Institutions Influence Preferences

Carlos Rodríguez-Sickert
Ricardo-Andrés Guzmán
Juan-Camilo Cárdenas

SFI WORKING PAPER: 2007-08-019

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu

SANTA FE INSTITUTE

INSTITUTIONS INFLUENCE PREFERENCES: EVIDENCE FROM A COMMON POOL RESOURCE EXPERIMENT

Carlos Rodríguez-Sickert
Instituto de Sociología
Universidad Católica de Chile
and Santa Fe Institute

Ricardo Andrés Guzmán*
Instituto de Economía
Universidad Católica de Chile

Juan Camilo Cárdenas§
Facultad de Economía
Universidad de los Andes, Colombia
and Santa Fe Institute

September 28, 2006

* Corresponding author. Instituto de Economía, Facultad de Ciencias Económicas y Administrativas, Pontificia Universidad Católica de Chile. Avda. Vicuña Mackenna 4860, Macul, Santiago, Chile. Tel. (56 2) 354-4303. Fax (56 2) 553-2377. E-mail: rnguzman@puc.cl.

§ We are indebted to Sam Bowles, Marcos Singer, Rodrigo Harrison, Rodrigo Troncoso, Bob Rowthorn and Will Mullins.

ABSTRACT

We model the dynamic effects of external enforcement on the exploitation of a common pool resource. Fitting our model to experimental data we find that institutions influence social preferences. We solve two puzzles in the data: the increase and later erosion of cooperation when commoners vote against the imposition of a fine, and the high deterrence power of low fines. When fines are rejected, internalization of a social norm explains the increased cooperation; violations (accidental or not), coupled with reciprocal preferences, account for the erosion. Low fines stabilize cooperation by preventing a spiral of negative reciprocation.

1. INTRODUCTION

It is now widely agreed that social preferences, such as altruism, reciprocity, and guilt, are strong motives for behavior. Without a state to enforce property rights (or the disciplining hand of reputation) the selfish *homo economicus* engages in a war of all against all. Not the *homo sapiens*: social preferences help him avert chaos and cooperate.

Economists usually assume away the influence institutions exert on social preferences. Often the assumption is harmless, but occasionally it may result in unexpected or even disastrous consequences. English health authorities learned this the hard way. They decided to incentivize blood donations by paying donors. Instead of increasing, blood donations plummeted (Titmuss 1969).[1]

Experiments indicate institutions affect social preferences. For example, Gneezy and Rustichini (2000) studied day-care centers in Haifa, where a fine was imposed on parents who picked up their children late. Unexpectedly, tardiness more than doubled in those centers. A plausible explanation is that, by transforming a misdemeanor into a commodity that parents could buy cheaply, the fine eroded their sense of duty. Another example is Falk and Kosfeld's (2004) experimental study of principal-agent relations. They gave principals the option to set a lower bound on the effort of agents. Falk and Kosfeld found that agents who were not restricted by their principals worked harder than those who were. Agents seemed to punish distrust.

In this paper we explore the dynamic effects of external enforcement on the exploitation of a common pool resource (CPR).[2] As the previous evidence suggests,

---

[1] See Bowles (1998, 2005) for an extensive discussion of endogenous preferences and their policy implications.

[2] In a CPR game each player chooses privately how many tokens she will extract from a common pool. A player's material payoff depends positively on the number of tokens she extracts and negatively on the aggregate level of extraction. Thus, individual and social interest conflict.

external enforcement may change the preferences of players, so we begin by developing a model of CPR games that captures that possibility. The ingredients of the model are:

1. **Heterogeneous preferences.** We distinguish three types of players: (*i*) *selfish*, who only care about their own material payoffs; (*ii*) *unconditional cooperators*, who feel guilty when they violate a social norm; (*iii*) *conditional cooperators*, who experience guilt with an intensity that declines when others violate the norm.

2. **State-dependent preferences.** When institutions change, player types may change as well. Institutions comprise such things as the enforcement of a norm by an external authority.

3. **Stochastic behavior.** A player will choose with higher probability those actions that give her a higher expected utility.

4. **Adaptive expectations.** Each player has an estimate of how many tokens her peers will extract from the common pool, and updates that estimate as she observes what they actually do.

