SUPPLEMENTARY MATERIALS
for online Publication

Markussen, Putterman and Tyran (2012):

Self-Organization for Collective Action

An Experimental Study of Voting on Sanction Regimes

Appendix A - Experimental instructions

This appendix reproduces instructions for DC treatment (in condensed layout). Instructions for other treatments are available upon request from the authors.

Welcome

You are now taking part in an economic experiment. Depending on your decisions and the decisions of other participants, you will be able to earn money. How you can earn money is described in these instructions. Please read them carefully. You will have to answer control questions, to check that you understand the instructions. You can only continue the experiment when you have answered these questions correctly.

During the experiment you are not allowed to communicate with other participants. If you have a question, raise your hand. One of us will come to answer your question. Sometimes you may have to wait a short while before the experiment continues. Please be patient.

During the experiment your earnings will be calculated in points. Points will be converted to Danish kroner at the following rate: 5 points = 1 DKK.

At the end of the experiment your total earnings will be paid out to you in cash.

The experiment has seven phases (that means, seven sets of 4 periods, in total, 28 periods). The following instructions explain the details of phase 1. The details of the subsequent phases will be explained later.

Instructions for Phase 1

In the experiment, all participants are randomly divided into **groups of 5**. This means that you are in a group with four other participants. **You will be part of the same group throughout the entire experiment**. Nobody knows which other participants are in their group, and nobody will be informed who was in which group after the experiment.

Phase 1 is divided into 4 periods. In each period, each group member, yourself included, will be given an **endowment of 20 points**. In each period you will have to make one decision.

Your decision

You and the four others in your group simultaneously decide how to use the endowment. There are two possibilities:

1. **You can allocate points to a group account.**
2. **You can allocate points to a private account.**

You will be asked to indicate the number of points you want to allocate to the group account. Only integers between 0 and 20 are allowed for this purpose. The remaining points will automatically be allocated to your private account. Your earnings depend on the total number of points in the group account, and the number of points in your private account.

How to calculate your earnings

Your earnings from your private account are equal to the number of points you allocate to it. That is, **for each point you allocate to your private account you get 1 point as earnings**. For example, your earnings from the private account equal 3 points if you allocate 3 points to it. The points you allocate to your private account do not affect the earnings of the others in your group.

Your earnings from the group account equal the **sum** of points allocated to the group account by all 5 group members multiplied by 0.4. **For each point you allocate to the group account you and all others in your group each get 0.4 points as earnings**. For example, if the sum of points in the group account is 30, then your earnings from the group account and the earnings of each of the others in your group from the group account are equal to 12 points.

Your earnings can be calculated with the following formula:

20 – (points you allocated to the group account) + 0.4 * (sum of points allocated by all group members to the group account)

Note that you get 1 point as earnings for each point you allocate to your private account. If you instead allocate 1 extra point to the group account, your earnings from the group account increase by 0.4 * 1 = 0.4 points and your earnings from your private account decrease by 1 point. However, by allocating 1 extra point to the group account, the earnings of the other 4 group members also increase by 0.4 points. Therefore, the total group earnings increase by 0.4 * 5 = 2 points. Note that you also obtain earnings from points allocated to the group account by others. You obtain 0.4 * 1 = 0.4 points for each point allocated to the group account by another member.

Example: Suppose you allocate 10 points to the group account, the second and third members of your group each allocate 20 points to the group account, and the remaining two individuals allocate 0 points each. In this case, the sum of points in the group account is 10 + 20 + 20 + 0 + 0 = 50 points. Each group member gets earnings of 0.4 * 50 = 20 points from the group account.

Your total earnings are: 20 – 10 + (0.4 * 50) = 10 + 20 = 30 points. The second and third members' earnings are: 20 – 20 + (0.4 * 50) = 0 + 20 = 20 points. The fourth and fifth members' earnings are: 20 – 0 + (0.4 * 50) = 20 + 20 = 40 points.

Do you have any questions? (Please raise your hand.)

Instructions for Phase 2 to 7 *(for treatment DC)*

Please read these instructions carefully. Again, you will have to answer control questions to check that you understand the instructions.

The next six phases are like the previous one in that you continue to interact with the same four individuals and in each period you make a decision about allocating 20 points to either a private account or a group account. The earnings consequences of your decisions are also as before.

However, there will now be **three** different **rule sets**, two of them new to these phases, which affect your earnings in different ways. In each phase, your group will use **one** of these rule sets.

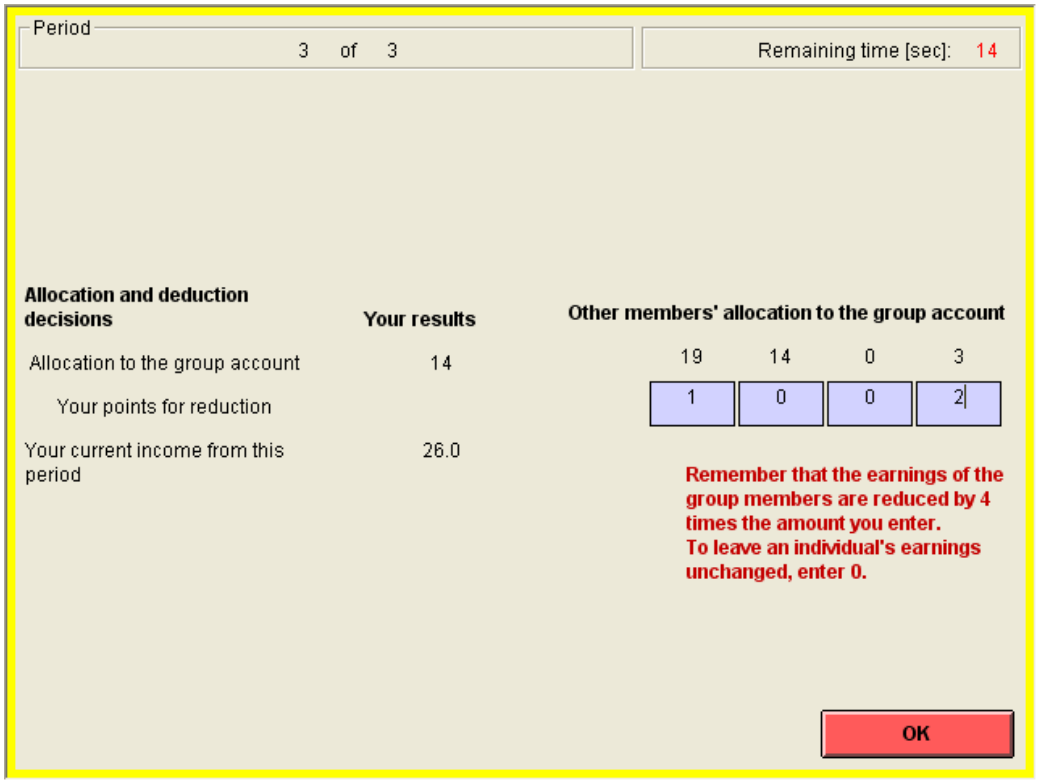
Now, we describe the three rule sets.

RULE SET 1 (*no point reductions)*: In rule set 1, earnings are determined in exactly the same way as in Phase 1 of the experiment.

RULE SET 2 (*individual point reductions)*: In rule set 2, there are two stages in each period. In the first stage, you make your allocation decision and learn the decisions of the other group members along with your earnings. In the second stage, you have an opportunity to reduce the earnings of others in your group at a cost to you. Here is how it works.

After the first stage of each period, you will be shown the amount allocated to the group account by each of the others in your group, **in a random order**, and in a box below that information you will be asked to enter a number of points (if any) that you wish to use to reduce the earnings of the individual who made that allocation decision (see below). Each point you allocate to reducing another's earnings **reduces your own earnings by 1 point** and **reduces that individual's earnings by**

4 points. Your own earnings can be reduced in the same way by the decisions of others in your group. You are free to leave any or all others' earnings unchanged by entering 0s in the relevant boxes.



Note: Numbers shown are for illustration only.

Earnings in each period are calculated as follows:

20 – (points you allocate to group account)
+ 0.4*(sum of points allocated by all in group to group account)
– (points you use to reduce others' earnings)
– 4*(sum of reduction points directed at you by others in your group),

For example, suppose that you use 0 points to reduce the earnings of the first and second group members whose allocations appear on the screen, you use 1 point to reduce the earnings of the third, and you use 2 points to reduce the earnings of the fourth. Suppose further that these individuals use 0, 1, 0 and 3 points to reduce your earnings. Then the third and fourth individuals' earnings for the period will be reduced by 4 and by 8 points, respectively, in addition to any reductions due to the decisions of others. Your own earnings for the period will be reduced by 3 points = your cost of imposing reductions on others, plus (1x4)+(3x4)=16 points = the reductions imposed on your earnings by others. At the end of the reduction stage, you will learn that your earnings were reduced by others by a total of 16 points, but you will not be told which individuals reduced your earnings or by how much any given individual reduced your earnings. Others will also not know who in particular reduced their earnings by how much.

