Second-Party and Third-Party Punishment in a Public Goods Experiment[*]

By Yan Zhou, Peiran Jiao, and Qilin Zhang

This version: February, 2016

Abstract

We experimentally investigate whether third-party punishment is more effective than second-party punishment to increase public goods contribution. In our experiment, third parties first played the standard public goods game and then made punishment decisions as independent bystanders. We find that third parties punished more frequently, severely, and less antisocially, resulting in a higher contribution level than that driven by second party punishment. The third party's exaggerated emotion towards free riders is proposed to explain their superior punishment effectiveness.

Keywords: third-party punishment, second-party punishment, public goods experiment, free rider

JEL Classification Codes: H41, C92, D70

Running title: Second-party and third-party punishment

[*] Yan Zhou: School of Management, Harbin Institute of Technology, Harbin, China 150001; Peiran Jiao: Department of Economics and Nuffield College, University of Oxford; Manor Road Building, Manor Road, Oxford, OX1 3UQ, UK; Qilin Zhang: School of Economics and Finance, University of Hong Kong, Room 908, K.K. Leung Building, Pokfulam Road, Hong Kong. We thank Simon Gächter, Joshua Tasoff, Pinghan Liang, and two anonymous referees for their comments and Prof. Mingjian Zhou for his assistance. This study benefited from participation in 2013 North-American Economic Science Association Conference. This study was financially supported by National Natural Science Foundation of China (71203046).

1. Introduction

Extant literature has documented the importance of peer punishment in enhancing cooperation, coordination, and contribution (e.g. Fehr and Gächter, 2000; Nikiforakis, 2008; Reuben and Riedl, 2013). While the main focus was on second-party (or peer) punishment by punishers who are directly affected by those whom they punish (for a review, see e.g. Chaudhuri, 2011), a growing literature switched attention to another non-negligible type of punishment, the one by third parties who are independent of or only indirectly affected by those whom they punish (e.g., Fehr and Fischbacher, 2004; Leibbrandt & López-Pérez, 2012). We will refer to peer punishment as second-party punishment (*SPP*) hereafter. Regarding punishment effectiveness, second parties could punish more severely because they are directly harmed by bad behaviors (e.g., free riding); however, they are also found to suffer from antisocial punishment which may deteriorate effectiveness (Herrmann et al., 2008). In an unaffected position, third parties may have less incentive to punish, but could potentially punish more impartially. Since both *SPP* and third-party punishment (*TPP*) have pros and cons, no conclusion has been drawn in the literature regarding which type is more effective. The present paper contributes to this literature by uncovering evidence in favor of TPP in a controlled experiment.

Several experimental studies have shed some lights on this puzzle but have not reached agreement. In the one-shot modified dictator game, Fehr and Fischbacher (2004) report 61% of the third parties did punish, compared with 74% of the second parties; and the ratio was 54% vs. 60% in Leibbrandt and López-Pérez (2012). Fehr and Fischbacher's one-shot prisoners' dilemma suggests similar pattern: third parties punished less frequently and less severely than second parties.

By contrast, the repeated One-way Treatment in Carpenter and Matthews (2009) shows that the average expenditures on *TPP* and *SPP* were 0.67 and 0.50 points respectively, indicating that third parties punished more severely. However, they failed to control for the potential scale effect: a second party was able to punish only 3 others while a third party was able to punish 4. Besides, each punisher simultaneously played both roles of second and third party, so the potential interaction effect remains unclear, and hence we cannot compare the two types of punishment directly. Our experiment controls for the scale effect and lets each punisher only take one role at a time, so as to provide a more precise and persuasive comparison.

2. The Experiment

Using between-subject design, the experiment consisted of two treatments: *SPP* (treatment *S*) as a baseline and *TPP* (treatment *T*). Each treatment had two stages: the *Contribution* stage followed by the *Punishment* stage. Each session of both treatments lasted for 20 periods. In each period participants were matched anonymously and randomly, and played the game in groups of four.

The *Contribution* stage was the same in both treatments. At the beginning each subject received an endowment of 20 tokens, and then simultaneously decided how many tokens to contribute to the public account with the marginal per capita return equal to 0.4, meaning one token contributed increased each in-group member's payoff by 0.4 tokens. At the end of this stage, subjects were informed of their own payoffs.

In the beginning of the *Punishment* stage in *S*, subjects were informed of each in-group member's contribution. Then they decided whether to punish, whom to punish and how many tokens to assign for punishing. One token used to punish cost one token to the punisher, but reduced three tokens of the earning of the targeted subject.

Treatment *T* was the same as *S* except that in the *Punishment* stage players made punishment decisions as independent third parties. This technique was similar to that in Fehr and Fischbacher (2004). Groups were numbered 1 through 6, and each group could only punish members in the next adjacent group.[1] In the beginning of the *Punishment* Stage, third parties were only informed of contribution levels of the group they could punish, but no information regarding their own group. Hence, from a punisher's perspective second parties and third parties were in informationally equivalent position. To alleviate the potential scale effect, we allowed each third-party punisher to punish only the members of the target group who had different within-group ID numbers from the punisher, so each second and third party had the

[1] We disallowed the last group to punish the first group to avoid the punishment circle that may add reciprocity effect.

opportunity to punish a maximum of three others.

The experiment was conducted using z-Tree 3.3.12 (Fischbacher, 2007) in Shenzhen, China. Totally 36 graduate students participated in Treatment *S* and 24 in Treatment *T*; 35 were males and 25 females. Each session took about 1.5 hours. The show up fee was 30 RMB (about 4.9 USD). The exchange rate between experimental tokens and RMB was 10:1. The average earning was 70.4 RMB (11.5 USD). The experimental instructions can be found in *Supplemental Material A*.

3. Results

Figure 1 displays the average punishment and contribution to public goods by treatment. At individual level, third parties punished more severely than second parties (1.6 vs. 0.9 tokens on average, t-test, p=0.025). As to punishment frequency, 40.6% of third parties and 37.5% of second parties punished at least once. Each subject had the opportunity to punish three others per period. Out of these punishment opportunities, punishment frequency was 25.9% for *T* and 19.9% for *S* (Pearson-chi-square test, p<0.01), showing again that third parties punished more frequently. Though both average severity and average frequency decreased gradually over periods, the frequency difference in frequency persisted (see Figure B1 in *Supplemental Material B*).

Figure 2 depicts the distribution of punishment frequency and severity respectively against the contribution difference between the punished and the punisher, bracketed into 3 categories: [-20, -2), [-2, 2], (2, 20].[2] Figure 2 shows that both third and second parties had similar attitudes towards deviators, i.e. the more a person deviated, the more likely and severely the person was punished. Regression analyses confirmed these findings (see *Supplemental Material C*). We also graphed the punishment distribution against the deviation from the target's group average, and found similar pattern (see Figure B2 in *Supplemental Material B*). Besides, there was a strong asymmetry: third parties were more sensitive to negative deviations (i.e., free riders) than to positive ones (i.e., cooperators).

[2] For example, suppose *N* is the punisher and *M* is the punished. *M*'s contribution is 5 tokens while *N*'s is 15 tokens; then the deviation of *M* to *N* is -10.

In public goods game antisocial punishment (*ASP*) refers to punishment towards the player who contributes no less than the punisher (Herrmann et al., 2008). Figure 2 also shows that *ASP* existed among both second and third parties, implying third parties were not absolutely impartial. *ASP* amounted to about 27% in both treatments: 115 out of 430 in *S* and 83 out of 311 in *T*. However, third parties imposed more punishment on free riders (2.92 vs. 1.35) and less on contributors (1.18 vs. 1.80). This suggests *TPP* was less antisocial (also see regression in *Supplemental Material C*).