Next, we fit our model to experimental data. In our experiment, groups of five persons played a CPR game twenty times. In some treatments the experimenter fined players he caught extracting more than one token (he applied the fines in private to prevent shame from affecting behavior). Some groups were treated with a high fine, other groups with a low one. Both fines induced high levels of cooperation. The effect of the high fine accorded with our expectations. The deterrence power of the low fine, by contrast, could not be justified by any reasonable parameterization of selfish preferences. Even more surprising was what happened when the experimenter proposed the sanction mechanism to the players but they voted against it. Extraction fell sharply at first, and

then cooperation slowly unraveled back to its original low level.[3] One may infer the norm was internalized by some players even when it was not enforced. Without enforcement, moralization seemed to vanish over time.

Fitting our model to the experimental data we find that most selfish players adopt a cooperative type when the experimenter prescribes extracting one token. We also find that fewer people internalize the norm if the norm is enforced. We rationalize these findings as follows. Initially, there are very few cooperative players. When the experimenter prescribes the norm, the number of cooperators rises: we term this a "prescriptive effect". If the players learn that the norm will be enforced, the number of cooperators immediately falls, although it still remains higher than its initial level. Enforcement seems to relieve some players from the guilt of infringement: we call this a "guilt relief effect."

The existence of a prescriptive and a guilt relief effect also reconcile our findings with Gneezy and Rustichini's. In their experiment, the imposition of a fine alleviated the parent's guilty feelings. But, as parents knew beforehand that it was their duty to pick up their children on time, the prescriptive effect was absent. The result was a crowding out of cooperation. In our experiment, both effects act together. The prescriptive effect dominates the guilt relief effect, so cooperation crowds in.

Finally, our study reveals that a player who cooperates conditionally under no fine is likely to cooperate unconditionally when a fine is in force. This is probably because the fine relieves her of the desire to retaliate against uncooperative players in the only way she can: by ceasing to cooperate herself.[4]

---

[3] Ostrom, Gardner and Walker (1994) and Cárdenas (2000) also find unraveling in CPR games. The unraveling of cooperation has been reported in public good experiments as well. The earliest reports are in Kim and Walker (1984), and in Isaac, McCue and Plott (1985). See Fehr and Gaechter (2000) for a more recent treatment of the subject.

[4] Andreoni (1995) advanced a similar hypothesis in the context of public good games.

Our findings solve the two puzzles in the experimental results: the increase and later erosion of cooperation when commoners vote against the imposition of a fine, and the high deterrence power of low fines. When fines are rejected, moralization explains the increased cooperation; violations (accidental or not), coupled with reciprocal preferences, account for the erosion. Low fines, on the other hand, induce players to cooperate irrespective of the behavior of their peers. A spiral of negative reciprocation is prevented and, as a result, cooperation becomes stable.

2. A MODEL OF COMMON POOL RESOURCE GAMES

N persons play a finitely repeated common pool resource (CPR) game. The game is repeated *T* times. At the beginning of each round, every player decides privately how many tokens to extract from a common pool; the minimum being one token, and the maximum $x_{\max}$ tokens. Let $x_{it} \in \{1,...,x_{\max}\}$ be the number of tokens player $i \in \{1,...,N\}$ takes from the common pool in round $t \in \{1,...,T\}$.

A player's payoff from extraction depends positively on the number of tokens she extracts and negatively on the aggregate level of extraction. Denote by $\pi(x_{it}, \overline{x}_{-it})$ player *i*'s payoff from extraction in round *t*, where $\overline{x}_{-it} = \frac{1}{N-1}\sum_{j \neq i} x_{jt}$. Function $\pi(x_{it}, \overline{x}_{-it})$ is increasing in x_{it} and decreasing in $\overline{x}_{-it}$. The sum of the payoffs of all players is maximized when they all extract the minimum amount (one token).

Assume that the social norm is to extract one token. At the end of each round, an external authority inspects each player with probability $p_t \in [0,1)$. If the authority discovers that a player violated the social norm, he fines that player with an amount $f_t \geq 0$ for every token she extracted in excess of one (the authority then casts the collected fine into the sea). Thus, the expected material payoff of player *i* in round *t* is $\pi(x_{it}, \overline{x}_{-it}) - p_t f_t \cdot (x_{it} - 1)$.

There are three types of players: selfish (S), unconditional cooperators (UC), and conditional cooperators (CC). A selfish player derives utility only from her own consumption. An unconditional cooperator also enjoys consumption, but feels guilty when she extracts more than the amount prescribed by the norm, an idea we borrow from Bowles and Gintis (2002). Finally, a conditional cooperator enjoys consumption and feels guilty when she infringes the norm, though her guilt diminishes as group extraction increases. Conditional cooperators relate our model to those of reciprocal preferences, such as Rabin's (1993), and Dufwenberg and Kirchsteiger's (2004). Fischbacher, Gaechter and Fehr (2004) report conditional cooperation is the most common behavior in

one-shot public goods games, and that suggests it may also be common in CPR games. The effect of diminishing guilt on norm compliance was recently explored by Lin and Yang (2005).