The earnings reduction process is subject to two limits. First, you cannot assign more than 10 reduction points to any one individual in your group. Second, the total effective reduction of your earnings due to others' decisions in a given period cannot be greater than your total earnings from the allocation stage of that period. For example, if your earnings after the allocation stage are 26 points and if others use a total of 7 points to reduce your earnings, you will lose only 26 points, not

7x4 = 28. However, the points that you spend to reduce the earnings of others are always costly to you, even if that brings your earnings for a period to less than zero. To continue with the example in which you earn 26 points before reductions and you lose 26 points due to the (28 points worth of) reductions others impose on you: if in the same period you have chosen to spend 3 points on reducing others' earnings, your total earnings for the period are – 3. Points lost in some periods are deducted from your accumulated earnings of other periods.

Remember that if no reductions are imposed (the reduction boxes are filled in with 0's), earnings after the reduction stage are the same as those before it.

RULE SET 3 (*automatic point reductions*): In rule set 3, each individual pays a fixed fee of **two** points in each period. The fee is deducted from your earnings **at the end of the period**. In addition, each individual pays a fine equal to **80 percent** of the amount of points allocated to the **private account.**

Earnings in each period are calculated as follows:

20 – (points you allocate to group account) + 0.4*(sum of points allocated by all in group to group account) – 0.8*(points you allocate to private account) – 2

For example, suppose you allocate **10 points** to the group account, the second and third members of your group each allocate 20 points to the group account, and the remaining two individuals allocate 0 points each. In this case, the sum of points in the group account is 10 + 20 + 20 + 0 + 0 = 50 points. Each group member gets earnings of 0.4 * 50 = 20 points from the group account.

Your total earnings are: 20 – 10 + (0.4 * 50) – (0.8*10) – 2 = 10 + 20 – 8 – 2 = 20 points.

The second and third members' earnings are: 20 – 20 + (0.4 * 50) – (0.8*0) – 2 = 0 + 20 – 0 – 2 = 18 points. The fourth and fifth members' earnings are: 20 – 0 + (0.4 * 50) – (0.8*20) – 2 = 20 + 20 – 16 – 2 = 22 points.

Notice that for each point you put in your private account, you gain 0.2 points (that is, you gain 1 point as income and lose 0.8 points in fines); and for each point you put in the group account, you gain 0.4 points.

This table presents an overview of the three rule sets:

RULE SET 1 *(no point reductions)*	RULE SET 2 *(individual point reductions)*	RULE SET 3 *(automatic point reductions)*
Same as in Phase 1	Each group member can reduce other group members' earnings after seeing the allocations of each individual to the group account. It costs 1 point to reduce the earnings of another group member by 4 points	There is a fixed cost of 2 points in each period, deducted from your earnings at the end of the period. Each individual pays a fine equal to 80 percent of the amount of points he or she allocated to the private account.

At the beginning of each phase, your group will **choose between two of these rule sets**, by **voting**. Each individual votes for the rule set that he or she prefers and the rule set receiving the highest number of votes in your group is implemented in the next four periods. At the beginning of the next phase, your group will again choose between two different rule sets, by means of voting.

Do you have any questions? (Please raise your hand.)

Appendix B – Derivation of theoretical predictions

This appendix derives the predictions (for contributions and punishment) under NS (no sanctions), IS (informal sanctions) and FS (formal sanctions), and for voting choices between these conditions. The treatments under formal sanctions differ by whether the sanction *s* is sufficiently strong to deter a rational and self-interested individual from free riding, and by the fixed cost for enacting the formal sanction (See table 1 in the main text).
Section I derives predictions based on the model by Fehr and Schmidt (1999) assuming aversion to inequality, section II based on the model of Charness and Rabin (2002) assuming maximization of social welfare without reciprocity preferences (their Appendix 1). Section III compares the predictions derived from the two accounts. Section IV briefly discusses the version of Charness and Rabin (2002) that assumes reciprocity. All derivations below refer to one-shot games of complete information.

I. Fehr and Schmidt (1999)

The Fehr and Schmidt assume a utility function as follows:

$$U_i\left(\pi\right)=\pi_i-\frac{\alpha_i}{n-1}\sum_{j\neq i}\max\left(\pi_j-\pi_i,0\right)-\frac{\beta_i}{n-1}\sum_{j\neq i}\max\left(\pi_i-\pi_j,0\right)$$

I.1 Contributions

NS: Proposition 4 in Fehr and Schmidt 1999 implies that equilibria with positive contributions exist when all group members are sufficiently averse to advantageous inequality. With our parameters inserted, the proposition reads:

a) $C_i=0\ \forall i$ if $\exists j$ with $\beta_j<0.6$,

b) $C_i=C\in[0,20]$ if $\beta_j\geq 0.6\ \forall i$

Using the distribution of β in Fehr and Schmidt Table III (which is similar to the distribution in Blanco et al. 2011, Table 2), the probability that condition b) is met is 0.4^5 = 1 percent.

IS: Proposition 5 in Fehr and Schmidt 1999 implies that equilibria with positive contributions exist when group members are sufficiently averse to advantageous or disadvantageous inequality, and that equilibria with positive contributions are more feasible in IS than in NS. With our parameters inserted, the proposition reads:

Suppose there are *n'* "conditionally cooperative enforcers" with preferences that obey
1) $\beta_i\geq 0.6$ and
2) $\frac{1}{4}<\frac{\alpha_i}{4\left(1+\alpha_i\right)-\left(n'-1\right)\left(\alpha_i+\beta_i\right)}$

Where *n'* is the number of group members with $\beta_i\geq 0.6$

Then there is a sub-game perfect Nash equilibrium with the following characteristics:
a) Each player contributes $C_i=C\in[0,20]$.
b) No punishment.

If anybody (off the equilibrium path) chooses $C_i < C$, each conditionally cooperative enforcer reduces the deviators earnings *by* $(C-C_i)/(n'- 1/4)$.

Conditions 1) and 2) are potentially met with as little as two "conditionally cooperative enforcers" ($n' = 2$), if these individuals are sufficiently averse to disadvantageous inequality. In NS, cooperative equilibria are only feasible if all group members fulfill $\beta_i \geq 0.6$. Therefore, cooperative equilibria are feasible for a wider range of social preferences in IS than in NS.

What is the probability that conditions 1 and 2 are met, given empirical estimates of the distribution of α and β? Assume, as in Fehr and Schmidt 1999, that all individuals who have $\alpha \geq 1$ also have $\beta \geq 0.6$. Conditions 1 and 2 then boil down to the condition that there must be either i) at least two group members with $\alpha \geq 3.4$, or ii) at least three group members with $\alpha \geq 1.4$ or iii) at least four group members with $\alpha \geq 1$. Following the exact, discrete distribution presented in Fehr and Schmidt, Table III, the probability that conditions 1) and 2) are met is 14.5 percent. However, the data behind the table is consistent with assuming a distribution of social preference parameters, which implies that the probability of meeting conditions 1 and 2) is at least 33.9 percent.[1] These probabilities are much higher than the probability of meeting the conditions for cooperative equilibria in NS, but also much lower than the observed share of groups with high levels of contributions under IS in the experiment.

FS, deterrent: This regime has not been analyzed in Fehr and Schmidt (1999). The following proposition describes behavior under deterrent FS (with our parameter values inserted):

Proposition B1:

a) $C_i = E = 20 \ \forall i$ is always an equilibrium.

b) For a player *i* with $\alpha_i \leq \frac{m}{1-s} - 1 = 1$, $C_i = E$ is a dominant strategy.

c) Let the number of players with $\alpha_i \leq \frac{m}{1-s} - 1$ be denoted by *h*. The remaining $n - h$ players choose C_i = E if they fulfill:

$$\alpha_i \leq \frac{(n-1)(m+s-1)}{(1-s)(n-1-h)} + \frac{\beta_i h}{(n-1-h)} = \frac{0.8}{0.8-0.2h} + \frac{\beta_i h}{4-h} \qquad \text{(B1)}$$

[1] In particular, Fehr and Schmidt estimate that 30 percent of subjects have α=1. If we assume that these individuals instead have α=1.4, the probability of meeting conditions 1) and 2) is 33.9 percent. The estimates of α are based on rejection thresholds in ultimatum games. α = 1 implies a rejection threshold of 0.33 (i.e. a responder only accepts offers of at least a third of the pie). α =1.4 implies a rejection threshold of 0.37. Results from ultimatum games are generally consistent with both of these alternative assumptions; see e.g. Camerer 2003, Table 2.3.