Potentially due to the more frequent, severer, and less antisocial *TPP*, the average contribution was higher in *T* than in *S* (9.8 vs. 6.4, $p<0.01$), as shown in Figure 1. This indicates a higher effectiveness of *TPP*. Only 0.7 more tokens spent on punishment by third parties resulted in a 3.4-token higher contribution on average, compared with SPP, a net benefit of 2.7 tokens.

Our results are consistent with Carpenter and Matthews (2009) but seemingly contradictory with Fehr and Fischbacher (2004). One explanation would be strategic punishment. However, this should not be a serious issue in our experiment because (1) we used a stranger-matching design; (2) the substantial amount of punishment at the end of experiment cannot be explained by strategic motivation (see Figure B1 in Supplemental Material B); (3) second and third parties faced the same probability of being re-matched with those they punished, so neither should have a stronger strategic motive. A plausible explanation we propose is the emotion effect: third parties showed stronger negative emotions toward bad behavior (e.g., Bosman and Van Winden, 2002; Nelissen and Zeelenberg, 2009). In Fehr and Fischbacher (2004) the third parties were purely independent without the chance of experiencing elevated emotions. In our treatment *S*, second parties only experienced being free ridden; in *T*, however, third parties not only experienced being free ridden but also observed free riding on others. Such kind of doubled experience should associate with stronger negative emotions that triggered heavier punishment.[3] To provide solid evidence supporting such explanation, a supplemental quasi-experiment was conducted (see Supplemental Material D). The results

[3] An anonymous referee mentioned blind revenge as an explanation. We tested it and find it was significant only in *SPP* decision model, but it didn't significantly influence any punishment intensity.

reveal that stronger negative experience indeed associated with stronger negative emotion that triggered heavier punishment.

4. Conclusion

In a public goods experiment, we let third parties play the standard contribution game and then make punishment decisions, which were compared to second-party punishment in the same setup. Since third parties could not punish in-group free riders, their anger might have been elevated so that they sanctioned more heavily towards out-group free riders. As expected, third parties indeed punished more frequently, severely, and less antisocially, leading to a higher level of contribution.

Our findings suggest that incorporating third-party punishers with similar prior experience into punishment institutions would lead to beneficial consequences. For example, in the US a jury of a criminal trial selected from the community has some power to judge whether a wrongdoing is guilty. If the jury were composed of people who had similar prior experience, it could potentially help defend justice, yet the issue arises regarding whether third parties who suffered from prior negative experience may be too emotional and over-punish. Besides, emotion is just a conjectured mechanism to explain why third parties punished more. The lack of direct evidence linking experience, emotion, and heavier punishment calls for further investigation.

References

Andreoni, J. and Gee, L. K. (2012) Gun for hire: delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics,* 96, 1036-1046.

Carpenter, J. and Matthews, P. H. (2009) What norms trigger punishment? *Experimental Economics,* 12, 272-288.

Chaudhuri, A. (2011) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14, 47-83.

Fehr, E. and Fischbacher, U. (2004) Third-party punishment and social norms. *Evolution and human behavior,* 25, 63-87.

Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980-994.

Fischbacher, U. (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics,* 10, 171-178.

Herrmann, B., Thöni, C. and Gächter, S. (2008) Antisocial punishment across societies. *Science,* 319, 1362-1367.

Leibbrandt, A. and López-Pérez, R. (2012) An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization,* 84, 753-766.

Nelissen, R. M. and Zeelenberg, M. (2009) Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making,* 4, 543-553.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*, 91-112

Reuben, E. and Riedl, A. (2013) Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior,* 77, 122-137.

Figure 1 Mean punishment and contribution by treatment

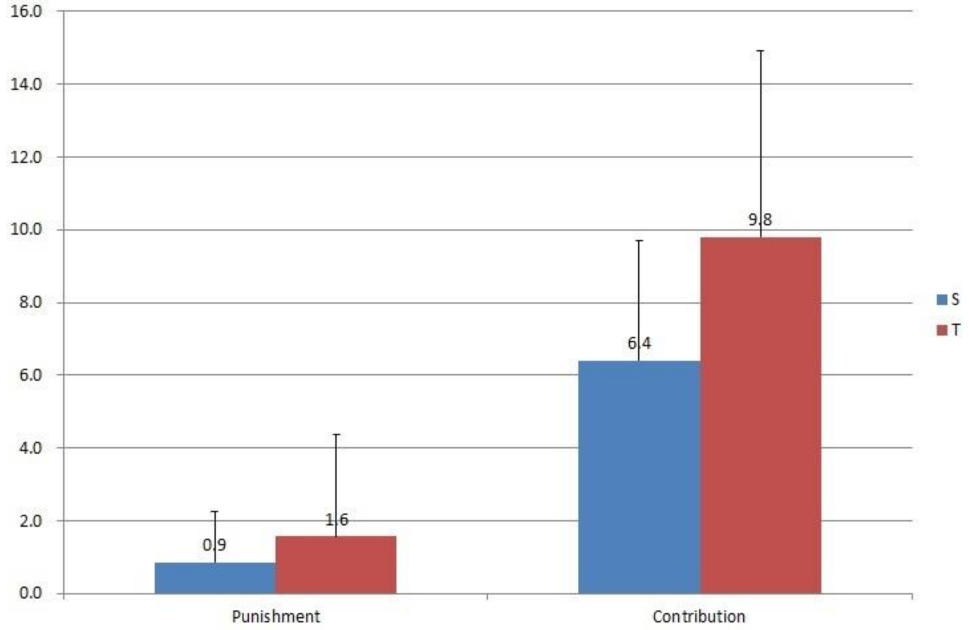
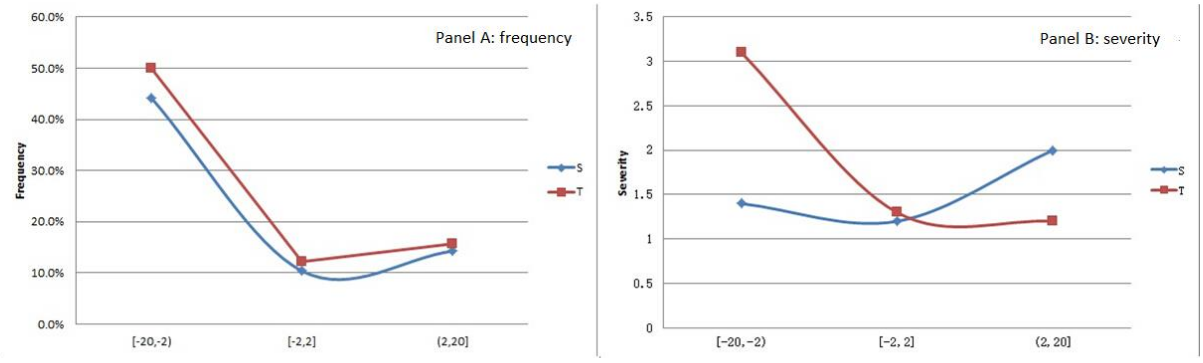


Figure 2 Punishment distribution: frequency and severity



Supplemental Material A: Experimental Instructions

Instruction for S treatment

Welcome to this economic experiment. Be sure that it is safe, both physically and mentally.

If you follow the rules in this experiment and make your decision on your own, then you may earn a considerable amount of money. Therefore, please carefully read this instruction and learn these rules. Talking and any form of communication among you are not allowed when experiment is on. If you have any question, please raise your hand.

The money in the experiment is named ‘token’. The experiment will run 20 periods, namely the game will repeat 20 times. You final earning is the total earnings in 20 periods. Earning will be calculated in token, and your final earning will be exchanged into RMB (USD) at the end of the experiment. 10 tokens equals to 1 RMB (about 0.15 USD). You will be paid immediately by cash plus 30 RMB (5 USD) show-up fee. Whole experiment lasts about 60-90 minutes.

