in a situation where participants do not expect reciprocal benefit, and that even in this situation, "fairness" can drive the altruistic punishment. These result support other studies (e.g., Yamagishi et al 1999; Kiyonari et al 2000), and are consonant with our previous research (Morimoto & Watabe 2005; Shimizu 2005) .

If punishing free-riders can yield a solution to social dilemmas, one possible solution might be that that the first blow of punishment is given by a fair-minded person, and human society profits from that person's altruistic behavior. However, evolutionary psychology theory tells us that purely altruistic behavior is not sustainable if the altruistic actor does not receive any benefits. Thus, further research is needed to understand precisely which social elements or conditions enable us to "institutionalize" individual punishment.

References

Axelrod, R., & Hamilton, W.D. 1981 The evolution of cooperation, *Science*, 211, 1390-1396.

Alexander, R.D. 1987 *The biology of moral systems*, Hawthorne, NY: Aldine de Gruyter.

Baston, C.D. 1987 Prosocial motivations: Is it ever truly altruistic? In L.Berkovitz (Eds.), *Advances in experimental social psychology*, vol.20, Orlando: Academic Press, 65-122.

Bernahard, H., Fehr, E., & Fischbacher., U. 2006 Group Affiliation and altruistic norm enforcement, *American Economic Journal*, 5/1/2006. vol.96, 217-221.

Bowles, S. 2004 *Micoroeconomics: Behavior, Institutions, and Evolution*, Princeton University Press.

Buller, D. J. 2005 *Adapting Minds, Evolutional Psychology and the Persistent Quest for Human Nature*, MIT Press.

Cosmides, L., & Tooby, J. 1989 Evolutionary psychology and the generation of culture: part □, A computation theory of social exchange, *Ethology and Sociobiology*, 10, 51-97.

－[2000], “Evolutionary Psychology and the Emotions”, *Handbook of Emotions*, edited

by M.Lewis and J-M. Haviland-Jones, The Guiford Press.

de Waal, F. 1996 *Good Natured, The Origin of Right and Wrong in Humans and Other Animals*, Harvard University Press..

Fehr, E., & Simon, G.. 2000a "Cooperation and Punishment in Public Goods Experiments", *American Economic Review*, vol.90, no.4, 980-994.

------ 2000b "Fairness and Retaliation: The Economics of Reciprocity", *Journal of Economic Perspectives*, vol.14, no.3, 159-181.

Gintis, H. 2000 Strong reciprocity and human society, *Journal of Theoretical Biology*, 206, 169-179.

Goette, L., Huffman, D., & Meier, S. 2006 The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups, *American Economic Review* (papers and proceedinds), 96(2), 212-16.

Hamilton W. 1964a The Genetical Evolution of social behaviour, *Journal of Theoretical Biology*, 7, 1-16.

----1964b The genetical evolution of social behaviour, *Journal of Theoretical Biology*, 7, 17-32.

Henrich, J., & Boyd, R. 2001 Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas,

Journal of Theoretical Biology, **208**, 79-89.

Kiyonari, T., Tanida.,S and Yamagishi, T. 2000 Social exchange and reciprocity: confusion or a heuristic?, *Evolution and Human Behavior*, 21, 41-427.

Luce R.D., & Raiffa, H. [1967] 1989 *Games and Decisions, Introduction and Critical Survey*, Dover.

Trivers, R. 1971 The evolution of reciprocal altruism, *Quartely Review of Biology*, **46**, 35-57.

Nowak, M.A., & Sigmund, K. 1998 Evolution of indirect reciprocity by image scoring, *Nature*, 393, 573-577.

Tajfel, H., & Turner, J.C. 1979 An itegrative theory of intergroup conflict. In W.G. Austin & S. Worchel (Eds.) *The social psychology of intergroup relations,* (pp.33-47), Monterey, CA: Brooks/Cole.

Yamagishi, T., and Kiyonari, T. 2000 The group as the container of generalized reciprocity, *Social Psychology Quarterly*, **63** (2), 2000, 116-132.

Yamagish, T., and Yamagishi. M. 1994 Trust and commitment in the United States and Japan. *Motivation and Emotion* 18(2), 1994, 129-166.

Yamagishi, T., Jin, N., & Kiyonari, T. 1999 Bounded generalized reciprocity: in-group boasting and in-group favoritism, *Adavances in Group Processes*, **16**, 161-197.

