ON THE DESIGN OF PEER PUNISHMENT EXPERIMENTS

Marco Casari *

Universitat Autònoma de Barcelona

Abstract:

We discuss how technologies of peer punishment might bias the results that are observed in experiments. A crucial parameter is the "fine-to-fee" ratio, which describes by how much the punished subjects income is reduced relatively to the fee the punishing subject has to pay to inflict punishment. We show that a punishment technology commonly used in experiments embeds a variable fine-to-fee ratio and show that it confounds the empirical findings about why, whom, and how much subjects punish.

Keywords: sanctions, public goods, cooperation, experiments

JEL Classification: C91, C92

* I want to thank Dirk Engelman, Ernst Fehr, Gianandrea Staffiero, and two anonymous referees for their comments. Any remaining errors are mine. We gratefully acknowledge the support of a Marie Curie Individual Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect nor those of the European Commission.

ON THE DESIGN OF PEER PUNISHMENT EXPERIMENTS

Several experimental studies have convincingly argued that the cooperation level of a group of agents in prisoner's dilemma-type games can be dramatically increased when each agent has a costly opportunity to punish others in the group (beginning with Ostrom et al., 1992, Fehr and Gaechter, 2000). Agents do punish others and that can push cooperation up to nearly efficient levels. The power of this mutual sanctioning is even more remarkable because it is achieved also with a random matching protocol among subjects, where incentives for building an individual reputation are very low. These studies point to the willingness of many agents to pay a small fee in order to lower earnings of others by a larger amount (fine). There is evidence that this attitude is subject to a "price effect": the smaller the fee necessary to produce a given amount of punishment (fine), the higher is the empirical frequency of requests to punish (Ostrom et al., 1992, Carpenter, 2002, Andreoni et al., 2003, Putterman and Anderson, 2003).

There are competing explanations on what drives people to punish. Among them, scholars have mentioned inequality-aversion, emotions, reciprocity, confusion, spite, and social norms. Moreover, there have been investigations about second-order effects on punishment behavior, for instance regarding the effect of group size or deviations from the group average contribution. *We argue that a specific punishment technology that is widely used in the literature is unfit for both these tasks.* The reason being that it embeds a variable fine-to-fee ratio, which is known to strongly influence demand for punishment and hence it confounds, sometimes fundamentally, the interpretation of the experimental results.

First, we present a typology of punishment technologies (Section I), which allows to better characterize the specific punishment technology under scrutiny. Then in Sections II and III we report of few experimental studies that have adopted this type of technology (Fehr and Gaechter, 2000, Bowles et al., 2001, Carpenter, 2002, Masclet et al., 2003, Nikisforakis, 2004) and use

Fehr and Gaechter (2000) data to illustrate the impact of the fine-to-fee ratio. Conclusions and implications for future research are then spelled out (IV).

I. A TYPOLOGY OF PUNISHMENT TECHNOLOGIES

Consider the following stylized punishment game with two agents. Each period is divided into two stages. Agents at the first stage make simultaneous decision on whether to cooperate or defect in a prisoner's dilemma situation. One could replace a prisoner's dilemma situation with other social dilemma situation such as the voluntary contribution to a public good or the appropriation of a common-pool resource. Before the second stage, each agent learns about the action taken by the other and then has the opportunity to punish.

Experimental studies with decentralized punishment technologies - where subjects have the opportunity to pay a cost c to reduce earnings of others by an amount R - have shown that subjects do punish others. Moreover, the evidence suggests that the frequency of the sanctions is related to the fine-to-fee ratio $\vartheta = R/c$ (Ostrom et al., 1992, Carpenter, 2002, Andreoni et al., 2003, Putterman and Anderson, 2003). All other things equal, the "demand" for punishment is higher when by paying \$1 one can reduce others' earnings by \$4 (ϑ=4) compared to a situation where the fine-to-fee ratio is 1:2 (ϑ=2).

We now introduce a typology of punishment technologies using the 2x2 first stage of the punishment game. The classification in four types reported in Table 1 is somewhat arbitrary but useful to identify how a manipulation of the fine-to-fee ratio can have predictable consequences on aggregate cooperation levels. In particular, it will help to clarify the two dimensions in which the punishment technology introduced by Fehr and Gaechter (2000) is biased.