This experiment is a game played by 4 persons in one group. At the beginning of each period, each of you will be randomly assigned into a group, and you will play with your three group members anonymously.

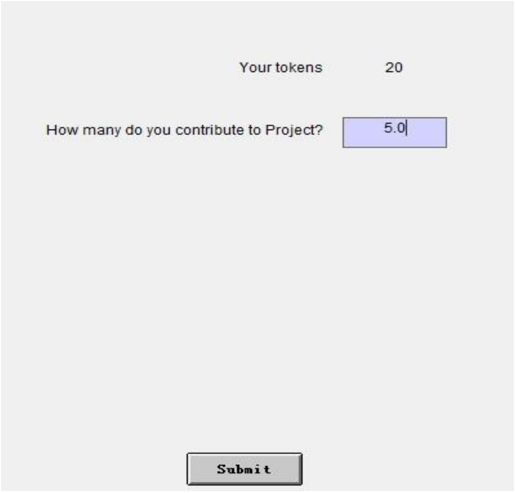
Questionnaire

Before we start, you have 7 questions to answer on the computer to make sure you understand the rules in the game. If and only if everyone correctly answers all the questions, we can start.

The First Stage

Each period consists of two stages. After you are randomly and anonymously assigned to a group, the first stage begins, and each of you will receive 20 tokens.

There is a project (the only one) in your group. If you contribute one token in this project, each member in your group will get 0.4 token. You have to decide how many of the 20 tokens you contribute to a project and how many you keep. See a sample screenshot in Fig. A.1.



Your tokens 20

How many do you contribute to Project? 5.0

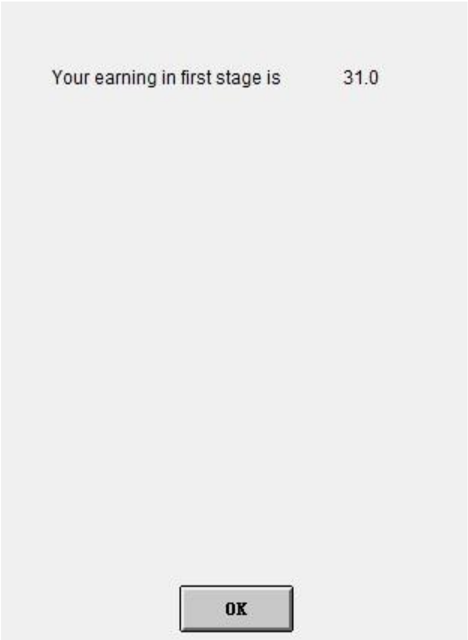
Fig. A.1 Make contribution

In Fig. A.1, the proposed contribution is 5.0 tokens. After all participants have submitted contributions, your earning in your group in this stage will be calculated by following formula:

*Your earning at first stage = (20 tokens - Your contribution to the project in your group) + 0.4*Total contribution to the project in your group*

Notice that your earning comes from two parts: the token you keep (20 tokens - Your contribution to the project) and the income from the project (0.4*Total contribution to the Project). For example, if you and other three members contribute 5, 15, 20, and 0 respectively, your earning in first stage is 31 = (20-5) + 0.4*(5+15+20+0) tokens.

At the end of the first stage, your earning will be displayed on the screen. See Fig. A.2 for an example.



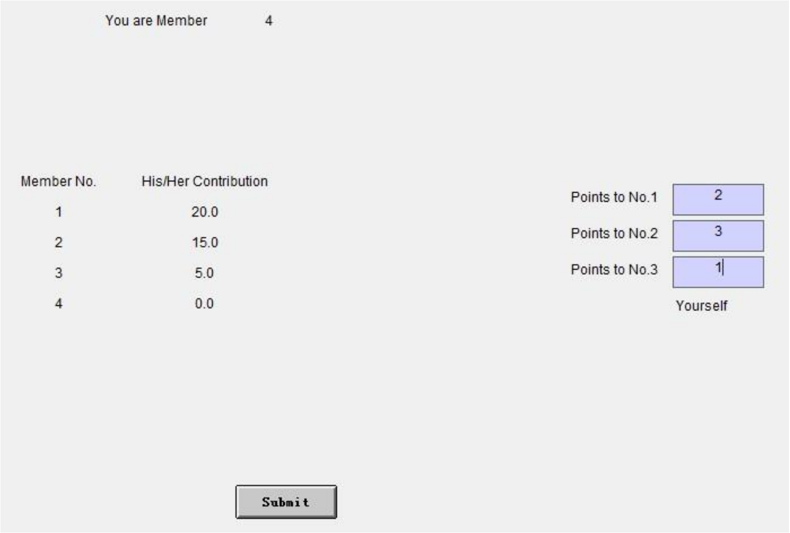
Your earning in first stage is 31.0

Fig. A.2 Earning in the first stage

The Second Stage

After the first stage, you will come to the second stage and will be informed about your number in your group, group members' contributions in first stage, including yours. Please notice that your number in your group will randomly change in each period, so do others.

In this stage you are able to use the 'Points' function. 'Points' makes you be able to reduce member's earning by assigning points to them. If you assign 1 point to a member, you are able to reduce his/her earning by 3 tokens, but you have to pay 1 token for purchasing 1 point, so using Points is costly. See a sample screenshot in Fig. A.3.



You are Member 4

Member No.	His/Her Contribution
1	20.0
2	15.0
3	5.0
4	0.0

Points to No.1 2

Points to No.2 3

Points to No.3 1

Yourself

Fig. A.3 Displaying and Points

In the sample given in the Fig. A.3, participant (he) is Member Four. He decides to assign 2, 3 and 1 point(s) to Member One, Two and Three respectively. So he reduces earnings of Member One, Two and Three by 6, 9 and 3 tokens respectively. Meanwhile, he will pay 2+3+1=6 tokens for using 6 Points.

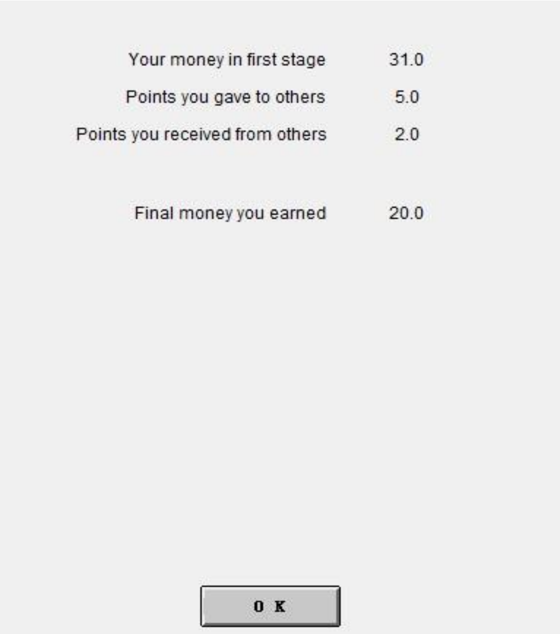
Please notice that (1) the points you assign to any member can't exceed 8 tokens; (2) total points you assign to all other members can't exceed your earning in first stage; (3) off course, other group members in your group can also reduce your earning by using Points if they want to; (4) if tokens reduced exceeds your earning, your earning will equal to 0 automatically, namely we let your earning can NOT be negative, so we do to others.

After all participants have submitted, your earning in your group will be calculated by following formula:

*Your earning at second stage = earning in first stage – total points you assign to others – 3 * total points you receive from others.*

For example, if your earning in first stage is 31 tokens, you assign 5 points to others, and you receive 2 points from others, your earning in second stage is 20=31 – 5 – 3*2 tokens.

Calculation work will be done by computer. After that, you will be informed about the earning in first stage, the total amount of points you assign to others, the total amount of points you receive from others and your earning in second stage. See a sample screenshot in Fig. A.4.