Let $u(x_{it}, \bar{x}_{-it}, \theta_{it})$ be the utility function of player *i* in round *t* when she is of type $\theta_{it} \in \{\mathrm{S}, \mathrm{UC}, \mathrm{CC}\}$. We define $u(x_{it}, \bar{x}_{-it}, \theta_{it})$ as follows:

$$u(x_{it}, \bar{x}_{-it}, \theta_{it}) = \pi(x_{it}, \bar{x}_{-it}) - p_t f_t \cdot (x_{it} - 1)$$
$$- \mathrm{I}(\theta_{it} \neq \mathrm{S}) \beta_1 \pi_{\max} \frac{x_{it} - 1}{x_{\max} - 1} \left\{ 1 - \mathrm{I}(\theta_{it} = \mathrm{CC}) \beta_2 \frac{\bar{x}_{-it} - 1}{x_{\max} - 1} \right\},$$

where $\pi_{\max} = 880$ is the maximum material payoff a player may obtain in one round, β_1 and β_2 are positive constants, and function $\mathrm{I}(s)$ is 1 if statement *s* is true, and 0 otherwise. This means that an unconditional cooperator that extracts $x_{\max}$ tokens experiences guilt equivalent to β_1 times $\pi_{\max}$. A conditional cooperator feels as guilty as an unconditional one, provided everybody else abides by the norm and extracts one token. If $\beta_2 > 1$ and aggregate extraction is high, a conditional cooperator will enjoy violating the norm.

We allow a player's type to depend on institutions. We shall postpone the definition of institutions until the next section. For the time being, bear in mind that institutions may comprise such things as the enforcement of a norm by an external authority, and that institutions may change over time. Each player is born a certain type (S, UC, or CC), and she may only switch types when institutions change. If we denote the institution in force during round *t* as ω_t, that means that $\theta_{it} = \theta_{i(t-1)}$ unless $\omega_t \neq \omega_{t-1}$. Denote as $\mathrm{q}(\theta \mid \omega)$ the probability that a player will become type θ at the beginning of institutional regime ω.

Player *i* will choose with higher probability those actions that give her a higher expected utility. Let ε_{it} be her expectation of how much other players will extract in

round t. The probability that player i will extract x tokens on round t is a logistic function of her expected utilities:

$$P_{it}(x) = \frac{\exp \lambda \cdot u(x, \varepsilon_{it}, \theta_{it})}{\sum_{y=1}^{x_{\max}} \exp \lambda \cdot u(y, \varepsilon_{it}, \theta_{it})},$$

where $\lambda \geq 0$ represents her tendency to maximize. If $\lambda = 0$, the player will choose all extraction levels with equal probability. As λ approaches infinity, the player will tend to extract with probability one the number of tokens that maximizes her utility.

Finally, player i updates her estimate of how much others will extract as she observes what they actually do. Player i's expectations are adaptive:

$$\varepsilon_{it} = \begin{cases} \varepsilon(\omega_t) & \text{if } t = 1 \text{ or } \omega_t \neq \omega_{t-1} \\ \phi\varepsilon_{i(t-1)} + (1-\phi)\overline{x}_{-i(t-1)} & \text{otherwise,} \end{cases}$$

where $\phi \in [0,1]$ measures the persistence of expectations, and $\varepsilon(\omega)$ is an exogenous initial expectation. Initial expectations depend on ω because a change in institutions may induce a change in what players expect. Stochastic choice combined with adaptive learning make our model a close cousin of Camerer and Ho's (1999) EWA learning model. Our work is also linked to Janssen and Ahn's (2003), who fit an EWA learning model to the results of two public good experiments. They find that heterogeneous preferences are essential to account for their experimental evidence.

The steady state of $\overline{x}_t$, the mean extraction level of the group in round t, has one important property. If there are no conditional cooperators in a group, $\overline{x}_t$ has a unique stable steady state. But, if enough conditional cooperators are added to the mix, the reciprocal nature of their preferences may cause a second steady state to emerge (a feature shared by other models of reciprocal preferences, like Rabin's [1993], and Lin and Yang's [2005]). The intuition is simple: if conditional cooperators expect group extraction to be low, they will be inclined to extract few tokens. On the other hand, if

they expect a high group extraction, conditional cooperators will tend to extract many tokens. Hence, there will be two attracting poles of self-fulfilling expectations: one where players cooperate a lot, and another with little cooperation.