[Table 1 about here]

The first distinction is between neutral and non-neutral punishment technologies (Type A versus all others in Table 1). A neutral technology allows an agent to punish the other with a

constant fine-to-fee ratio in all circumstances. For instance, a defector can punish a cooperator with the same efficacy that a cooperator can punish a defector. Ostrom et al. (1992) report the example of a fisherman that could damage overnight the nets or boat of another fisherman. The cost to damage the boat of a fisherman that has overused the village fishery is the same than the cost to damage any other boat. Several experimental studies have implemented a neutral punishment technology (Fehr and Gaechter, 2002, Sefton et al., 2002, Page et al., 2002, Carpenter, 2002, Andreoni et al., 2003, Putterman and Anderson, 2003).

A punishment technology is non-neutral when the fine-to-fee ratio varies with the first-stage action of the punisher or of any other agent in the group. An important type of non-neutral technology is when only defectors can be punished but not cooperators (Type B in Table 1). This feature is a characteristic of legal sanctioning systems. If you file a case against a user of a common forest that is known to have fully complied with the established appropriation rules, he will not be convicted to pay fines. Nevertheless, a defector can bring to court another defector and get it punished (Casari and Plott, 2003).

If they do expect a milder punishment, free riders may increase cooperation. That would be the case under a legal system even when punishment is just due to confusion or trembling hand. The aggregate result would be a higher cooperation level. A neutral punishment system would raise aggregate cooperation only when punishment is intentionally directed toward free riders. Alternatively, when subjects are responsive to the "price" of punishment, an increase in aggregate cooperation rate may be generated by a simple positive differential in the fine-to-fee ratio between defectors and cooperators. Hence, any sanctioning system with differential "pricing" as the legal type B of Table 1 is biased toward promoting group cooperation.

A different logic applies to a punishment by cooperators system (Type C in Table 1). It is non-neutral because a defector has no opportunity to inflict punishment. In a sense, you need to be virtuous, a cooperator, to have the opportunity to punish others. In the context of a public good

game, a free rider that wants to punish another agents must first improve the group surplus by contributing and only then she can inflict a punishment.[1]

Finally, under social conformity only deviant behavior could be sanctioned (Type D in Table 1). An example could be workers organizing a strike or board members of a company proposing and deciding on their own pay increase (Hung and Plott, 2001). When the two agents either both cooperate or both defect, nobody can be punished.

All sanctioning systems in this game can be obtained as a linear combination of these four basic types.

II. THE FG TECHNOLOGY

The FG punishment technology (Fehr and Gaechter, 2000) is non-neutral[2] because it includes elements of a legal sanctioning system and of a punishing by cooperator system. Its implicit fine-to-fee ratio – the amount of punishment inflicted on the punished subject, relative to the cost for the punisher – is not constant but depends from the first stage choices of each one of the agents in a group. In this section we spell out the implications of a variable fine-to-fee ratio, reanalyze the FG data, and show how this approach could change the interpretation of the observed receive punishment (Figure 1).

The first-stage game of Fehr and Gaechter (2000) is a voluntary contribution to a linear public good in groups of four persons. In one treatment, group composition was changed after each round by randomly matching the 24 participants ("Stranger") while in another treatment groups were stable across the ten rounds ("Partner"). In the second-stage each agent could assign punishment points p_i^j to others in the group. The costs of punishment points, $c(p_i^j)$, for the punishing subject i are shown in Table 2.

[Table 2 about here]

The first-stage earnings of the punished subject *j* are lowered by 10% for each punishment point received, P^i. Then, individual payoffs are as follows:

(1) $\pi_i = Max\{ 0,\ (20 - g_i + a \cdot \Sigma_{j=1\ldots N}\ g_j) \times [1 - (1/10)P^i] \} - \Sigma_{j \neq i}\ c(p_i^{\,j})$

where a=0.4, N=4, the individual contribution to the public good is $g_i \in [0, 20]$, P^i are the punishment points received by agent *i*, and $p_i^{\,j}$ are the punishment points given by agent *i* to agent *j*. When agent *i* punishes agent *j*, the fine-to-fee ratio ϑ is defined for any $c(p_i^{\,j}) \neq 0$ as:

(2) $\vartheta = R/c = (1/10)\ p_i^{\,j}\ (20 - g_i + a \cdot \Sigma_{j=1\ldots N}\ g_j) / c(p_i^{\,j})$

which clearly varies according to the contribution level of agent *i*, g_i, and of the contributions of *any* of the three other agents in the group.