Proof:

Without loss of generality, we ignore the fixed cost of FS.

With our pay-off function, the Fehr and Schmidt utility function is:

$$U_i(C) = (E - C_i)(1-s) + mC_i + m\sum_{j\neq i} C_j - c \\ -\frac{\alpha_i(1-s)}{n-1}\sum_{j\neq i}\max(C_i - C_j, 0) - \frac{\beta_i(1-s)}{n-1}\sum_{j\neq i}\max(C_j - C_i, 0) \quad \text{(B2)}$$

To understand the inequality aversion terms, consider that the difference in pay-off between *i* and *j* is

$$\pi_i - \pi_j = (E - C_i)(1-s) + m\sum_k C_k - c - \left((E - C_j)(1-s) + m\sum_k C_k - c\right) \\ = (C_j - C_i)(1-s)$$

a) When all players contribute *E*, a deviation lowers own pay-off and also generates advantageous inequality (shifting one point from the public to the private account decreases the subject's own earnings by *1-s* and the earnings of each other group member by *m > 1-s*). Therefore, universal, full contribution is always an equilibrium.

b) Consider an arbitrary contribution profile {$C_1, C_2, .., C_n$}. Player 1's utility from choosing $C_1 = E$ is

$$U_1(C_1 = E) = mE + m\sum_{j=2}^{n} C_j - \frac{\alpha_1(1-s)}{n-1}\sum_{j=2}^{n}(E - C_j) \\ = mE + m\sum_{j=2}^{n} C_j + \frac{\alpha_1(1-s)}{n-1}\sum_{j=2}^{n} C_j - \alpha_1(1-s)E$$

Note that $0 \le \beta_i \le \alpha_i$. Assume that at least one player other than 1 chooses $C_j < E$. Let players be labeled in descending order according to contribution levels. Player 1's utility from choosing $C_1 \in [C_{k+1}; C_k]$ is:

$$U_1(C_1 < E) = (E - C_1)(1-s) + mC_1 + m\sum_{j=2}^{n} C_j - \frac{\alpha_1(1-s)}{n-1}\sum_{j=2}^{k}(C_i - C_j) - \frac{\beta_1(1-s)}{n-1}\sum_{j=k+1}^{n}(C_j - C_i) \\ \le (E - C_1)(1-s) + mC_1 + m\sum_{j=2}^{n} C_j - \frac{\alpha_1(1-s)}{n-1}\sum_{j=2}^{k}(C_i - C_j) + \frac{\alpha_1(1-s)}{n-1}\sum_{j=k+1}^{n}(C_j - C_i) \\ = (E - C_1)(1-s) + mC_1 + m\sum_{j=2}^{n} C_j + \frac{\alpha_1(1-s)}{n-1}\sum_{j=2}^{k} C_j - \alpha_1(1-s)C_1 \\ = U(C_1 = E) + (E - C_1)(1-s) + mC_1 - \alpha_1(1-s)C_1 - mE + \alpha_1(1-s)E \\ = U(C_1 = E) + (E - C_1)((1-s) - m + \alpha_1(1-s))$$

Hence, player 1 deviates to $C_1 = E$ if

$$(1-s) - m + \alpha_1(1-s) \le 0 \Leftrightarrow$$
$$\alpha_1 \le \frac{m}{(1-s)} - 1$$

which proves the claim.

c) Assume that h players have $\alpha_i \le \frac{m}{1-s} - 1$. Part b) shows that these players choose $C_j = E$.

Player i's utility from choosing $C_i = E$ is

$$U_i(C_i = E) = m(h+1)E + m\sum_{j=i+1}^{n} C_j - \frac{\alpha_i(1-s)}{n-1}\sum_{j=i+1}^{n}(E - C_j)$$
$$= m(h+1)E + m\sum_{j=i+1}^{n} C_j + \frac{\alpha_i(1-s)}{n-1}\sum_{j=i+1}^{n} C_j - \frac{\alpha_i(1-s)(n-i)}{n-1}E$$

Again, let players be ordered by contribution level. Assume that player i chooses the highest, contribution level strictly below E, $E = C_{i-1} > C_i \ge C_{i+1}$. Note that $i-1 = h$. Player i's utility is then:

$$U_i(C_i) = (E - C_i)(1-s) + mC_i + mhE + m\sum_{j=i+1}^{n} C_j - \frac{\alpha_i(1-s)}{n-1}\sum_{j=i+1}^{n}(C_i - C_j) - \frac{\beta_i(1-s)}{n-1}\sum_{j=1}^{i-1}(E - C_i)$$
$$= (E - C_i)(1-s) + mC_i + mhE + m\sum_{j=i+1}^{n} C_j$$
$$-\frac{\alpha_i(1-s)(n-i)}{n-1}C_i + \frac{\alpha_i(1-s)}{n-1}\sum_{j=i+1}^{n} C_j - \frac{\beta_i(1-s)(i-1)}{n-1}(E - C_i)$$

Deviation to E is optimal if:

$$U_i(E) \ge U_i(C_i) \Leftrightarrow mE - (E - C_i)(1-s) - mC_i +$$
$$\frac{\beta_i(1-s)(i-1)}{n-1}(E - C_i) - \frac{\alpha_i(1-s)(n-i)}{n-1}(E - C_i) \ge 0 \Leftrightarrow$$
$$\frac{\beta_i(1-s)(i-1) - \alpha_i(1-s)(n-i)}{n-1} \ge 1 - s - m \Leftrightarrow$$
$$\frac{\beta_i(1-s)h - \alpha_i(1-s)(n-1-h)}{n-1} \ge 1 - s - m \Leftrightarrow$$
$$\alpha_i \le \frac{(m+s-1)(n-1)}{(1-s)(n-1-h)} + \frac{\beta_i h}{(n-1-h)}$$

which proves the claim.

To interpret part c) of Proposition B2, note that aversion to advantageous inequality (positive value of β) decreases the incentive to choose contributions below E, because doing so generates

advantageous inequality vis-à-vis the full contributors. To investigate the lowest values of α, which are consistent with players contributing less than *E*, set $\beta_i = 0$ in B1. Consider the 5 - *h* group members with α ≥ 1. Part c) of the proposition shows that these players choose *C* = *E* with certainty if they have $\alpha_i < \alpha(h)$, where α(0) = 1, α(1) = 1.4, α(2) = 2, α(3) = 4, α(4) = ∞.

Assuming the distribution of α presented in Fehr and Schmidt Table III, the probability of meeting the conditions for equilibria other than universal, full contribution is equivalent to the probability that all group members have α ≥ 1, which equals $0.3^5 = 0.2$ percent. For simplicity, we assume below that groups using deterrent FS always obtain universal, full contribution.

FS, non-deterrent

An analogous proposition to Proposition 4 in Fehr and Schmidt 1999 can be proved:

a) If $\beta_i < 1 - m/(1-s) = 1/3$ for player *i* then it is a dominant strategy for that player to choose $C_i = 0$.

b) Let *k* denote the number of players with $\beta_i < 1/3$. If $k > (n-1)m/2(1-s) = 4/3$, then there is a unique equilibrium with $C_i = 0$ for all *i*.

   This means that positive contributions are possible even if there is one player who cares little about advantageous inequality, but not if there is more than one.

c) If $k < (n-1)(m/(1-s) + \beta_i - 1)/(\alpha_i + \beta_i) = (4\beta_i - 4/3)/(\alpha_i + \beta_i)$ for all players with $\beta_i \geq 1/3$, then equilibria with positive contribution levels exist. In these equilibria all players with $\beta_i < 1/3$ must choose $C_i = 0$, while all other players contribute $C_j = C \in [0, 20]$.

Proofs are completely analogous to those in Fehr and Schmidt.

Equilibria with positive contributions are more feasible in non-deterrent FS than in NS for two reasons: first, the *β*-threshold for being a potential contributor is lower (*β* ≥ .6 in NS and *β* ≥ .33 in non-deterrent FS). Second, in NS equilibria with positive contributions are only feasible if all players are above the threshold. In non-deterrent FS, such equilibria may occur even if one player is below the threshold of *β* ≥ .33.

The ultimatum game results used by Fehr and Schmidt to derive their table III are fully consistent with assuming that 70 percent of subjects have β ≥.33. In this case, the probability of meeting the conditions for equilibria with positive contributions is 52.8 percent.