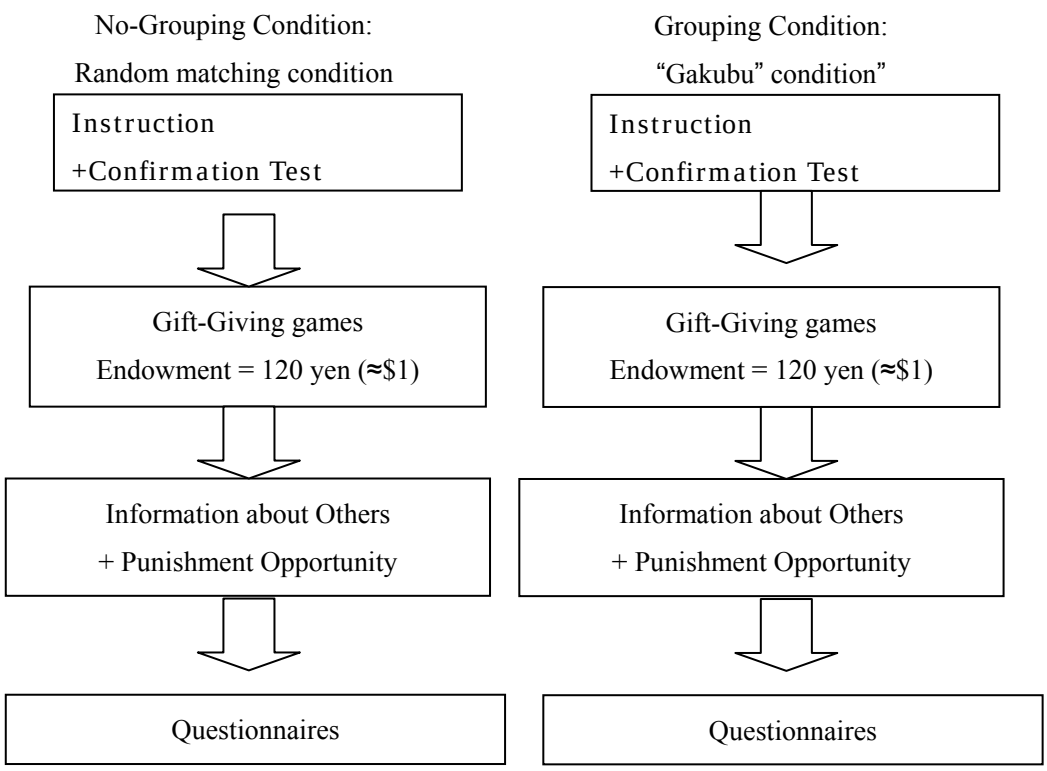

Figure 1. Basic Structure of the Experiment.

Table1. Mean Amount of Punishment against Free-Rider

N=47	In-Group	Out-Group
No-Grouping	A)16.4 yen (Sd=32.2)	B)18.6 yen (Sd=35.2)
Grouping	C)17.7 yen (Sd=38.6)	D)13.5 yen (Sd=36.3)

Table 2. Percentage of Participants who Punish: logit model

N=47	In-Group	Out-Group
No-Grouping	36% (17/47)	38% (18/47)
Grouping	34% (16/47)	26% (12/47)

Table 3. Mean of Cooperation Rate: T-test

	In-Group	Out-Group
No-Grouping	2.06/5 (Sd=1.66)	∅
Grouping	2.21/5 (Sd=1.77)	∅

Table 4. Correlation between “Fairness” and Punishment Cost

	Sum of Punishment against Free-Rider (Total)	Sum of Punishment against FR (In-Group)	Sum of Punishment against FR (Out-Group)
“Fairness” sentiment	**0.365****	**A) 0.413****	**B) 0.386****

∗ **p < 0.01.

∗ There is no significant difference between A) and B).

∗ Spearman’s correlation coefficient

Table 5. Correlation between Cooperation rate and Supportiveness

	Supportiveness
Cooperation rate	**0.421****
“Fairness” sentiment	**0.218***

∗ **$p < .01$.

∗ * $p<.05$.

∗ Spearman’s correlation coefficient.

Appendix1: Sentiment Measures

Fairness Measure

- I think I always behave fairly.
- I never want to do anything dishonest.
- I always keep the spirit of fair play regardless of the circumstances.
- I never do any mean things.
- I always try to act fairly.
- I am a fair person.

Sympathy Measure

- When I see people in trouble, I do not feel much sympathy.
- When I see someone crying for joy, it turns me off.
- When I see someone crying, I get irritated rather than empathetic.
- I cannot understand why others get so upset.
- When I see someone hurt, I usually stay calm.
- I can stay calm and uninterested even when others around me are in distress.
- When my friend starts talking about his/her trouble, I feel like changing the subject.
- I can remain uninterested and calm even when others around me are unhappy.

Perspective Taking Measure

- In order to understand my friends well, I try to put myself in their shoes.
- Before criticizing someone, I try to think what if I were him/her.
- If someone makes me feel irritated, I take his/her perspective to understand him/her.
- When I make a decision, I try to think of the opposite perspective of mine.
- Since I think every problem has two opposed viewpoints, I always consider both of them.
- I tend to be empathetic to others.
- I feel happy when I see someone feeling happy even though I have no direct relation with the person.
- I think other people would be pleased with my happiness.
- I think many people would enjoy others' unhappiness.
- I feel sad when I see people or animals being abused on TV.
- I always try to think from another's point of view.

Appendix 2： Earnings at gift-giving game

N=47	Min	Max	Av. Sum	Sd
No-Grouping	480 yen	1080 yen	827 yen	195.8 yen
Grouping	480 yen	1080 yen	784 yen	204.6 yen