3. A COMMON POOL RESOURCE EXPERIMENT

In our common pool resource (CPR) experiment all subjects were adult villagers from five communities in Colombia. The communities exploited a common resource, such as fish or water. To control for the effect of kin altruism, no two members of the same household were admitted into the same experimental group.

Here we briefly describe the experiment and discuss its results.[5]

3.1 Experimental design

Groups of five persons ($N = 5$) play the CPR game of the previous section. The game is repeated twenty times ($T = 20$), and the players know the number of repetitions beforehand. In each round every player decides privately how many tokens to extract from a common pool; the minimum being one token, and the maximum, eight ($x_{\max} = 8$). The experimenter then informs players of the aggregate level of extraction, but does not reveal individual levels. Player *i*'s payoff from extraction in round *t* is given by

$$\pi(x_{it}, \overline{x}_{-it}) = 800 + 40x_{it} - \frac{5}{2}x_{it}^2 - 80\overline{x}_{-it}.$$

A simple calculation shows that a player maximizes her material payoff by extracting eight tokens. The aggregate payoff, on the other hand, is maximized when each player extracts only one. After the final round, players cash their tokens. Prizes range between one and two days' wages.

At the end of round 10 the experimenter may introduce the following sanction mechanism: after each round he will randomly inspect one player; if he discovers that the player took more than one token, he will fine her in private. The experimenter may force the sanction mechanism on the players, or let them vote on it. In either case, he first

[5] See Cárdenas (2004) for a detailed description of the experiment.

explains to the players that having a fine is in their best interests because it discourages extracting more than one token, and because when everybody extracts only one token the material welfare of each player is maximized.

We identify four institutions:

NF: No fine has ever been imposed on, or approved by, the players.

HF: A high fine regime is in force.

LF: A low fine regime is in force.

RF: A fine regime was proposed to, and rejected by the players.

We do not distinguish between fines imposed by the experimenter and fines approved by player vote, because the distinction made no difference to the behavior of the players.[6] Since the experimenter may affect the preferences of players when he proposes a fine and they vote against it, we do distinguish between the no fine (NF) and the rejected fine (RF) regimes.

Let $f(\omega)$ be the fine in force when the institution is ω:

$$f(\omega) = \begin{cases} 0 & \text{if } \omega \in \{\text{NF,RF}\} \\ 175 & \text{if } \omega = \text{RF} \\ 50 & \text{if } \omega = \text{LF.} \end{cases}$$

The expected material payoff of player *i* in round *t* is therefore $\pi(x_{it}, \overline{x}_{-it}) - \frac{1}{5} f(\omega_t) \cdot (x_{it} - 1)$, where $\frac{1}{5}$ is the probability she will be inspected.

[6] We performed a Kruskall-Wallis test on the hypothesis that mean extraction levels are the same under voted and externally-imposed fine regimes. The test for high fines produced a p-value of 0.78. The test for low fines produced a p-value of 0.80.

Sixty-four groups of players received one of four different treatments:

Control: (8 groups) The institution is NF for all twenty rounds.

High fine: (14 groups) The institution is NF for the first ten rounds, and HF for the last ten rounds.

Low fine: (26 groups) The institution is NF for the first ten rounds, and LF for the last ten rounds.

Rejected Fine: (16 groups) The institution is NF during the first ten rounds, and RF for the last ten rounds.

The standard prediction for this version of the CPR game is its subgame perfect equilibrium. Table 1 summarizes the predictions for each institution. According to the predictions, only a high fine should have enough deterrence power to reduce individual extraction to its socially optimal level. Also, in the case of the low fine and the rejected fine institutions, the equilibrium extraction levels are far above the socially optimal level (one token). Thus, if one observes players complying with the social norm, one should feel less inclined to deem their compliance a mistake. Note that the equilibrium levels of extraction are close to, or coincide with, either the minimum or the maximum number of tokens that players are allowed to extract. This is intended to avoid the confusion that may arise among players if the equilibria were interior.

Institution	Predicted extraction
No fine	8
High Fine	1
Low Fine	6
Rejected Fine	8

Table 1 Predicted levels of extraction.