I.2 Voting

Assume that subjects vote for the scheme yielding the highest, expected utility.[2] In case the same set of equilibria exist in two institutions, assume that groups coordinate on the same equilibrium in both.

IS vs. NS: Groups in which the distribution of social preference parameters implies that cooperative equilibria are feasible in either both or none of the institutions should be indifferent between IS and NS. Groups in which cooperative equilibria are feasible in IS but not in NS should vote for IS. We have shown above that cooperative equilibria are feasible for a wider range of social preferences in IS than in NS. Therefore, at least half of groups should choose IS.

FS vs. NS

Deterrent FS vs. NS

Groups in which the distribution of social preference parameters implies cooperative equilibria are feasible in NS should vote for NS if and only if voters believe they will be able to coordinate on equilibria with contribution levels of at least 20 – *c* (assuming that the equilibrium with universal, full contribution applies in deterrent FS).

Proof:
$U^{NS}(C) \geq U^{FS,\ deterrent}(E) \iff (E - C) + nmC \geq nmE - c \iff C \geq E - c / (nm-1) = 20 - c$

In particular, groups must coordinate on contributing at least 18 to make NS the preferred option in DC, and at least 12 in DE. If abilities to coordinate on high equlibria are smoothly distributed, and therefore some groups exist which coordinate on contributing more than 12 but less than 18 in NS, then this implies that more groups vote for FS in DC than in DE. Groups where no cooperative equilibria exist in NS should vote for FS.

Non-deterrent FS vs. NS

Groups where cooperative equilibria are feasible in either both or none of the institutions should vote for NS, to save the fixed cost of FS (and avoid sanctions in the case of the non-cooperative equilibrium). Groups where symmetric, cooperative equilibria are feasible in FS but not in NS should vote for FS if they are able to coordinate on symmetric equilibria with contributions of at least (8 + *c*) / 1.4, equal to 7.1 in NC and 11.4 in NE.[3]

Proof:
$U^{FS,\ non\text{-}det.}(C) \geq U^{NS}(0) \iff (E - C)(1 - s) + nmC - c \geq E \iff C \geq (sE + c) / (nm + s - 1) = (8 + c) / 1.4$

[2] Note that all equilibria in NS, IS and deterrent FS are symmetric. In those cases, therefore, utility equals own, pecuniary pay-off.

[3] In non-deterrent FS, equilibria may occur where one individual free rides while the others contribute. In those cases, a cooperator *i* should vote for FS if cooperators are able to coordinate on $C \geq (8+c)/(1-\alpha_i/4)$. The free rider *j* should vote for FS if $C \geq (8+c)/(1.6-\beta_j/4)$.

We have shown above that cooperative equilibria are feasible for a wider range of social preferences in non-deterrent FS than in NS, implying that some groups potentially fall into this category. More groups should vote for FS in NC than in NE. Again, the reason is that groups needs to coordinate on higher levels of contributions to make FS profitable in NE than in NC.

FS vs. IS

Deterrent FS vs. IS

Groups in which cooperative equilibria are feasible in IS should vote for IS if and only if they can coordinate on equilibria with contribution levels of at least 20-*c*, again assuming universal, full contribution in deterrent FS. Other groups should vote for FS. Again, more groups should vote for FS in DC than in DE because groups needs to coordinate on higher levels of contributions to make FS profitable in DE than in DC.

Non-deterrent FS vs. IS

Groups in which cooperative equilibria are feasible in either both or none of the institutions should vote for IS, to save the fixed cost of FS (and avoid sanctions in case the non-cooperative equilibrium is selected). Groups where cooperative equilibria are feasible in FS but not in IS should vote for FS if they are able to coordinate on symmetric equlibria with contributions of at least (8+*c*)/1.4.[4] Again, more groups should vote for FS in NC than in NE.

Comparing voting predictions for FS vs. NS with those for FS vs. IS, it is easy to see that FS is predicted to be on average more popular when pitted against NS than when pitted against IS. The reason is that cooperative equilibria are feasible for a wider range of social preferences in IS than in NS. Therefore, for some groups IS is an attractive alternative to FS, even when NS is not.

II. Charness and Rabin (2002), model without reciprocity

This section derives predictions from the model of social welfare maximization presented in Appendix 1 of Charness and Rabin (2002). There are two versions of this model. We first consider the simple version, which ignores reciprocity preferences, and then discuss the extended model, which includes reciprocity. Charness and Rabin assume preferences as follows:

$$U_i^{CR}(\pi) = (1-\lambda_i)\pi_i + \lambda_i\left(\delta_i \min_j(\pi_j) + (1-\delta_i)\sum_j \pi_j\right) \qquad \text{(B3)}$$

where λ and δ are parameters between 0 and 1. We assume that preference parameters are individual-specific, but suppress subscripts below. Intuitively, λ measures the weight an individual attaches to social welfare maximization, relative to maximization of own pay-off. The model considers two aspects of social welfare preferences, namely maximin preferences (i.e. the

[4] See footnote above on asymmetric equilibria in non-deterrent FS.

desire to maximize the earnings of the worst-off individual), and preferences for efficiency (i.e. the desire to maximize aggregate pay-off). The parameter δ measures the weight attached to maximin preferences, relative to preferences for efficiency.

II. 1 Contributions

NS and non-deterrent FS:

(To save space, these two regimes are considered together)

Individual profits in case formal sanctions apply are (see eq. 2 in the main text)

$$\pi_i = \left(E - C_i\right)\left(1-s\right) + m\sum_{j=1}^{n} C_j - c\,.$$

Consider the incentive for player *i* to contribute a positive amount, given that all others contribute zero. Note that as long as sanctions are non-deterrent, *i* is necessarily the worst off person in this situation. The derivative of U_i^{CR} with respect to C_i is positive if:

$$\begin{aligned} &\frac{\partial U_i^{CR}}{\partial C_i} = \left(1-\lambda\right)\left(m-(1-s)\right) + \lambda\delta\left(m-(1-s)\right) + \lambda(1-\delta)\left(nm-1\right)\right) \geq 0 \\ &\Leftrightarrow \delta \leq 1 - \frac{1-m-s}{\lambda\left(nm-m-s\right)} = 1 - \frac{0.6-s}{\lambda\left(1.6-s\right)} \equiv \delta^* \end{aligned} \tag{B4}$$

where our parameter values for *n* and *m* have been inserted in the last term.

NS: set $s = 0$. Positive contributions increase utility if $\delta \leq 1 - 3/8\lambda$. This implies that λ must be at least 3/8 for positive contributions to be optimal. Higher values of λ increase the threshold value of δ, δ^*. δ must be lower than 5/8 (the threshold value if $\lambda = 1$).

Non-deterrent FS: Note that δ^* is increasing in *s*. Therefore, higher sanctions widen the range of social welfare preferences for which cooperation is the best strategy. With $s = 0.4$ (our non-deterrent value), cooperation is potentially possible if $\lambda \geq 0.2/1.2 = 1/6$, which is lower than the threshold without sanctions. δ must be lower than 5/6 (the threshold value if $\lambda = 1$), which is higher than the threshold without sanctions. Hence, non-deterrent sanctions increase the scope for cooperative strategies.

Deterrent FS: Full contribution is the only equilibrium. The reason is that deviations reduce own payoff, aggregate payoff, and the payoff of the worst-off person.

IS: Since punishment under IS reduces both own and aggregate pay-off, and does not increase the pay-off of the worst-off group member, the punishment option is never utilized (see Sutter, Haigner and Kocher 2010, Appendix B). As with standard theory, predictions for IS are therefore the same as for NS.

II. 2 Voting

IS vs. NS: Since punishment is never used (on or off the equilibrium path), expected payoffs are identical under IS and FS. Voters should therefore be indifferent between the two institutions.

NS vs. deterrent FS:

Assuming universal, full contribution, utility under deterrent FS in, respectively, the DC and the DE treatments are:

$$U_{FS}^{DC} = (1-\lambda+\lambda\delta)38+\lambda(1-\delta)190 \quad \text{(B5)}$$

$$U_{FS}^{DE} = (1-\lambda+\lambda\delta)32+\lambda(1-\delta)160 \quad \text{(B6)}$$

(each subject earns 0.4*5*20 – *c* , equal to 38 in DC and 32 in DE. Total group earnings are 5*38 = 190 in DC and 5*32 = 160 in DE).

Table B.1 shows utility levels for free riders and cooperators under NS.