Your money in first stage	31.0
Points you gave to others	5.0
Points you received from others	2.0
Final money you earned	20.0

Fig. A.4 Displaying in second stage

After the second stage one period ends, and then another period begins unless it is the 20th period.

If you have question, please raise your hand. One of our experimenters will come to help you.

Control Questions for S Treatment

1. If you, Member A, Member B and Member C contribute 0, 0, 0 and 0 tokens to Project, what are your earnings at the end of first stage?

You: ______

A: ______

B: ______

C: ______

2. If you, Member A, Member B and Member C contribute 20, 20, 20 and 20 tokens to Project, what are your earnings at the end of first stage?

You: ______

A: ______

B: ______

C: ______

3. If you, Member A, Member B and Member C contribute 12, 8, 16 and 4 tokens to Project, what are your earnings at the end of first stage?

You: ______

A: ______

B: ______

C: ______

4. Suppose your earning in first stage is 40 tokens. You don’t assign any point to others, but you receive 6 points from them. What’s your earning at the end of second stage? ______

5. Suppose your earning in first stage is 20 tokens. You don’t assign any point to others, but you receive 12 points from them. What’s your earning at the end of second stage? ______

6. Suppose your earning in first stage is 35 tokens. You assign 2, 4, 2 points to Member A, Member B and Member C respectively, and you receive 4 points from them. What’s your earning at the end of second stage? ______

7. If your earning in first stage is 20 tokens, can you successfully do following things?

(1) Assign 8, 8, 6 points to Member A, Member B and Member C respectively.

(i) YES (ii) NO

(2) Assign 3, 5, 0 points to Member A, Member B and Member C respectively.

(i) YES (ii) NO

(3) Assign 2, 10, 5 points to Member A, Member B and Member C respectively.

(i) YES (ii) NO

Instruction for T treatment

Welcome to this economic experiment. Be sure that it is safe, both physically and mentally.

If you follow the rules in this experiment and make your decision on your own, then you may earn a considerable amount of money. Therefore, please carefully read this instruction and learn these rules. Talking and any form of communication among you are not allowed when experiment is on. If you have any question, please raise your hand.

The money in the experiment is named ‘token’. The experiment will run 20 periods, namely the game will repeat 20 times. You final earning is the total earnings in 20 periods. Earning will be calculated in token, and your final earning will be exchanged into RMB (USD) at the end of the experiment. 10 tokens equals to 1 RMB (about 0.15 USD). You will be paid immediately by cash plus 30 RMB (5 USD) show-up fee. Whole experiment lasts about 60-90 minutes.

This experiment is a game played by 4 persons in one group. At the beginning of each period, you will be randomly assigned into a group, and you will play with others anonymously.

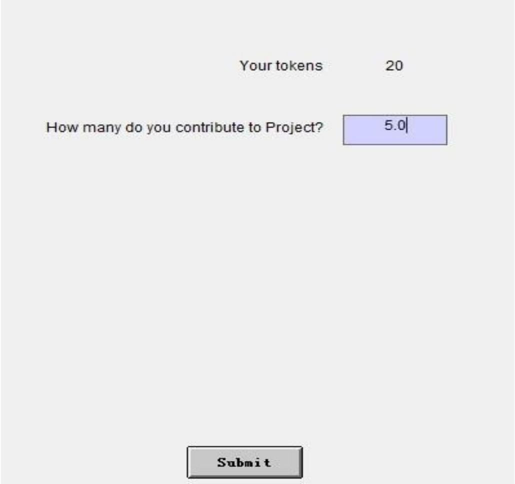
Questionnaire

Before we start, you have 7 questions to answer on the computer to make sure you understand the rules in the game. If and only if everyone correctly answers all the questions, we can start.

The First Stage

Each period consists of two stages. After you are randomly and anonymously assigned to a group, the first stage begins, and each of you will receive 20 tokens.

There is a project (the only one) in your group. If you contribute one token in this project, each member in your group will get 0.4 token. You have to decide how many of the 20 tokens you contribute to a project and how many you keep. See a sample screenshot in Fig. A.5.



Your tokens 20

How many do you contribute to Project? 5.0

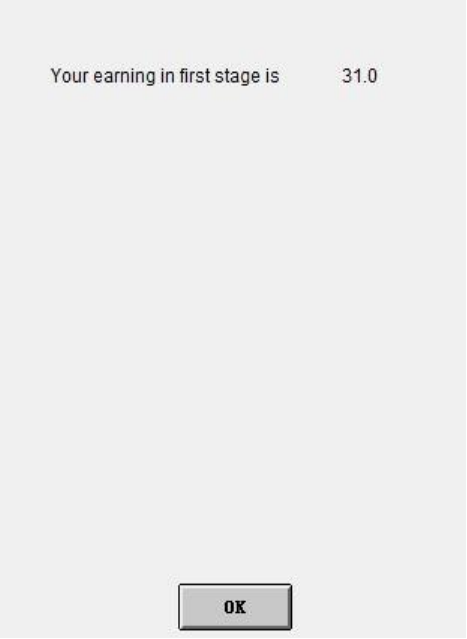
Fig. A.5 make contribution

In Fig. A.5, the proposed contribution is 5.0 tokens. After all participants have submitted contributions, your earning in your group in this stage will be calculated by following formula:

*Your earning at first stage = (20 tokens - Your contribution to the project in your group) + 0.4*Total* *contribution to the project in your group*

Notice that your earning comes from two parts: the token you keep (20 tokens - Your contribution to the project) and the income from the project (0.4*Total contribution to the Project). For example, if you and other three members contribute 5, 15, 20, and 0 respectively, your earning in first stage is 31 = (20-5) + 0.4*(5+15+20+0) tokens.

At the end of the first stage, your earning will be displayed on the screen. See Fig. A.6 for the example.



Your earning in first stage is 31.0

Fig. A.6 Earning in the first stage

The Second Stage

After the first stage, you will come to the second stage and will be informed about your group number, your in-group number, and the contributions of another group in the first stage. Please notice the contribution information is NOT your group's contribution information.

In this stage you are able to use the 'Points' function. Towards whom you can use Points is restricted, say if you are in Group i, you can only use Points towards members in Group i+1, meanwhile members in Group i-1 also can use Points towards you. We have to mention that if you are in the first Group, nobody can use Points toward you; but if you are in the last group, you can use points towards nobody (you will not get in Stage 2 as well). However, since you are randomly assigned into a group at the beginning of every period, each of you gets equal chance to be in the first or last group. Another Restriction is that you can't use Points towards the person whose in-group number is the same to you. Namely you can use Points towards 3 persons rather than 4 persons. See Fig. A.7 for a clear sense.