3.2 Results of the CPR experiment

Figure 1 displays the aggregate behavior of players under each institutional regime. Note that:

1. Groups start at low levels of cooperation, extracting about 4.5 tokens on average. The mean level of extraction remains fairly constant during the first 10 rounds. In the control treatment, extraction stays around 4.5 tokens until the end of the game.

2. Under all treatments other than the control, cooperation increases on round 11. The social optimum, however, is never reached. Nonetheless, extraction falls even when the players vote against the fine.

3. Cooperation remains high after round 11 only when a fine, be it high or low, is in force. If the players reject the fine, cooperation slowly unravels.

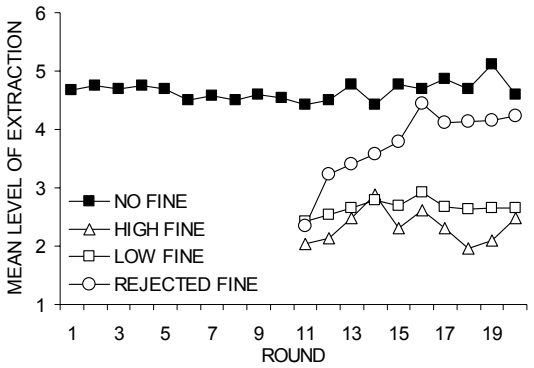


Figure 1 Experimental results: aggregate behavior.

Compare the results of the experiment with the predictions of Table 1. According to the predictions, initial extraction levels should be 60% higher than they actually are. Under the high fine, extraction should drop to one. Instead, it stays over two. We expected a low fine to exert little deterrence. However, the low fine and the high fine work almost as well. A rejected fine should have no effect whatsoever. But it has.

Table 2 shows mean extraction levels under each institution, along with group and individual deviations from the mean. The high individual deviations suggest that players randomize or experiment.

	Institution			
	No fine	High fine	Low Fine	Rejected fine
Mean extraction	4.6	2.3	2.7	3.7
Group deviation	2.3	1.9	2.1	2.3
Average individual dev.	1.8	1.0	1.2	1.8

Table 2 Summary statistics from the CPR experiment.

Figure 2 shows histograms of individual extraction levels under different treatments. Under both fine treatments extraction is concentrated in the vicinity of one token. The histogram representing the no fine treatment is almost flat. If all players were identical, that would imply that they choose strategies completely at random, as if indifferent to material payoffs. A complementary explanation for the flatness is that players are heterogeneous along the moral dimension: some feel strongly that they should not take more than one token; others have no qualms and maximize their material payoff by taking eight. Also note how the histograms that represent the rejected fine treatment get flatter on rounds 15 and 20, as cooperation deteriorates.

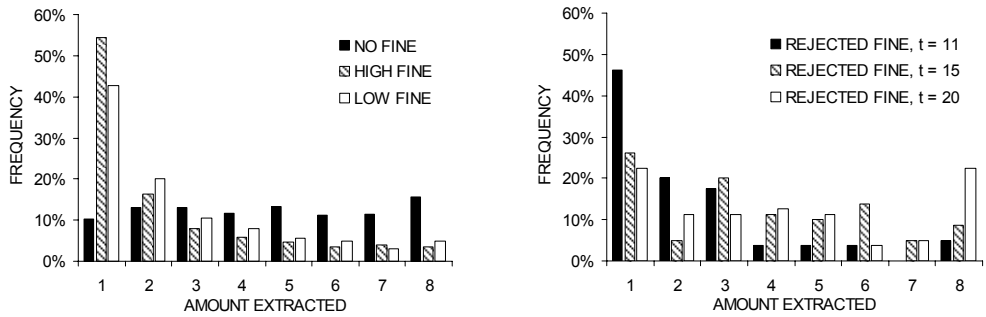


Figure 2 Experimental results: distribution of individual extraction levels.

The unraveling process is better understood by examining, one by one, the groups that rejected a fine. Figure 3 shows four such groups. Group 1 extracts a high amount from the first period until the end. Groups 2, 3, and 4 initially extract a low amount, but only group 4 cooperates until the last round. The most common pattern of behavior is represented by groups 2 and 3: both start by cooperating, but somewhere along the way they abruptly cease to cooperate (first group 2 and later group 3). The smooth, concave line representing the rejected fine treatment in Figure 1 results from averaging many groups like 2 and 3.