Table B.1 Utility under NS

Number of free riders	Cooperators	Free riders
0	$(1-\lambda+\lambda\delta)40+\lambda(1-\delta)200$	
1	$(1-\lambda+\lambda\delta)32+\lambda(1-\delta)180$	$(1-\lambda)52+\lambda\delta32+\lambda(1-\delta)180$
2	$(1-\lambda+\lambda\delta)24+\lambda(1-\delta)160$	$(1-\lambda)44+\lambda\delta24+\lambda(1-\delta)160$
3	$(1-\lambda+\lambda\delta)16+\lambda(1-\delta)140$	$(1-\lambda)36+\lambda\delta16+\lambda(1-\delta)140$
4	$(1-\lambda+\lambda\delta)8+\lambda(1-\delta)120$	$(1-\lambda)28+\lambda\delta8+\lambda(1-\delta)120$
5		$(1-\lambda+\lambda\delta)20+\lambda(1-\delta)100$

Table B.2 Utility under non-deterrent FS, NC treatment

Number of free riders	Cooperators	Free riders
0	$(1-\lambda+\lambda\delta)38+\lambda(1-\delta)190$	
1	$(1-\lambda+\lambda\delta)30+\lambda(1-\delta)162$	$(1-\lambda)42+\lambda\delta30+\lambda(1-\delta)162$
2	$(1-\lambda+\lambda\delta)22+\lambda(1-\delta)134$	$(1-\lambda)34+\lambda\delta22+\lambda(1-\delta)134$
3	$(1-\lambda+\lambda\delta)14+\lambda(1-\delta)106$	$(1-\lambda)26+\lambda\delta14+\lambda(1-\delta)106$
4	$(1-\lambda+\lambda\delta)6+\lambda(1-\delta)78$	$(1-\lambda)18+\lambda\delta6+\lambda(1-\delta)78$
5		$(1-\lambda+\lambda\delta)10+\lambda(1-\delta)50$

Table B.3 Utility under non-deterrent FS, NE treatment

Number of free riders	Cooperators	Free riders
0	$(1-\lambda+\lambda\delta)32+\lambda(1-\delta)160$	
1	$(1-\lambda+\lambda\delta)24+\lambda(1-\delta)132$	$(1-\lambda)36+\lambda\delta 24+\lambda(1-\delta)132$
2	$(1-\lambda+\lambda\delta)16+\lambda(1-\delta)104$	$(1-\lambda)28+\lambda\delta 16+\lambda(1-\delta)104$
3	$(1-\lambda+\lambda\delta)8+\lambda(1-\delta)76$	$(1-\lambda)20+\lambda\delta 8+\lambda(1-\delta)76$
4	$(1-\lambda+\lambda\delta)0+\lambda(1-\delta)48$	$(1-\lambda)12+\lambda\delta 0+\lambda(1-\delta)48$
5		$(1-\lambda+\lambda\delta)4+\lambda(1-\delta)20$

Comparisons between utility levels under NS and deterrent FS show that:

In DC, groups vote for NS when all five group members are cooperators, and otherwise choose FS (in case there are one or two free riders, the free riders may vote for NS).

In DE, groups vote for NS when there are zero or one free riders. With three free riders, groups vote for NS if free riders value social welfare sufficiently little (e.g. if they have $\lambda = 0$). In cases with zero or one free riders, all group members support NS. With two or three free riders, only the free riders (may) support NS. Only with three free riders does this group constitute a majority.

The number of free riders depends on the distribution of social preference parameters. Hence, voting depends on preference distributions. Voting for NS is more likely in DE than in DC, in the sense that voting for NS is optimal under a wider spectrum of preferences in DE than in DC.

NS vs. non-deterrent FS

As demonstrated above, the number of cooperators is potentially higher in non-deterrent FS than in NS. Therefore, there are three types of voters: 1) those who are cooperators in both regimes, 2) those who are free riders in both regimes, 3) those who shift from free riding in NS to cooperating in FS.

Utility levels of, respectively, cooperators and free riders in each regime are given by:

$$\begin{aligned} U_{Coop}^{NS} &= (1-\lambda+\lambda\delta)mn'E+\lambda(1-\delta)\big((n-n')E+nmn'E\big) \\ &= (1-\lambda+\lambda\delta)8n'+\lambda(1-\delta)(100+20n') \end{aligned} \tag{B7}$$

$$\begin{aligned} U_{FR}^{NS} &= (1-\lambda)\big(E+mn'E\big)+\lambda\delta mn'E+\lambda(1-\delta)\big((n-n')E+nmn'E\big) \\ &= (1-\lambda)\big(20+8n'\big)+\lambda\delta 8n'+\lambda(1-\delta)(100+20n') \end{aligned} \tag{B8}$$

$$\begin{aligned} U_{Coop}^{FS} &= (1-\lambda+\lambda\delta)\big(mn''E-c\big)+\lambda(1-\delta)\big((n-n')E(1-s)+nmn''E-nc\big) \\ &= (1-\lambda+\lambda\delta)\big(8n''-c\big)+\lambda(1-\delta)\big(60+28n''-5c\big) \end{aligned} \tag{B9}$$

$$\begin{aligned} U_{FR}^{FS} &= (1-\lambda)\big(E(1-s)+mn''E-c\big)+\lambda\delta\big(mn''E-c\big)+\lambda(1-\delta)\big((n-n')E(1-s)+nmn''E-nc\big) \\ &= (1-\lambda)\big(12+8n''-c\big)+\lambda\delta\big(8n''-c\big)+\lambda(1-\delta)\big(60+28n''-5c\big) \end{aligned} \tag{B10}$$

Where *n'* is the number of cooperators in NS and *n''* is the number of cooperators in FS. The values of these utility functions, conditional on numbers of cooperators, are tabulated for NS and for FS in NC and NE, in tables B.1-B.3. The condition for voting for NS in each of the three groups of voters are:

Cooperators in both regimes: $U_{Coop}^{FS} \geq U_{Coop}^{NS}$

Free riders in both regimes: $U_{FR}^{FS} \geq U_{FR}^{NS}$

Free riders-turning-cooperators: $U_{Coop}^{FS} \geq U_{FR}^{NS}$

Clearly, NS is preferred as long as *n'* = *n''*. However, when *n''* exceeds *n'* by a sufficient amount, voting for FS becomes attractive. This condition is more difficult to meet in NE than in NC, although comparison of Tables B.1 and B.3 show that for some parameter constellations, FS should in fact be preferred to NS even in NE.

In sum, the Charness-Rabin model predicts that, depending on the distribution of social preference parameters, some groups vote for NS and some for FS in NC and NE. The share of groups voting for FS is (weakly) higher in NC than in NE. Comparison of equation B4 (B5) and Table B.2 (B.3) shows that voting for FS is more likely in DC (DE) than in NC (NE).

Since contribution behavior is predicted to be the same under IS and NS, and the punishment option is not used in IS, FS is predicted to be equally popular when pitted against IS and NS.

III. Comparison of the predictions derived from the two models

In sum, the two behavioral theories, and the standard model of rational egoism, yield distinct predictions on contributions behavior and voting. While standard theory predicts zero contributions in NS, IS and non-deterrent FS, the theories of both Charness and Rabin and Fehr and Schmidt predict positive contributions for some empirically plausible distributions of social preference parameters. Only Fehr and Schmidt's theory predicts an effect of informal sanctions on contributions. Standard theory predicts that the popularity of FS is driven by the deterrence level of formal sanctions but not by the fixed cost of FS. In contrast, both behavioral theories predict that fixed cost as well as deterrence shape voting for FS. Fehr and Schmidt differ from Charness and Rabin by predicting that FS is more popular when the alternative is NS than when it is IS. Predictions are summarized in table B.4