We introduce some simplifications that do not alter the substance of the FG punishment game and allow a comparison with the framework in Table 1. First, consider only two first-stage strategies, either full contribution, g_i=20, or no contribution, g_i=0. Second, assume that the second-stage decision is whether to assign or not the first punishment point, $p_i^{\,j}$=0 or $p_i^{\,j}$=1. Third, in order to compare a four-person with a two-person game, assume an identical first-stage contribution level among three of the agents. Table 3 shows the associated fine-to-fee ratio of the FG punishment technology.[3]

[Table 3 about here]

This punishment technology is a linear combination of a neutral technology (A), a legal sanctioning system (B), and of a punishing by cooperator system (C):[4]

(3) $$\begin{pmatrix} 2 & 4.4 \\ 0.8 & 3.2 \end{pmatrix} = A + B + C = \begin{pmatrix} 0.8 & 0.8 \\ 0.8 & 0.8 \end{pmatrix} + \begin{pmatrix} 1.2 & 1.2 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 2.4 \\ 0 & 2.4 \end{pmatrix}$$

While embedding a legal sanction bias, punishment in FG is strictly informal, in the sense that punishment does not follow a code of law. We will now relate two empirical findings of FG with the punishment technology decomposition in expression (3). First, FG showed that cooperators frequently use the punishment opportunity to sanction free riders; they carry out the vast

majority of punishment acts, imposing them on free riders. Provided that people are prepared to punish at all, the legal sanction bias of their punishment technology (B) predicts this result. Given the knowledge that punishment frequency is positively correlated with the fine-to-fee ratio, the FG finding might have also another interpretation from the one FG suggests. Our interpretation is largely independent from the detailed motives behind punishment; they could be fairness, reciprocity, spite, or envy. It is simply based on the behavioral correlation between frequency of sanctions and fine-to-fee ratio.

Second, FG conclude that the strength of the punishment levied on the defectors empirically increases in proportion to a defector's deviation from the average contribution of the other group members. Observed punishment patterns were similar in "Partner" and "Strangers". The legal sanction bias of their punishment technology (B) predicts this result as well. In comparison with the average contributor in the group, a lower-than-average contributor will be punished more heavily and a higher-than-average contributor will be punished more lightly.[5] To derive these predictions, we need to generalize scheme (3) for a first-stage game with a continuum strategy space. Consider an average public good contributor that lowers her contribution from $g_i = \overline{g}$ to $g_i = \overline{g} - \Delta$, where $\Delta > 0$. A punishment point assigned to her produces now a decrease in her earnings that is larger than before, $\overline{R} <$ R'. In fact, $\overline{R} = 0.1(20 - \overline{g} + aN\overline{g}) < 0.1(20 - \overline{g} + \Delta + a(N-1)\overline{g} + a(\overline{g} - \Delta)) = R'$ if and only if $0 < a < 1$.[6]

To summarize, provided that people are prepared to punish at all, FG design is biased in favor of the findings claimed in their paper and hence makes their conclusion from this dataset problematic. We know from their other Nature article (Fehr and Gaechter, 2002) that the main results still stand and it would be interesting to compare the two datasets.

To assess whether the fine-to-fee ratio has any explanatory power in the FG data, we reanalyze their data. The fine-to-fee ratio ϑ has a rather strong influence on the amount of received punishment as it can be detected from looking at Figure 1 and at the result of the clustered OLS

regression reported in Table 4.[7] As already shown in other studies, this analysis confirms that the "law of demand" also holds for punishment behavior.

[Figure 1 about here]

[Table 4 about here]

The most important point of the paper is that there is a serious problem of multi-colinearity between the fine-to-fee ratio and other variables of interest of FG that makes it rather problematic to evaluate the relative importance in explaining punishment behavior. The correlation between fine-to-fee ratios and "others' average contribution" is 0.78; with "positive deviations from the other's average contribution" is -0.70; and with "absolute negative deviation" is 0.66.[8]

The same criticism may apply to other studies that have adopted the FG setup (Bowles, 2001, Carpenter, 2002, Masclet et al., 2003, Nikisforakis, 2004). Only a neutral punishment technology, i.e. that maintains a constant fine-to-fee ratio across all conditions can allow to evaluate the impact of these other variables on punishment behavior.