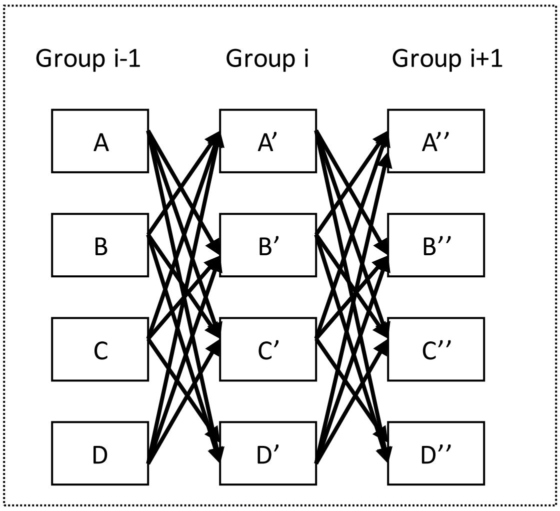
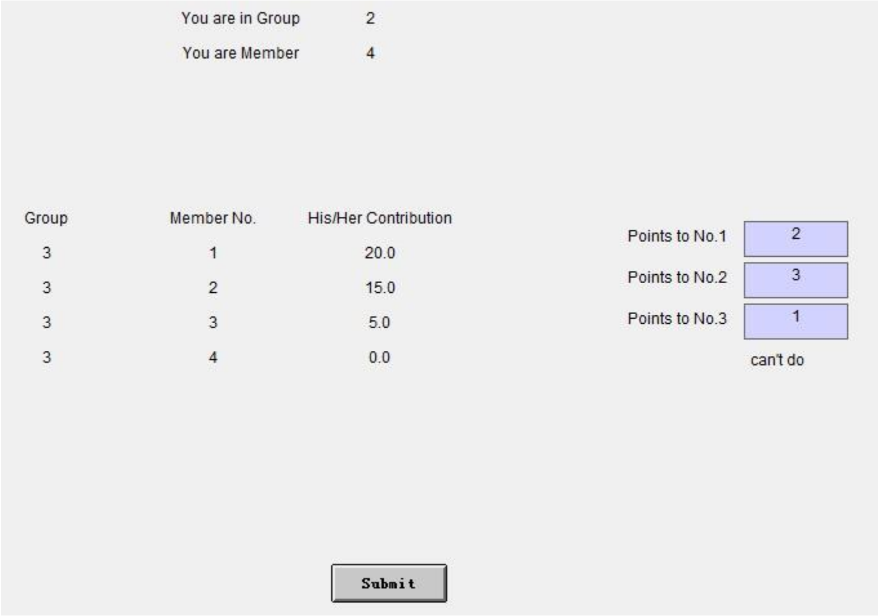


Fig. A.7 Points Restriction

‘Points’ makes you be able to reduce member’s earning by assigning points to them. **If you assign 1 point to a member, you are able to reduce his/her earning by 3 tokens, but you have to pay 1 token for purchasing 1 point,** so using Points is costly. See a sample screenshot in Fig. A.8.



You are in Group 2

You are Member 4

Group	Member No.	His/Her Contribution
3	1	20.0
3	2	15.0
3	3	5.0
3	4	0.0

Points to No.1 2

Points to No.2 3

Points to No.3 1

can't do

Fig. A.8 Displaying and Points

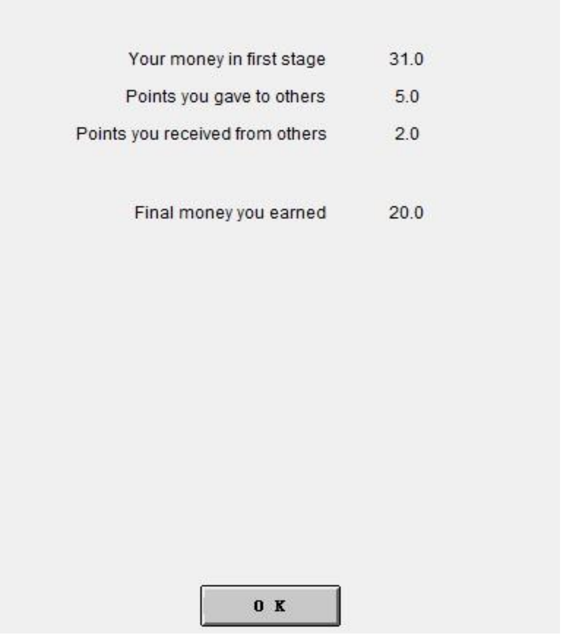
In the sample given in Fig. A.8, participant (he) is Member Four in Group 2. He decides to assign 2, 3 and 1 point(s) to Member One, Two and Three in Group 3 respectively. So he reduces earnings of them by 6, 9 and 3 tokens respectively. Meanwhile, he will pay 2+3+1=6 tokens for using 6 Points.

Please notice that **(1) the points you assign to any member can’t exceed 8 tokens; (2) total points you assign to all other members can’t exceed your earning in first stage;** (3) off course, members in other group can also reduce your earning by using Points if they want to; (4) if tokens reduced by Points exceeds your earning, your earning will equal to 0 automatically, namely we let your earning can NOT be negative, so we do to others.

After all participants have submitted, your earning in your group will be calculated by following formula:

*Your earning at second stage = earning in first stage – total points you assign to others – 3 * total points you receive from others.*

For example, if your earning in first stage is 31 tokens, you assign 5 points to others, and you receive 2 points from others, so your earning in second stage is 20=31 – 5 – 3*2 tokens. Calculation work will be done by computer. After that, you will be informed about the earning in first stage, the total amount of points you assign to others, the total amount of points you receive from others and your earning in second stage. See the sample screenshot in Fig. A.9.



Your money in first stage 31.0

Points you gave to others 5.0

Points you received from others 2.0

Final money you earned 20.0

Fig. A.9 Displaying in second stage

After the second stage one period ends, and then another period begins unless it is the 20th period.

If you have question, please raise your hand. One of our experimenters will come to help you.

Control Questions for T Treatment

1. If you, Member A, Member B and Member C in the same group contribute 0, 0, 0 and 0 tokens to Project, what are your earnings at the end of first stage?

You: ______

A: ______

B: ______

C: ______

2. If you, Member A, Member B and Member C in the same group contribute 20, 20, 20 and 20 tokens to

Project, what are your earnings at the end of first stage?

You: ______

A: ______

B: ______

C: ______

3. If you, Member A, Member B and Member C in the same group contribute 12, 8, 16 and 4 tokens to Project, what are your earnings at the end of first stage?

You: ______

A: ______

B: ______

C: ______

4. Suppose you are in Group 3 and your earning in first stage is 40 tokens. You don't assign any point to others, but you receive 6 points from others. What's your earning at the end of second stage? ______

5. Suppose you are in Group 3 and your earning in first stage is 20 tokens. You don't assign any point to others, but you receive 12 points from others. What's your earning at the end of second stage? ______

6. Suppose you are in Group 3 and your earning in first stage is 35 tokens. You assign 2, 4, 2 points to Member A, Member B and Member C in Group 4 respectively, and you receive 4 points from others. What's your earning at the end of second stage? ______

7. If you are in Group 3 and your earning in first stage is 20 tokens, can you do following things?

(1) Assign 8, 8, 6 points to Member A, Member B and Member C in Group 4 respectively.