Table B.4 Theoretical Predictions

Panel (A) – Contributions	**Standard theory**	**Charness and Rabin 2002 (Social welfare maximization)**	**Fehr and Schmidt 1999 (Inequity aversion)**
NS	$C_i = 0 \ \forall i$	$C_i = 20$ if $\delta_i \leq 1-3/(8\lambda_i)$, otherwise 0	F&S Proposition 4: $C_i = 0 \ \forall i$ if $\exists j$ with $\beta_j < 0.6$, $C_i \in [0,20]$ if $\beta_j \geq 0.6 \ \forall i$
IS	$C_i = 0 \ \forall i$ No punishment (on or off equilibrium path)	$C_i = 20$ if $\delta_i \leq 1-3/(8\lambda_i)$, otherwise 0 No punishment (on or off equilibrium path)	F&S Proposition 5: Suppose there are *n'* "conditionally cooperative enforcers" with preferences that obey 3) $\beta_i \geq 0.6$ and 4) $\frac{1}{4} < \frac{\alpha_i}{4(1+\alpha_i)-(n'-1)(\alpha_i+\beta_i)}$ Where n' is the number of group members with $\beta_i \geq 0.6$ Then there is a sub-game perfect Nash eq. with the following characteristics: c) Each player contributes $C_i = C \in [0,20]$. d) No punishment. If anybody (off the equilibrium path) chooses $C_i < C$, each conditionally cooperative enforcers reduces the deviator's earnings by $(C - C_i)/(n' - 1/4)$.
FS, deterrent	$C_i = E \ \forall i$	$C_i = E \ \forall i$	$C_i = E \ \forall i$ is always an equilibrium, although other equilibria exist if agents are very averse to disadvantageous inequality.
FS, non-deterrent	$C_i = 0 \ \forall i$	$C_i = E$ if $\delta_i \leq 1-1/(6\lambda_i)$, otherwise 0	$C_i = 0 \ \forall i$ if there are at least two players with $\beta_i < 0.33$ If there are at least four players with $\beta_i \geq 0.33$, then equilibria with positive contribution levels may exist. In these equilibria the players with $\beta_i < 0.33$ choose $C_i = 0$, while all others contribute $C_j = C \in [0,20]$.
Panel (B)- Voting			
IS vs. NS	Indifferent	Indifferent	A majority vote for IS
FS vs. NS	Vote for FS in DC and DE. Vote for NS in NC and NE	Depends on distribution of δ and λ. The share of groups voting for FS increases with the level of deterrence and decreases with the fixed cost of FS.	Depends on distribution of β. The share of groups voting for FS increases with the level of deterrence and decreases with the fixed cost of FS.
FS vs. IS	As in FS vs. NS	As in FS vs. NS	Depends on distribution of α and β. The share of groups voting for FS increases with the level of deterrence and decreases with the fixed cost of FS. In all treatments, FS is less popular when pitted against IS than when pitted against NS.

Table B.5 summarizes theory predictions on the parameters estimated in the regressions reported in Table 2.

Table B.5 Predictions on parameters in voting regressions (cf. equation 4 in the main text)

Parameter	**Standard Theory**	**Charness and Rabin**	**Fehr and Schmidt**
γ_1	+	+	+
γ_2	0	+	+
γ_3	0	0	-

IV. Charness and Rabin (2002), model with reciprocity

We now discuss how the predictions of the Charness and Rabin model change if the richer version which includes reciprocity preferences, is used. The utility function in this case is:

$$U_i(s,d) = (1-\lambda)\pi_i + \lambda\left\{\delta \min_j\left[\pi_i, \min_{j\neq i}\left(\pi_j + bd_j\right)\right] + (1-\delta)\left[\pi_i + \sum_{j\neq i}\max\left(1-kd_j, 0\right)\pi_j\right] - f\sum_{j\neq i} d_j\pi_j\right\} \quad \text{(B11)}$$

where s is strategies, d_j is the "demerit" of player j – a measure of the degree to which j fails to meet acceptable standards of pro-social behavior. b, k and f are non-negative parameters. We consider a simplified version of this model.

First, assume that $\delta = 0$[5], the function then simplifies to:

$$U_i(s,d) = \pi_i + \lambda\sum_{j\neq i}\max\left(1-kd_j, 0\right)\pi_j - \lambda f\sum_{j\neq i} d_j\pi_j \quad \text{(B12)}$$

d_j is defined as max (λ^*- g, 0), where λ^* is the standard of kindness and g is endogenously derived as the weight that j's behavior implies that he puts on social welfare, given what others players believe j believes about their strategies. Now, assume that k is a large number (in particular, $k \geq 1/d_j$ for all $d_j > 0$). (B12) then simplifies to:

[5] This might be a reasonable simplification in our context: If player i is not the worst-off person, then there is no trade-off between efficiency and maximizing the income of the worst-off person – increased contribution to the PG achieves both. If player i *is* the worst-off person, then there is such a trade-off: increased contribution improves efficiency but leaves the worst-off person worse off. However, this is the same trade-off as the one between maximizing own pay-off and social welfare, which is already captured by the parameter λ, which measures the weight that i attaches to social welfare.

$$U_i(s,d) = \pi_i + \lambda \sum_{j|d_j=0} \pi_j - \lambda f \sum_{j|d_j>0} d_j \pi_j \qquad \text{(B13)}$$

The difficulty lies in determining the demerit values, because *i*'s assessment of the weight *j* puts on social welfare depends not only on *j*'s behavior but also on *i*'s beliefs about *j*'s beliefs.

We focus on showing that, in contrast with the Charness-Rabin model without reciprocity, predictions on contributions are not identical under NS and IS. In particular, in some groups IS generates equilibria with higher contributions and earnings than NS.

Assume the following: a) players who choose $C = 0$ are assumed by other players to not value social welfare and get $d = \lambda^*$. This is not necessarily the only rational belief, but it is one rational belief. b) Players who choose $C = 20$ get $d = 0$. Again, other beliefs may be rational, but this belief is never irrational.

IV.1 Contributions

NS and non-deterrent FS:

> *Proposition B2:* Universal non-contribution is an equilibrium, for any distribution of λ and f.

Proof: If all other group members are assumed to contribute nothing to the public good, individual *i*'s utility, and first-order-condition with respect to C_i, are given by:

$$\begin{aligned} U_i(C) &= \left(E - C_i\right)\left(1-s\right) + mC_i - c - \lambda f \left(n-1\right)\lambda^* \left(E\left(1-s\right) + mC_i - c\right) \\ \frac{\partial U_i}{\partial C_i} &= m + s - 1 - \lambda f \left(n-1\right)\lambda^* m < 0 \end{aligned} \qquad \text{(B14)}$$

Increasing contributions strictly decreases utility, so universal non-contribution is an equilibrium.

> *Proposition B3:* Universal full contribution is an equilibrium if $\lambda \geq \frac{1-m-s}{\left(n-1\right)m}$ for all group members.

Proof: If all other group members are assumed to contribute everything to the public good, individual *i*'s utility, and first-order-condition with respect to C_i, are given by:

$$\begin{aligned} U_i(C) &= \left(E - C_i\right)\left(1-s\right) + mC_i - c + \lambda\left(n-1\right)\left(\left(n-1\right)mE + mC_i - c\right) \\ \frac{\partial U_i}{\partial C_i} &= m + s - 1 + \lambda\left(n-1\right)m \geq 0 \Leftrightarrow \lambda \geq \frac{1-m-s}{\left(n-1\right)m} \end{aligned} \qquad \text{(B14)}$$

(As in the C&R model without reciprocity, described above, this means that the threshold value for choosing high contributions is lower in non-deterrent FS ($\lambda \geq 1/6$) than in NS ($\lambda \geq 3/8$)).

Deterrent FS: Universal, full contribution is an equilibrium. Proof: for $s \geq 1\text{-}m$, $\partial U_i / \partial C_i$ is positive for any value of λ (see B14 above).

IS: Assume that group members other than *i* contribute either *E* or zero. Solve the game backwards. Consider first *i*'s incentive to punish free riders (punishment by other group members than *i* is ignored). $\tilde{\pi}_i$ denotes the stage-one (before punishment) pay-off of individual *i*:

$$U_i(C,p) = \tilde{\pi}_i - \sum_j p_{ij} + \lambda \sum_{j|d_j=0} \tilde{\pi}_j - \lambda f \lambda^* \sum_{j|d_j=\lambda^*} \left(\tilde{\pi}_j - \sigma p_{ij}\right)$$
$$\frac{\partial U_i}{\partial p_{ij}} = -1 + \lambda f \lambda^* \sigma \geq 0 \Leftrightarrow \lambda f \lambda^* \geq \frac{1}{\sigma} \qquad \text{(B16)}$$

If at least one group member has $\lambda f \lambda^* \geq 1/\sigma$, all free riders will see their earnings reduced to zero in the second stage. Therefore, if this condition is fulfilled, equilibria with zero contributions disappear.

Members of groups where this condition is fulfilled have an incentive to vote for IS over NS (note that both egoists and subjects with social preferences have this incentive – in equilibrium, free riders benefit from the higher over-all contributions brought about by the "reciprocators"). Note that equilibria with full contribution are at least as likely as in NS, and might be more likely, depending on belief formation, the strength of reciprocity preferences and on which standard of pro-social behavior a group applies.

IV.2 Voting

For some groups, namely those where at least one member fulfils $\lambda f \lambda^* \geq 1/\sigma$, FS is more likely to be outperformed by IS than by NS, in terms of earnings. Therefore, FS should on average be more popular when pitted against NS than when pitted against IS. This means that voting predictions from the Charness-Rabin model converge, at least qualitatively, toward those emerging from the Fehr-Schmidt model when reciprocity preferences are included in the model (see table B4 above).