III. GROUP SIZE AND OTHER EFFECTS

Every time first-stage payoff changes, the fine-to-fee ratio of the FG punishment technology (Fehr and Gaechter, 2000) varies. For example, in the linear public good setup of expressions (1), the fine-to-fee ratio increases with higher marginal per capita return of the public good a or with larger group sizes N. That is the case of the experimental study of Carpenter (2002), which aims at measuring the effect of group size on cooperation and punishment behavior using the FG setup. In one of his treatments the marginal per capita return of the public good is a=0.30 and in another is a=0.75. He compares performances of small groups of N=5 with large groups of N=10 and finds no significant differences in punishment patterns.

This comparison is done by implicitly varying the fine-to-fee ratios from 2.5 with N=5 to 4 with N=10. These ratios are computed at a contribution level of half the endowment, g_i=10 for every agent *i* and for *a*=0.30. The analogous ratios for *a*=0.75 are 4.75 with N=5 and 8.5 with N=10. All these variations in fine-to-fee ratios across treatments are in addition to the in-treatment variations already discussed in Section III. His conclusions about the no effect of group size on punishment patterns are conditional on having changed the fine-to-fee ratio by the magnitude specified above.

IV. CONCLUSIONS

We discuss alternative designs of peer punishment experiments and make a methodological contribution. Early studies about punishment (Ostrom et al., 1992, Fehr and Gaechter, 2000) provided support to the persistent tendency of subjects to make use of opportunities to punish others even when that choice was costly. One important factor in explaining the frequency of punishment turned out to be the fine-to-fee ratio – the amount of punishment inflicted on the punished subject relative to the cost of punishment. There likely are other, powerful factors that help to explain why people do punish and who they target. Those factors are yet not entirely clear but we argue that they could be better identified by adopting experimental designs with a constant fine-to-fee ratio.

While there already are many studies that hold the "price" of punishment constant, we show that a specific punishment technology that is used in the literature (Expression (1) and Table 2) does not satisfy this criterion and that is reflected in the empirical results. A re-analyses of Fehr and Gaechter (2000) data shows that the tendency of cooperators to punish free-riders and to do so with more strength in proportion to a defector's deviation from the average contribution of the other group members is confounded with the effect of a variable fine-to-fee ratio. Moreover, this punishment technology is not invariant to changes in group size and marginal per capita return of a public good (Carpenter, 2000).

We have three conclusions. First, one may argue that if the cooperation-enhancing effect of punishment opportunities only holds in the special case of the particular technology, results in previous studies might be less general than previously thought. Although a superficial comparison between Fehr and Gaechter (2000, 2002) shows that some basic findings are not reversed, the virtuous effects of peer punishment were probably enhanced by the use of a special variable fine-to-fee punishment technology. This consideration leads to the second conclusion. When the main research question was whether people punish at all in a one-shot situation – since punishment is always costly and therefore never in the interest of a selfish player – the exact parameters of the punishment technology were somewhat secondary. Instead, when studying either the motivations to punish or the extent to the improvement in group cooperation, a punishment technology with a constant fine-to-fee ratio is definitely recommended, in particular when there is a positive correlation between the variables of interest with the fine-to-fee ratio. Third, studies should explicitly report the fine-to-fee ratio of the adopted punishment technology and explain how that might bias the interpretation of their results.

---

[1] Given that the demand for punishment depends on its "price", the punishment by cooperator technology (C) in Table 1 has an interesting features when N>2. By cooperating an agent provides a second externality to the group: a more powerful sanctioning technology and hence we expect more sanctions as the average group contribution rises. The punishment by cooperator structure could generate a virtuous reinforcement cycle. Suppose in round 0 everybody defects. If in round 1 one agent fully cooperates, she or another agents could find it now attractive to sanction another agent. If, as a result, the targeted agent increases its contribution, the fine-to-fee ratio further increases, and so on. This mechanism is similar to a market where the demand function depends on fashion (the consumer gains more utility if more people buy units of the same good).

[2] Sefton et al. (2002, p.27), Andreoni et al. (2003, footnote 4) and Carpenter (2002, footnote 16) have also mentioned this point.