(i) YES (ii) NO

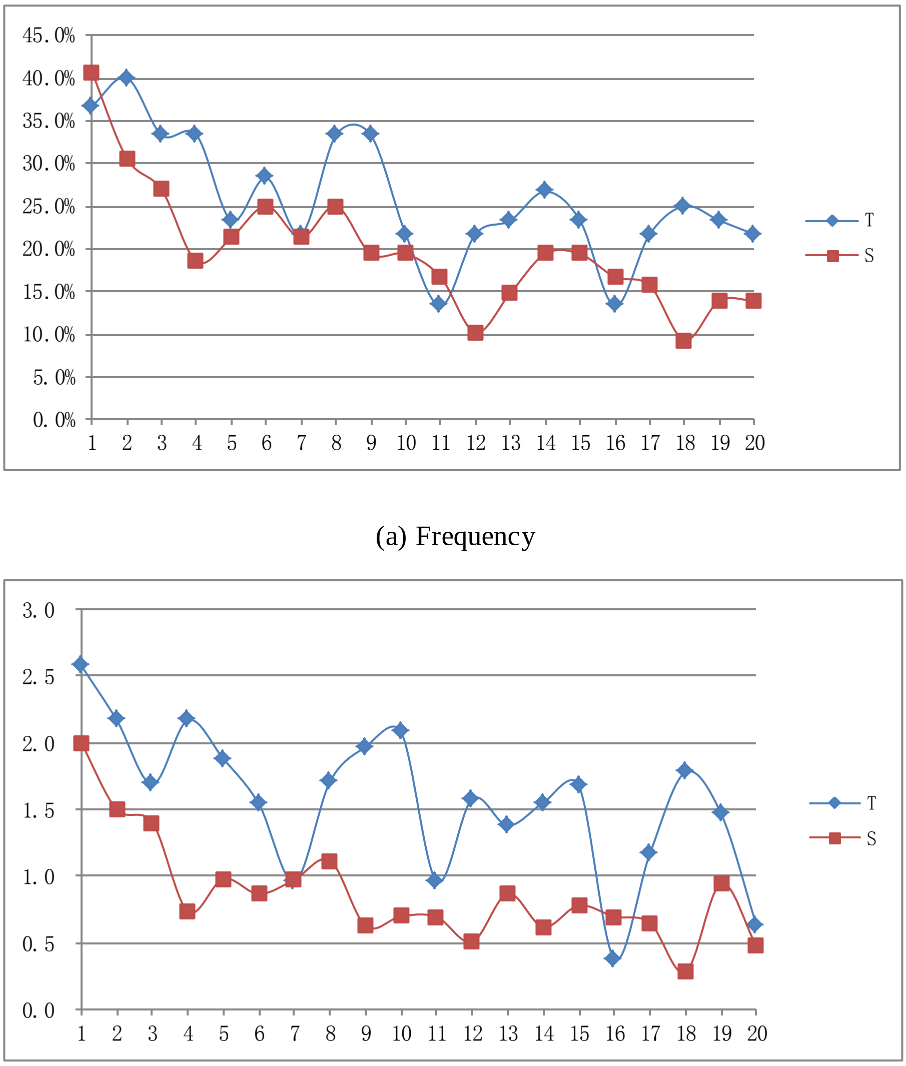
(2) Assign 3, 5, 0 points to Member A, Member B and Member C in Group 4 respectively.

(i) YES (ii) NO

(3) Assign 2, 10, 5 points to Member A, Member B and Member C in Group 4 respectively.

(i) YES (ii) NO

Supplemental Material B: Punishment Dynamics



(a) Frequency

(b) Severity

Figure B1 punishment dynamics: frequency and severity over periods

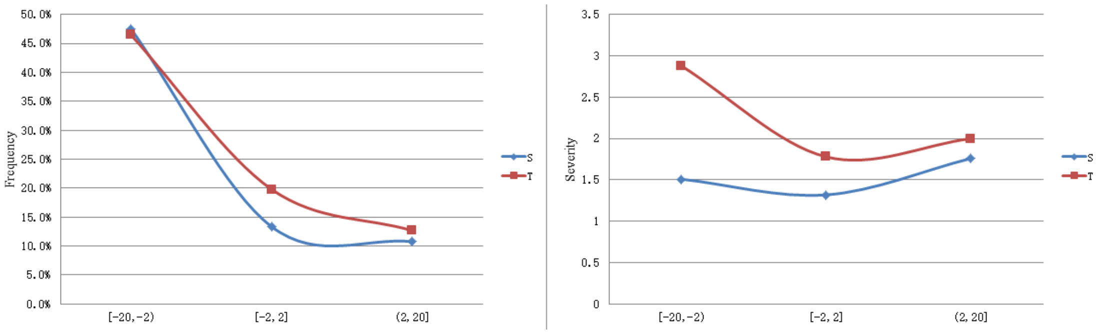


Figure B2 punishment distribution: group average contribution as benchmark

Supplemental Material C: Punishment Decision Model and Punishment Severity Model

We employ the inequity-aversion model of Fehr and Schmidt (*Fehr, E., and K.M. Schmidt. "A Theory of Fairness, Competition, and Cooperation." Quarterly Journal of Economics, 114(3), 1999, 817–68*), the Equity Reciprocity Competition (ERC) model of Bolton and Ockenfels (*Bolton, G.E., and A. Ockenfels. "ERC: A Theory of Equity, Reciprocity and Competition." American Economic Review, 90(1), 2000, 166–93*), and the contribution norm of Reuben and Riedl (2013) to investigate second-party and third-party punishment.

Fehr and Schmidt (1999) show that inequity-averse people suffer from psychological utility loss for both disadvantageous inequality caused by those who contribute less to public goods (namely lower contributors) and advantageous inequality caused by those who contribute more to public goods (namely higher contributors), hence as the second-party and third party they should punish both types of contributors. If we observe both lower and higher contributors are punished, then we find experimental evidence consistent with inequity-aversion model.

The ERC theory of Bolton and Ockenfels (2000) predicts that people with equity preference suffer from utility loss because of the violation of equity, which is measured by the payoff deviation from the group average. Therefore, a second party or a third party with equity preference would punish those who contribute both less and more than the group average. In fact equity preference is conceptually close to inequity-aversion. The difference is that the ERC model emphasizes equity at group level while the inequity-aversion model emphasizes individual inequality[4].

From the perspective of contribution norm (Reuben and Riedl 2013), a person should contribute to the public good as much as possible (efficiency rule) or should contribute a fair amount compared to others' contributions (relative contribution rule). Hence, a third party holding such norm cares about the overall efficiency of the whole group. As the overall efficiency (measured by average contribution) increases, a second party or a third party would become less likely to punish others. If we identify a negative correlation between a group's average contribution level and the average punishment level they received, then we find evidence suggesting that third parties also rely on the contribution norm to punish.

We run separate regressions for punishment decision and punishment severity to test how they are related to factors such as social preference and norm. In the test, *NegDiffI* and *PosDiffI* denote the disadvantageous and advantageous inequality at individual level respectively. *PosDiffG* and *NegDiffG* denote positive and negative violations of equity at group average level respectively. Contribution norm is denoted by *GroOthA*. See Table C.1 for the variable definitions. We include *Period*, the integer period number, to control for the time trend of punishment, as well as some demographic variables such as *gender*, *family wealth*, and *major.*

Table C.1 The definition of variable

[4] For example, A, B, C, and D contribute 20, 15, 5, and 0 tokens respectively. From the perspective of inequity aversion, B, C, and D are lower contributors than A. However, from the perspective of equity, only C and D are lower contributors (lower than average 10). Hence, inequity-aversion model predicts that A would punish B, C, and D, while ERC model predicts that A would punish only C and D.

variables	definition
GroOthA	Average contribution of the other group members (the punished member excluded).
PosDiffG	Positive contribution difference from the punished member to the average contribution of the other group members.
NegDiffG	Negative contribution difference from the punished member to average contribution of the other group members.
PosDiffI	Punished individual's contribution difference to punishing individual if the former contributes more.
NegDiffI	Punished individual's contribution difference to punishing individual if the later contributes more.
P_decision	Whether one person gets punished (0=NO; 1=Yes).
P_received	The amount of punishment received if any.
Punishing	The amount of punishment assigned to others if any
Gender	Male or female (0=male; 1=female).
Major	The student's major (0=non-economics; 1=economics).
Wealth	Economic status (0=rich; 1=poor).

Since punishment decision is binary variable, we use logit regression. Because the data has panel structure, we also run panel regression to double check. Since *PosDiffG* strongly correlates with *PosDiffI*, as well as *NegDiffG* with *NegDiffI*, we exclude *PosDiffG* and *NegDiffG* in our regressions[5]. This is reasonable as inequality-aversion model and ERC model are quite close in term of fairness. We also exclude constant term as incorporating constant term lowers model's performance.