Appendix C – Additional tables and figures

Table C.1 Voting outcomes (percent)

		Group level					
Share of groups		*Vote*					
adopting:	*Treatment*	1	2	3	4	5	6
No Sanctions	DC	78.6	28.6	-	50.0	14.3	-
	DE	75.0	83.3	-	50.0	66.7	-
	NC	85.7	42.9	-	35.7	57.1	-
	NE	75.0	91.7	-	33.3	83.3	-
Informal Sanctions	DC	21.4	-	42.9	50.0	-	42.9
	DE	25.0	-	83.3	50.0	-	66.7
	NC	14.3	-	64.3	64.3	-	71.4
	NE	25.0	-	100.0	66.7	-	100.0
Formal Sanctions	DC	-	71.4	57.1	-	85.7	57.1
	DE	-	16.7	16.7	-	33.3	33.3
	NC	-	57.1	35.7	-	42.9	28.6
	NE	-	8.3	0.0	-	16.7	0.0
		Individual level					
Share of subjects		*Vote*					
voting for:	*Treatment*	1	2	3	4	5	6
No Sanctions	DC	65.7	37.1	-	40.0	24.3	-
	DE	65.0	66.7	-	53.3	61.7	-
	NC	68.6	40.0	-	45.7	55.7	-
	NE	70.0	65.0	-	40.0	60.0	-
Informal Sanctions	DC	34.3	-	42.9	60.0	-	51.4
	DE	35.0	-	73.3	46.7	-	68.3
	NC	31.4	-	64.3	54.3	-	64.3
	NE	30.0	-	71.7	60.0	-	81.7
Formal Sanctions	DC	-	62.9	57.1	-	75.7	48.6
	DE	-	33.3	26.7	-	38.3	31.7
	NC	-	60.0	35.7	-	44.3	35.7
	NE	-	35.0	28.3	-	40.0	18.3

Table C.2 Punishment under informal sanctions

	Treatment				
	DC	DE	NC	NE	All
Share of punishment options used	0.08	0.12	0.14	0.19	0.14
Share of individuals who punished at least once	0.84	0.71	0.83	0.95	0.83
Reduction points sent per person per period, mean:					
All	0.58	0.91	1.13	1.25	1.01
Punishers only	2.80	3.70	3.24	2.84	3.10
Share of individuals who received punishment at least once	0.70	0.96	0.79	0.93	0.85
Punishment points received per person per period, mean:					
All	0.58	0.91	1.13	1.25	1.01
Punished individuals only	3.31	2.81	3.37	2.79	3.00
Share of punishment directed at above-median contributors ("perverse punishment")	0.00	0.14	0.14	0.12	0.13

Note: The first line is based on the 9,120 punishment options opportunities. Lines 2 and 5 are based on the 235 participants who used informal sanctions in at least one phase. Lines 3 and 6 are based on 2,280 subject-by-period observations under informal sanctions. Line 4 is based on the 741 subject-by-period observations where punishment points were *sent.* Line 7 is based on the 766 subject-by-period observations where punishment points were *received.* The last line is based on the 1,255 punishment opportunities where punishment was actually used. (Note that lines 3 and 6 are by necessity identical.)

We report in the paper that informal sanctions were "**behaviorally deterrent**", i.e. that punishment was mainly well-targeted, so that subjects tended to earn more by contributing more. Our evidence for this claim is as follows. We estimate GLS regressions for all IS-condition observations at individual level, with earnings as dependent and contribution to group account as independent variable, controls for group average contribution and period, and standard errors clustered by group. There is one regression for each of four phases (2, 4, 5, 7) and for each treatment. The estimated coefficients on amount contributed are positive in 13 of 16 regressions with 6 of the 13 positive coefficients being statistically significant, half of these at the 1% level. Efficient targeting of punishment is also evident from the fact that individuals contributing above their group median received only 13% of all punishment given.

Table C.3 Voting and earnings

	Dependent variable:						
	Voted for *formal* sanctions (against informal) in **Vote 3**	Voted for *informal* sanctions (against no sanctions) in **Vote 4**		Voted for *formal* sanctions (against no sanctions) in **Vote 5**		Voted for *formal* sanctions (against informal) in **Vote 6**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Earn_fs/earn_is	0.709**					0.849***	0.677***
	(0.322)					(0.254)	(0.228)
Earn_is/earn_ns		0.865***	0.889***				
		(0.132)	(0.148)				
Earn_fs/earn_ns				0.566***	0.580***		
				(0.202)	(0.207)		
CoV_is/CoV_ns			0.002				
			(0.003)				
CoV_fs/CoV_ns					0.015		
					(0.035)		
CoV_fs/CoV_is							-0.043
							(0.059)
Treatment:							
DC		0.038	0.039	0.136	0.144	0.032	0.183
		(0.112)	(0.114)	(0.121)	(0.123)	(0.172)	(0.188)
DE		-0.079	-0.031	0.100	0.110	-0.173	0.113
		(0.097)	(0.095)	(0.157)	(0.157)	(0.177)	(0.219)
NC		-0.114	-0.111			-0.126	-0.08
		(0.110)	(0.111)			(0.145)	(0.162)
Pseudo-R^2	0.4	0.21	0.22	0.13	0.13	0.42	0.43
Observations	15	200	195	125	125	125	85

Note: Probit regressions, marginal effects reported. Standard errors in parentheses. Standard errors adjusted for within-group clustering, except in the regression for Phase 4, where only three groups are included. In the regression for Phase 6, the outcome does not vary within the **NE** treatment (all five individuals with the relevant experience voted for formal sanctions) and the observations in this treatment can therefore not be included. The earnings variables are always based on experience from *the most recent phase*. For example, "earn_fs/earn_is" is mean earnings in the most recent phase with formal sanctions, divided by mean earnings in the most recent phase with informal sanctions. "CoV" stands for "Coefficient of Variation". The number of observations in column (7) is less than that in column (6) due to the dropping of observations for which CoV_is = 0. * significant at 10%; ** significant at 5%; *** significant at 1%

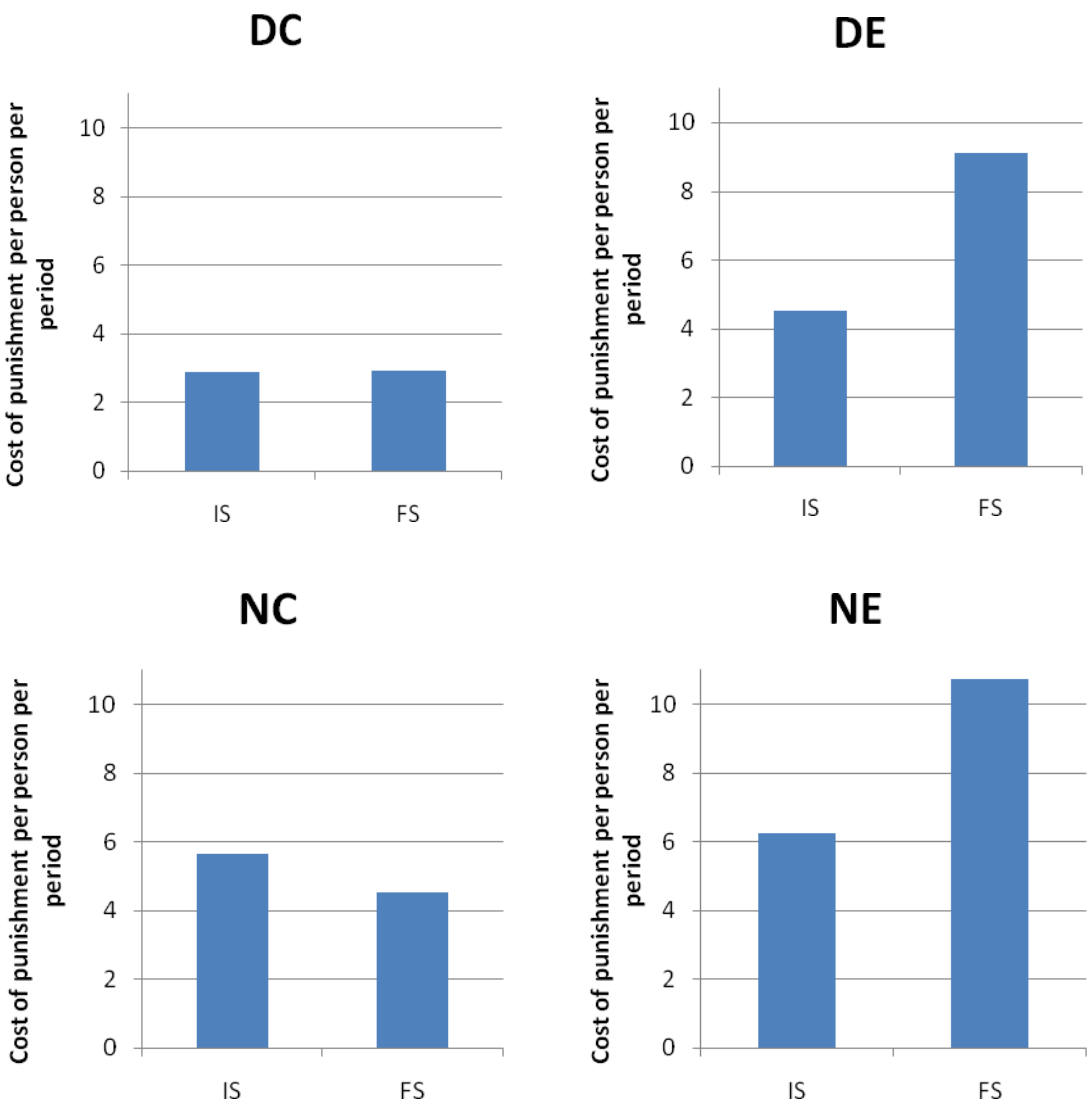
Table C.4 Contributions per person per period by treatment, condition and