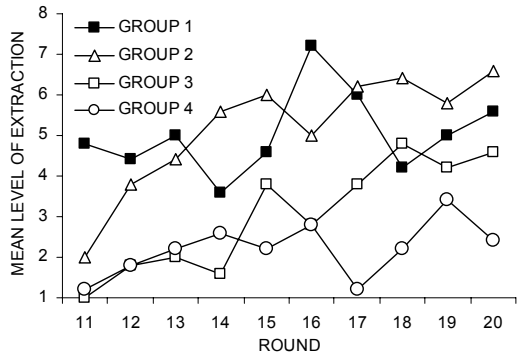


Figure 3 Experimental results: groups that voted against a fine.

4. MODEL ESTIMATION AND SIMULATION

We used maximum-likelihood to estimate the parameters of our model: λ, β_1, β_2, ϕ, $\varepsilon(\cdot)$, and $\mathrm{q}(\cdot|\cdot)$ (see the Appendix for a detailed account of the estimation procedure). Recall that λ is the players' tendency to maximize, β_1 and β_2 determine the social preferences of cooperators, $\varepsilon(\omega)$ is the initial expectation of players under institution ω, constant ϕ measures the persistence of expectations, and $\mathrm{q}(\theta|\omega)$ is the probability that a player will become type θ at the beginning of institutional regime ω. We based our estimations on the outcomes of the first 19 rounds of play, and left the final round to test the predictive accuracy of our model.

To simplify estimation, we made two assumptions regarding initial expectations:

1. If $\omega \in \{\mathrm{NF,HF,LF}\}$, $\varepsilon(\omega)$ coincides with a stable steady state of $\overline{x}_t = \sum_{i=1}^{N} x_{it}$ under institution ω. Two conditions must hold for $\varepsilon(\omega)$ to be a stable steady state. First, the average level of player extraction when they expect others to extract $\varepsilon(\omega)$ must coincide with $\varepsilon(\omega)$. That is, the following condition must hold:

$$\sum_{\theta}\left\{q(\theta|\omega)\sum_{x=1}^{x_{\max}}\frac{x\exp\lambda\cdot u(x,\varepsilon[\omega],\theta)}{\sum_{y=1}^{x_{\max}}\exp\lambda\cdot u(y,\varepsilon[\omega],\theta)}\right\}-\varepsilon(\omega)=0$$

   Second, the derivative of the left hand side of the equation with respect to $\varepsilon(\omega)$ must be negative.

2. If $\omega = \mathrm{RF}$, $\varepsilon(\omega)$ is a convex combination of the stable steady states of $\overline{x}_t$.

The first assumption is justified by the fact that mean extraction levels remain fairly constant through all rounds under the no fine, high fine and low fine institutions

(see Figure 1). With assumption number two we intend to capture the confusion that may arise among players when there is more than one steady state (as Figure 3 suggests).

Table 3 displays the estimated values of λ, β_1, β_2, and ϕ. Table 4 displays the estimated distribution of types, $q(\cdot|\cdot)$, under each institution. Finally, Table 5 displays the estimated initial expectations.

Parameter	Estimate	Parameter	Estimate
λ	0.0030 (0.0007)	β_1	4.00 (2.45)
ϕ	0.50 (0.03)	β_2	4.00 (0.00)

Table 3 Estimated parameters: λ, β_1, β_2, and ϕ.

		Institution (ω)			
		No fine	High fine	Low Fine	Rejected fine
Player types (θ)	Selfish	88% (2%)	20% (2%)	21% (5%)	2% (2%)
	Unconditional Cooperators	7% (2%)	63% (7%)	57% (2%)	30% (6%)
	Conditional Cooperators	5% (1%)	17% (9%)	22% (8%)	67% (4%)

Table 4 Estimated distribution of types, $q(\theta|\omega)$.

	Institution (ω)			
	No fine	High fine	Low Fine	Rejected fine
$\varepsilon(\omega)$	4.7	2.0	2.4	2.2
Stable steady states	4.7	2.0	2.4	1.7; 5.8

Table 5 Estimated initial expectations and implied stable steady states.

Perhaps the most striking result is the effect the institutional environment has on the distribution of types (Table 4). Under the no fine institution, only 12% of the players are cooperative. When a fine (high or low) is in force, the percentage rises to about 80%, and to 98% when the players reject a fine regime. Also, our results reveal that the

enforcement of the norm induces more players to cooperate unconditionally: unconditional cooperators are 30% when a fine is rejected, and approximately 60% when a fine (high or low) is in force.[7] Why? We hypothesize that fines relieve the cooperative player of the desire to retaliate against uncooperative ones in the only way she can: by ceasing to cooperate herself.