Table C.2 SPP decision model

Independent variables	Logit	Logit (robust)	Logit (cluster: subject)	Logit (cluster: period)	Panel (fixed effect)	Panel (random effect)
GroOthA	-.026	-.026	-.026	-.026	-.001	-.025
	(.041)	(.039)	(.059)	(.044)	(.064)	(.058)
PosDiffI	-.012	-.012	-.012	-.012	.028	.020
	(.045)	(.048)	(.057)	(.054)	(.056)	(.053)
NegDiffI	.536***	.536***	.536***	.536***	.680***	.674***
	(.067)	(.066)	(.069)	(.066)	(.090)	(.086)
Period	-.074***	-.074***	-.074***	-.074***	--	--
	(.014)	(.013)	(.016)	(.018)		
Gender	-.746***	-.746***	-.746***	-.746***	--	-1.074**
	(.187)	(.182)	(.349)	(.147)		(.431)
Wealth	.052	.052	.052	.052	--	-.595**
	(.160)	(.155)	(.307)	(.127)		(.267)
Major	.072	.072	.072	.072	--	.176
	(.274)	(.257)	(.552)	(.248)		(.635)
Observations	720	720	720	720	720	720
log-likelihood	-395.2	-395.2	-395.2	-395.2	-276.1	-374.8
Hausman test	--	--	--	--	*p*=0.557	

Note: *** significant at the 0.01 level; ** 0.05; * 0.10; standard error in bracket.

[5] Their correlation coefficients are greater than 0.97 in our panel sample data,

Table C.3 TPP decision model

Independent variables	Logit	Logit (robust)	Logit (cluster: subject)	Logit (cluster: period)	Panel fixed effect	Panel random effect
GroOthA	.029	.029	.029	.029	.060	.013
	(.041)	(.039)	(.047)	(.043)	(.048)	(.042)
PosDiffI	-.001	-.001	-.001	-.001	.112**	.030
	(.041)	(.042)	(.047)	(.035)	(.050)	(.048)
NegDiffI	.452***	.452***	.452***	.452***	.515***	.479***
	(.065)	(.063)	(.073)	(.070)	(.074)	(.069)
Period	-.066**	-.066**	-.066**	-.066**	--	--
	(.021)	(.022)	(.021)	(.026)		
Gender	-.274	-.274	-.274	-.274	--	-.732
	(.440)	(.462)	(.512)	(.476)		(.475)
Wealth	.108	.108	.108	.108	--	-.058
	(.223)	(.219)	(.228)	(.199)		(.248)
Major	-.794*	-.794*	-.794	-.794*	--	-1.206**
	(.450)	(.461)	(.553)	(.483)		(.497)
Observations	400	400	400	400	400	400
log-likelihood	-208.3	-208.3	-208.3	-208.3	-162.2	-213.2
Hausman test	--	--	--	--	*p*=0.000	

Note: *** significant at the 0.01 level; ** 0.05; * 0.10; standard error in bracket.

Table C.2 and Table C.3 report the regression results of the decision models. Second-party punishment and third-party punishment have similar coefficients on *NegDiffI*. All regressions indicate inequity aversion did affect people's punishment decisions. Namely, players were more likely to punish free riders. This finding is constant with the prediction of inequity aversion model. The coefficient of *GroOthA* are insignificant in all models *PosDiffI* seems to have no significant effect in the process of decision making, probably this is because the majority of punishment is normal ones, that's say, the larger proportion of normal punishment diluted the effect of a smaller proportion of antisocial punishment.

Table C.4 and Table C.5 display the results of the punishment severity regression models. The patterns of third-party and second-party punishment severity are similar. The difference is that third-party punishers assigned heavier punishment to free riders, as suggested by the positive coefficients of *NegDiffI*. The coefficients on *PosDiffI* in SPP and TPP severity models are positive and significant, indicating both SPP and TPP are not completely impartial. However, the coefficient of *PosDiffI* in SPP severity model is slightly higher than in TPP model, indicating that third-party punishment was comparably less antisocially than second-party punishment.. To sum up, the signs and magnitudes of these coefficients generally match the pattern as shown in Figure 2 in the manuscript.

Table C.4 SPP Severity model

Independent	OLS	OLS	OLS	OLS	Panel	Panel	Panel
variables		(robust)	(cluster:	(cluster:	fixed	random	random
			subject)	period)	effect	effect	effect
						(GLS)	(ML)
GroOthA	.004	.004	.004	.004	-.038**	-.031*	.004
	(.013)	(.015)	(.012)	(.021)	(.019)	(.017)	(.013)
PosDiffI	.048***	.048***	.048**	.048***	.029	.031*	.047***
	(.016)	(.017)	(.020)	(.014)	(.020)	(.016)	(.015)
NegDiffI	.099***	.099***	.099***	.099***	.111***	.095***	.099***
	(.015)	(.018)	(.020)	(.018)	(.019)	(.015)	(.015)
Period	.003	.003	.003	.003	--	--	--
	(.005)	(.004)	(.003)	(.004)			
Gender	-.012	-.012	-.012	-.012	--	-.038**	.007
	(.067)	(.058)	(.073)	(.058)		(.066)	(.066)
Wealth	.206***	.206***	.206***	.206***	--	.101	.221***
	(.057)	(.056)	(.039)	(.068)		(.063)	(.052)
Major	.038	.038	.038	.038	--	.074	.028
	(.099)	(.125)	(.111)	(.100)		(.098)	(.098)
Observations	301	301	301	301	301	301	301
R^2	.676	.676	.676	.676	--	--	--
F statistics	--	--	--	--	12.5	--	--
Wald Chi2	--	--	--	--	--	47.7	614.5
Hausman test	--	--	--	--	--	*p*=0.452	n.a.

Note: *** significant at the 0.01 level; ** 0.05; * 0.10; standard error in bracket.

Note that the coefficients of *PosDiffI* is almost the same in both SPP and TPP severity model, not as what is displayed in the right side of Figure 2 in the manuscript. This is because in severity models we incorporated both normal punishment and antisocial cases. After we excluded normal punishment cases, we find the difference becomes obvious as shown in the regression result in Table C.6. The sign of *PosDiffI* in antisocial SPP severity model is positive while it is significantly negative in antisocial TPP severity model.

To test whether blind revenge existed in our sample, we regress punishment in current period on punishment received in last period, controlling for *gender*, *wealth*, *major*, *and other variables that may affect punishment*. If blind revenge does exist, we would expect a subject's punishment spending to be positively related to the punishment received lagged by one period. Table C.7 and Table C.8 report the results for decision model and severity model respectively.

Table C.5 TPP severity model

Independent variables	OLS	OLS (robust)	OLS (cluster: subject)	OLS (cluster: period)	Panel fixed effect	Panel random effect (GLS)	Panel random effect (ML)
GroOthA	.001	.001	.001	.001	.008	.003	-.002
	(.017)	(.017)	(.020)	(.022)	(.022)	(.020)	(.016)
PosDiffI	.046**	.046***	.046***	.046***	.047	.050**	.047**
	(.023)	(.015)	(.014)	(.014)	(.030)	(.024)	(.023)
NegDiffI	.222***	.222***	.222***	.222***	.227***	.225***	.222***
	(.016)	(.020)	(.016)	(.017)	(.020)	(.017)	(.016)
Period	-.007	-.007	-.007	-.007	--	--	--
	(.009)	(.008)	(.008)	(.008)			
Gender	.161	.161	.161	.161	--	.226	.131
	(.192)	(.171)	(.112)	(.217)		(.298)	(.185)
Wealth	.043	.043	.043	.043	--	.061	.038
	(.100)	(.098)	(.093)	(.101)		(.115)	(.099)
Major	.017	.017	.017	.017	--	.077	-.009
	(.206)	(.183)	(.118)	(.217)		(.293)	(.200)
Observations	215	215	215	215	215	215	215
R^2	.767	.767	.767	.767	--	--	--
F statistics	--	--	--	--	53.1	--	--
Wald Chi2	--	--	--	--	--	208.5	706.9
Hausman test	--	--	--	--	--	*p*=0.905	n.a.