Treatment		*Phase*						
	Condition	1	2	3	4	5	6	7
DC	NS	*11.5*	*8.7*	6.6		6.7	13.4	
	IS		16.9		19.0	19.6		*19.7*
	FS			18.6	18.9		18.9	19.0
	*MW-test, p-value**		*0.02*	*0.00*	*1.00*	*0.00*	*0.04*	*0.53*
DE	NS	*12.6*	*11.6*	12.1		10.7	12.0	
	IS		14.3		16.6	18.9		*19.0*
	FS			17.0	18.4		19.5	18.6
	MW-test, p-value		*0.41*	*0.13*	*0.67*	*0.03*	*0.04*	*0.86*
NC	NS	*11.9*	*9.3*	9.9		9.1	8.4	
	IS		18.9		16.9	17.9		*19.1*
	FS			14.1	11.9		14.0	14.4
	MW-test, p-value		*0.03*	*0.04*	*0.01*	*0.00*	*0.01*	*0.01*
NE	NS	*9.1*	*6.3*	4.7		3.8	4.2	
	IS		11.8		12.7	16.2		*13.8*
	FS			17.8	-		10.8	
	MW-test, p-value		*0.02*	*0.11*	-	*0.01*	*0.03*	-
All	NS	*11.3*	*9.0*	8.4		7.8	8.3	
	IS		15.2		15.8	18.0		*17.4*
	FS			16.7	16.5		17.1	17.7
	MW-test, p-value		*0.00*	*0.00*	*0.52*	*0.00*	*0.00*	*0.93*

*Mann-Whitney tests of the hypothesis that contributions under the two conditions are equal.
Note: Tests are conducted at the *group* level, for phase averages. All Mann-Whitney tests reported in the paper and appendix are two-tailed.

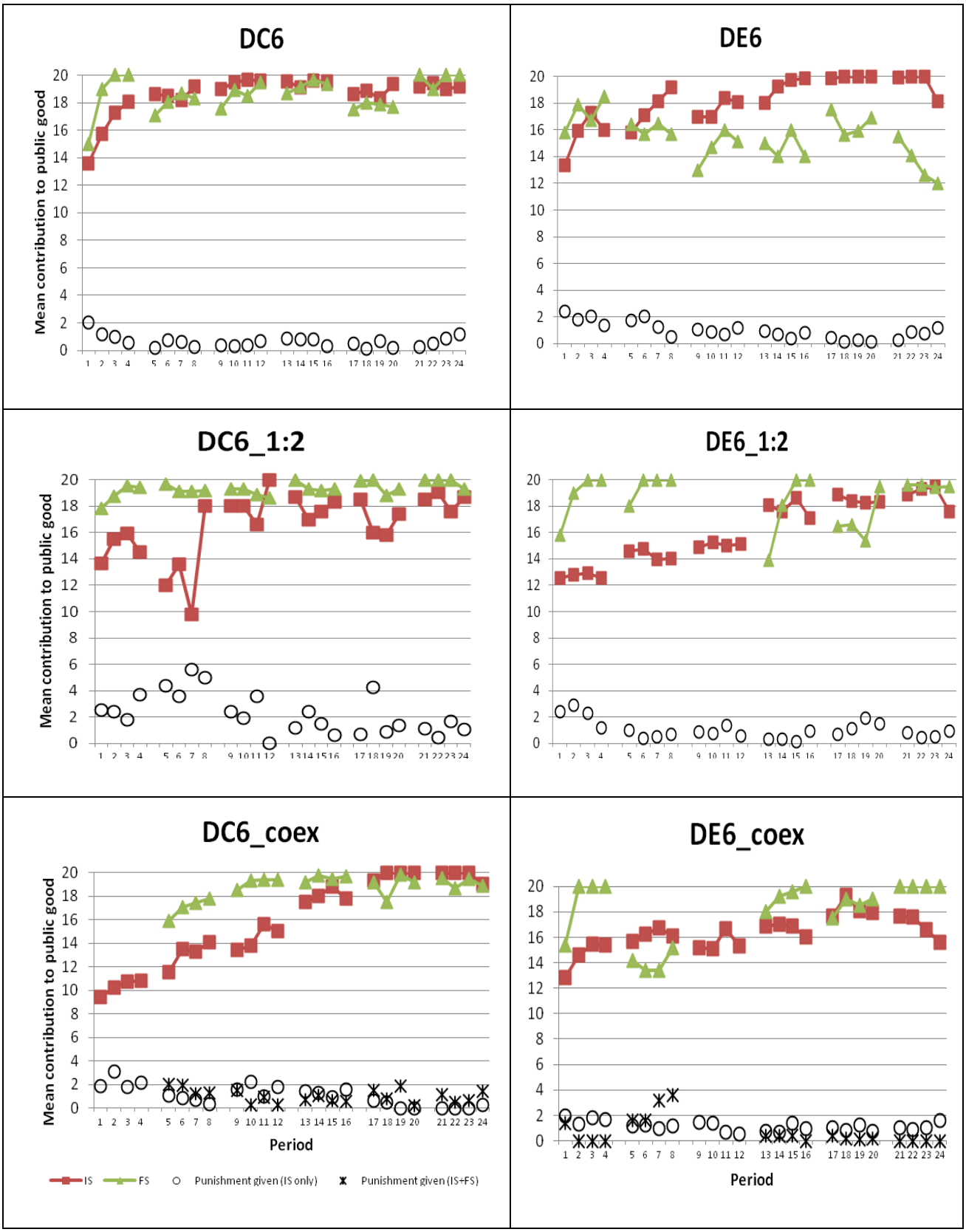
Comment to the table: This table shows that even under deterrent, formal sanctions, some subjects do not allocate their full endowment to the group activity. It is of some interest to understand why. In particular, contributions below 20 may either result from confusion about the rules, or from spite, for example related to being on the losing side of the group vote. We have run random effects regressions of contributions on a dummy for voting in favor of formal sanctions, for groups using formal sanctions in the DC and DE treatments. In DC, the voting dummy is not significant, but in DE, subjects who voted for FS contribute significantly more than those who votes against. This indicates that spite (dissatisfaction with the voting outcome) may contribute to explaining contributions below 20 in DE. Figure 4 in the main paper furthermore shows that contributions among groups using FS in the DE treatment tend to decline over time within each phase. Regression analyses of groups using FS in the DE treatment show that contributions decline over time for subjects who voted in favor of FS, but increase for those who voted against. So, one way to interpret the observation of declining contributions in DE is that the majority who voted in favor of FS initially choose high contributions, but in subsequent periods react to initially low contributions by the minority who voted against FS. The reaction by the majority may be motivated by a desire to punish low contributors or could simply be an expression of frustration.

Figure C.1. Absolute cost of sanctioning, by sanction scheme and treatment



The figure shows the total cost of sanctioning per person per period. Under IS, the cost of sanctioning includes costs to givers and receivers of punishment. Under FS, both the fixed cost of sanctions (2 or 8 points) and the point reductions received by free riders (sanction rate*points in private account).

Figure C.2 Contributions and punishment, robustness treatments



Note: *N* = 255