Table 5 also shows the stable steady states of $\overline{x}_t$ implied by the estimated parameters under each institutional environment. There is a unique stable steady state under the no fine, high fine and low fine institutions. That explains why players subject to those institutions rapidly cluster around the long run value of $\overline{x}_t$: where equilibria are unique, there is little scope for confusion. On the other hand, $\overline{x}_t$ has two stable steady states when players vote against the imposition of a fine. In that scenario, the intervention of the experimenter at the end of round 10 plays two complementary roles: moralizing players and coordinating expectations. In Schelling's (1960) terms, the experimenter makes the low extraction equilibrium a focal point.[8] The unraveling of cooperation is the transition from the high cooperation equilibrium to the low cooperation one.

In our model, fines affect behavior through two channels: material deterrence and moralization (i.e., the externally-induced change from selfish to cooperative type). To measure both channels separately, we simulated again the high and low fine regimes, but this time we kept preferences unaltered (i.e. using the NF distribution of types). In the new simulations, the low fine had no perceptible effect on extraction. The high fine, on the other hand, reduced the mean extraction level from 4.6 to 4.1 tokens. These results imply that, in our experiment, moralization accounted for the whole effect of low fines,

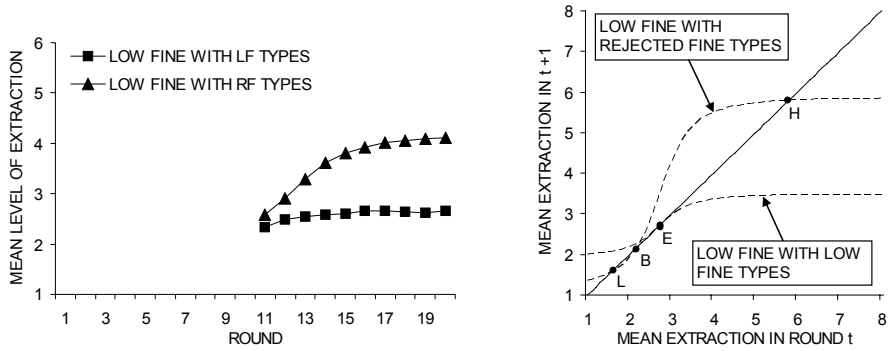
[7] These results are robust. We made 100 bootstrap estimations of the model, taking each group history as an independent observation. In all estimations we found that: $q(S \mid NF) > q(\theta \mid \omega)$ for all $\omega \in \{HF,LF,RF\}$, $q(S \mid RF) < q(\theta \mid \omega)$ for all $\omega \in \{NF,HF,LF\}$, and $q(CC \mid RF) > q(CC \mid \omega)$ for all $\omega \in \{HF,LF\}$.

[8] McAdams and Nadlery (2005) study coordination in a hawk-dove game. They find, as we do, that externally imposed norms signal focal points.

and for 78% of the effect of high fines. Material deterrence played but a minor role, and that explains why both fines work almost as well.

Our findings also explain the increase and later erosion of cooperation when commoners vote against the imposition of a fine. When players reject a fine, moralization explains the increased cooperation. Violations, coupled with a high share of conditional cooperators, account for the unraveling.

A low fine is able to prevent the unraveling by changing the nature of cooperation: from predominantly conditional to predominantly unconditional. The stabilizing role of unconditional cooperators becomes clear when one simulates what would happen if the experimenter imposed a low fine, but the distribution of types changed to that of the rejected fine institution (with most cooperators of the conditional kind). The triangle line in Figure 4a shows the result: extraction falls in round 11 and then slowly unravels.



(a) Dynamics. (b) Phase diagram.

Figure 4 The stabilizing role of unconditional cooperators.

To understand the mechanics of unraveling, look at Figure 4b. The figure displays two phase diagrams of mean extraction under a low fine: one with RF types and another one with LF types (see Table 4). With RF types the system has two steady states: one where extraction is low (L), and another one where extraction is high (H). The low extraction steady state has a small basin of attraction (to the left of point B), whereas the