Note: *** significant at the 0.01 level; ** 0.05; * 0.10; standard error in bracket.

Table C.6 Second-party and third-party antisocial punishment severity models

	Second-party antisocial punishment			Third-party antisocial punishment		
Independent variables	OLS	OLS (robust)	Panel random effect (ML)	OLS	OLS (robust)	Panel random effect (ML)
GroOthA	.050***	.050***	.028	-.011	-.011	-.021
	(.016)	(.015)	(.019)	(.020)	(.020)	(.019)
PosDiffI	.031*	.031**	.019	-.042**	-.042***	-.040**
	(.016)	(.015)	(.017)	(.023)	(.015)	(.020)
Period	.006	.006	--	-.015	-.015	--
	(.006)	(.006)		(.010)	(.008)	
Gender	-.052	-.052	.008	.946***	.946***	.911***
	(.091)	(.090)	(.112)	(.197)	(.230)	(.193)
Wealth	.098	.098	.217**	.209*	.209*	.187*
	(.078)	(.066)	(.088)	(.107)	(.110)	(.105)
Major	.130	.130	.082	.678***	.678***	.668***
	(.135)	(.190)	(.172)	(.209)	(.243)	(.205)
Observations	154	154	154	129	129	129
R^2	.612	.612	--	.604	.604	--
Wald Chi2	--	--	117.1	--	--	190.9

Note: *** significant at the 0.01 level; ** 0.05; * 0.10; standard error in bracket.

Table C.7 Test for blind revenge decision

Punishment decision	Second-party blind revenge decision			Third-party blind revenge decision		
	Logit	Panel fixed effect	Panel random effect	Logit	Panel fixed effect	Panel random effect
P_received(-1)	.021**	.193**	.199***	-.021	-.001	-.008
P_received	-.079	-.110	-.094	-.048	-.023	-.031
Punishing(-1)	.519***	-.005	.113	.126***	-.002	.053
Period	-.054***	--	--	-.036**	--	--
Gender	.609***	--	1.118*	-14.572	--	-18.11
Wealth	-.640***	--	-1.266**	.704***	--	.913**
Major	.617**	--	1.031	-14.90	--	-18.386
constant	.152	--	.185	13.750	--	16.600
Observations	720	720	720	480	480	480
log-likelihood	-402.6	-252.1	-352.3	-293.6	-219.2	-283.5
Hausman test	--	Fail		--	P<.01	

Note: *** significant at the 0.01 level; **0.05; *0.10. "*(-1)*" means lagging one period. Hausman test fails because it didn't meet asymptotic assumptions.

Table C.8 Test for blind revenge severity

Punishment	Second-party blind revenge severity			Third-party blind revenge severity		
severity	OLS	Panel	Panel	OLS	Panel	Panel
		fixed effect	random effect		fixed effect	random effect
P_received(-1)	.122	.115	.119	-.155*	-.058	-.141
P_received	.091	.079	.108	-.176**	-.033	-.137
Punishing(-1)	.476***	-.32	.479***	.568***	-.023	.577***
Period	-.031	--	--	-.087**	--	--
Gender	.334	--	.309	-1.452***	--	-1.469***
Wealth	-.534*	--	-.561**	.593	--	.450
Major	-.167	--	-.174	Omitted	--	Omitted
constant	2.201***	2.198***	1.940***	3.518***	4.035***	2.784***
Observations	270	270	270	195	195	195
Hausman test	--	P<.01		--	P<.01	
Adjusted R^2	.335	--	--	.438	--	--
F statistics	--	1.23	--	--	0.28	--
Wald chi square	--		139.7	--		155.9

Note: *** significant at the 0.01 level; **0.05; *0.10. "*(-1)*" means lagging one period. "Omitted" means some variables were omitted for collinearity.

The positive coefficient of *P_received(-1)* in *SPP* decision model suggests that some second-party punishers indeed made blind revenge decisions. However, blind revenge effect only existed in the *SPP* decision model but not in the severity model, i.e. punishment points received in the previous period influenced the decision of whether to punish, but not the punishment severity. Why blind revenge takes this pattern leads to an open question.

Supplemental Material D

Negative experience, emotion and punishment: a simple online quasi-experiment

Although the argument that negative experience and emotion associate with stronger punishment is quite intuitive, we did an additional simple online quasi-experiment to provide solid evidence. Two treatments, *Baseline* and *Experience* were designed under between-subject protocol.

In *Baseline*, subjects were lead to imagine that they were stolen 1000 dollars by a thief (framed as younger twin brother). The thief was caught. Subjects were given the power to punish the thief. Subjects were asked how much angry they felt and how many months they would like to put the thief in jail (i.e., second-party punishment). The *Experience* treatment was framed similarly to Baseline, except for adding that subjects couldn't punish the first thief who stole from them, but observed a second thief (the older twin brother) stealing from someone else. Subjects in *Experience* were asked their emotion state and how many months they would like to put the second thief in jail (i.e., third-party punishment). By this design, the subjects first had negative experienced and then again observed similar negative experience. We expect punishment in *Experience* to be heavier than *Baseline*, as the more negative Experience would trigger more negative emotions.

The quasi-experiment was constructed by using a convenient online questionnaire platform, a company known as Wen Juan Xing (http://www.wenjuan.com/). Methodologically, we conducted this quasi-experiment in a similar way as suggested by Horton et al (Horton, J. J., Rand, D. G., and Zeckhauser, R. J., "The online laboratory: conducting experiments in a real labor market" , *Experimental Economics, 14(3), 2011: 399-425*). First, we prepared the experiment via Wen Juan Xing; second, we posted online recruitment ads, and then the experiment was conducted automatically; thirdly, we collected data which were summarized automatically by Wen Juan Xing; fourthly, we paid subjects.

The quasi-experiment was conducted on Jing Guan Zhi Jia (http://bbs.pinggu.org/) on 27-28 Jan, 2016, a Chinese online platform where people could view, contribute, share, buy, and sell their resources, knowledges, and effort. Most of the users on the platform are students, teachers, and researchers. Each subject was paid 10 *Luntanbi,* which is the currency used on the platform. People can use *Luntanbi* to pay others for completing online surveys, finding materials, analyzing data, writing report, etc. From the perspective of the online market, Jing Guan Zhi Jia is similar to Amazon Mechanical Turk, though on a smaller scale. Totally 80 subjects took part in this quasi-experiment. We excluded 4 extreme cases, resulting 37 subjects in *Baseline* and 39 subjects in *Experience*, 35 females and 41 males.

Table D.1 displays the basic statistics. As expected, the average emotion level in *Baseline* was lower than the average level in *Experience* (8.0 vs. 7.5), though the t-test does not show significant difference. However, we observe much stronger punishment in Treatment (8.6 vs. 3.9).

Table D.1 Baseline vs. Treatment (Experience)

Treatments	Emotion (anger)	Punishment (month in jail)
Baseline	7.5 (0.31)	3.9 (0.85)
Experience	8.0 (0.28)	8.6 (1.42)
T-test	*p*=.246	*p*=.005

Note: standard error in bracket

OLS regression results show that punishment level is significantly positively correlated with anger, after controlling for other factors, as shown in Table D.2. When the emotion of anger is stronger, the heavier is the punishment.

Table D.2 Regression

Independent variables	Dependent variable: *Punishment*
Anger	1.009**
Treatment Dummy (0=baseline, 1=treatment)	4.435***
Gender (1=male, 2= female)	-.541
Past experience of being stolen (0=No, 1= Yes)	-3.219*
Constant	-12.175*
Adjusted R^2	.139***

Note: *** significant at the 0.01 level; ** 0.05; * 0.10.

The results suggest more negative experience triggers negative emotions, which triggers heavier punishment. This finding further implies experience and observation of more free riding could lead to heavier punishment.