[3] Table 3 does not consider other two sources of variability in the fine-to-fee ratio that are present in the FG technology. First, the marginal cost of punishment points varies five-folds between the first and the tenth point (Table 2). Second, if the received punishment of agent *i* is above her first stage earnings, the difference is reset to zero (expression (1)). Hence, the marginal punisher may pay a cost without obtaining any reduction on the target agent's earnings.

[4] An alternative decomposition with incentives for social conformity is:

(4) $$\begin{pmatrix} 2 & 4.4 \\ 0.8 & 3.2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1.6 \\ 0 & 3.2 \end{pmatrix} + \begin{pmatrix} 0 & 0.8 \\ 0.8 & 0 \end{pmatrix}$$

[5] One may conjecture that this effect on cooperation will be due to similar motivations than the ones at work in the tax-subsidy scheme studied in Falkinger et al. (2000).

[6] The opposite result follows when an average contributor increases her contribution. The empirical results reported in FG show a lower punishment toward subjects with positive deviations from the average contribution of other group members, although this difference was not statistically significant.

[7] As an aside point to Table 4, there are reasons why the computed fine-to-fee ratio ϑ could have a non-linear relationship with the level of punishment and not linear as it appears in the regression. First, the costs of punishment points is non linear (See footnote 3). Second, the willingness to pay of subjects for punishment points may be concave in quantity. While the second factor is not under the control of the experimenter, the first can be easily adjusted by modifying the punishment technology. A two-term polynomial of the fine-to-fee ratio ϑ generates the following results: Stranger-treatment Adj R-squared = 0.4163 and Partner-treatment Adj R-squared = 0.6085 [software package used: Stata].

[8] The fine-to-fee ratio is dropped from the regression with other three explanatory variables used in FG. To carry out a Variance Inflation Factor analysis one needs to use the natural logarithm of the fine-to-fee ratio (thetalog). The VIF values for the Partner (Stranger)-treatment are thetalog 174.21 (141.59), posdev_avg 49.17 (38.87), avgcothers 40.37 (23.78), negdev_avg 26.79 (18.63).

REFERENCES

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). "The Carrot or the Stick: Rewards, Punishment and Cooperation," *American Economic Review,* 93, 3, 893-902.

Bochet, O., Page, T., Putterman, L. (2002). "Communication and Punishment in Voluntary Contribution Experiments," Brown University, Department of Economics, *Working Papers no. 2002-29*

Bowles, S., Carpenter, J., Herbert G. (2001). "Mutual Monitoring in Teams: the Effects of Residual Claimancy and Reciprocity," *Working paper*.

Carpenter, J. (2002). "The Demand for Punishment," *Working Paper 0243*, Middlebury College, Department of Economics

Carpenter, J. (2002). "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods," *working paper*, Middlebury College.

Casari, M. and Plott, C.R. (2003). "Decentralized Management of Common Property Resources: Experiments with Centuries-Old Institutions," *Journal of Economic Behavior and Organization*, 51, 2, 217-247.

Falkinger, J., Fehr, E., Gächter, S., and Winter-Ebmer, R. (2000). "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence," *American Economic Review*, 90, pp. 247-264.

Fehr, E. and Gächter, S. (2000). "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 2000, 90, 4, pp. 980-994.

Fehr, E. and S. Gaechter (2002). "Altruistic Punishment in Humans," *Nature*, Vol.415, 137-140

Hung, A. and Plott, C.R. (2001). "Information Cascades: Replication and an Extension to Majority Rule and Conformity Rewarding Institutions," *American Economic Review,* 91, 5, 1508-20.

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval (2003). "Monetary and Nonmonetary Punishment in the Voluntary Contribution Mechanism," *American Economic Review*, 93, 1, 366-380.

Nikiforakis, N. S. (2004). "Punishment and Counter-punishment in Public Goods Games: Can we still govern ourselves?," March, University of London, Royal Holloway, Department of Economics, *Working paper*

Ostrom, E., Walker, J., and Gardner, R. (1992). "Covenants with and without a sword: self-governance is possible," *American Political Science Review*, 86, pp. 404-417.

Page, T., Putterman, L., Unel, B. (2002). "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency," Brown University, Department of Economics, *Working Papers no. 2002-19*.

Putterman, L. and Anderson, C. M. (2003). "Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism," Brown University, Department of Economics, *Working Papers no. 2003-15*.

Sefton, M., Shupp, R., and Walker, J.. (2002). "The effect of rewards and sanctions in provision of public goods," *CEDEX Working paper no. 2002-2*.

Table 1: A typology of punishment systems for 2x2 games

		Punishing agent							
		(A) NEUTRAL PUNISHMENT TECHNOLOGY		(B) LEGAL SANCTIONS		(C) PUNISHMENT BY COOPERATORS		(D) SOCIAL CONFORMITY	
		Defects	Cooperates	Defects	Cooperates	Defects	Cooperates	Defects	Cooperates
Agent target of punishment	Defects	ϑ	ϑ	ϑ	ϑ	0	ϑ	*0*	ϑ
	Cooperates	ϑ	ϑ	0	0	0	ϑ	ϑ	*0*

Note: ϑ is the fine-to-fee ratio of the punishment technology available for use

Table 2: Punishment levels and associated costs for the punishing subject

Punishment points p_i^j	0	1	2	3	4	5	6	7	8	9	10
Costs of punishment $c(p_i^j)$	0	1	2	4	6	9	12	16	20	25	30

Table 3: Fine-to-fee ratio of Fehr and Gaechter punishment technology

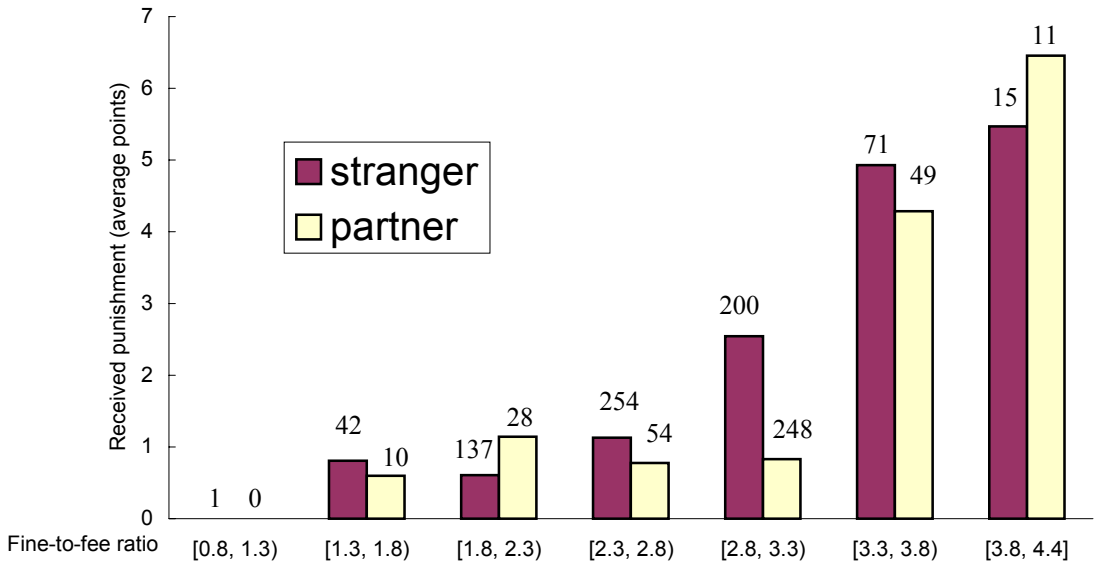
		All other agents	
		Defect	Cooperate
Agent *j*	Defects	*2*	*4.4*
target of punishment	Cooperates	*0.8*	*3.2*

Table 4: Determinants of getting punished

	Dependent variable: Received punishment points	
Independent variables:	Stranger-treatment	Partner-treatment
Constant	-3.92931 **	-4.540013 ***
	(.7102797)	(.6356664)
Fine-to-fee ratio ϑ	2.443645 **	2.929707 ***
	(.3117793)	(.3196917)
	Adj R-squared = 0.3726	Adj R-squared = 0.5440
	Number of obs = 720	Number of obs = 400

Note: Robust standard errors are in parenthesis. The OLS regression is clustered by session in the Stranger-treatment and by group in the Partner-treatment. * denotes significance at the 10-percent level, ** at the 5-percent level, and *** and the 1-percent level. To control for time and matching groups, the regression model also contains period dummies and dummies for matching groups (not reported).

Figure 1: Received punishment points for different fine-to-fee ratios



Note: Number of observations on top of each column.