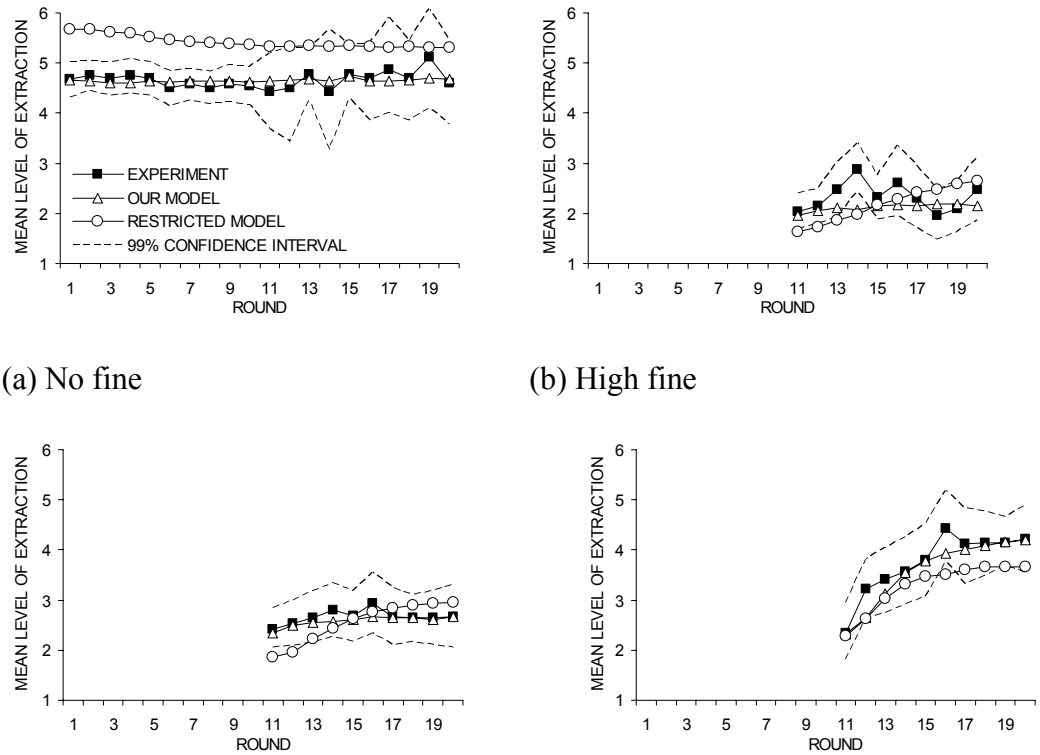
high extraction steady state has a large basin of attraction (to the right of point B). Since players randomize, the system will not stay close to L for a long time. Even a small shock will push the extraction level passed B and into the basin of attraction of H. As a result, cooperation will unravel. On the other hand, with LF types there is only one stable steady state, so extraction remains low regardless of shocks.

We've seen how a low fine stabilizes cooperation by preventing a spiral of negative reciprocation: when the norm is enforced, cooperation tends to be unconditional, and that eliminates the high extraction steady state that arises when the norm is prescribed but not enforced. Because the imposition of a low fine may moralize selfish players and induce unconditional cooperation, the "fine enough or don't fine at all" policy prescription of Lin and Yang (2005) must be qualified.

To test the descriptive accuracy of our model, we simulated each treatment 500 times, using the estimated parameters as inputs. Figure 5 displays the aggregate behavior of players under each treatment, actual and simulated. Table 6 shows mean extraction levels under each institution, along with group and individual deviations from the mean. The table pairs actual and simulated values. Figure 6 compares the actual and simulated histograms of individual extractions. The results of the experiment and the output of the simulation are very similar. Our model provides good account of the player's behavior, at both the group and the individual level.

		Institution			
		No fine	High fine	Low Fine	Rejected fine
Mean extraction	*Actual*	4.6	2.3	2.7	3.7
	Sim.	4.7	2.1	2.6	3.6
Group deviation from the mean	*Actual*	2.3	1.9	2.1	2.3
	Sim.	2.4	1.9	2.3	2.9
Average individual deviation	*Actual*	1.8	1.0	1.2	1.8
	Sim.	2.0	1.0	1.2	1.5

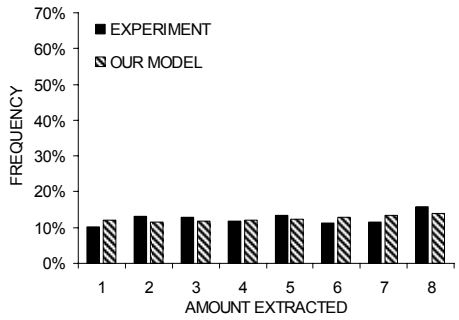
Table 6 Summary statistics, actual and simulated, from the CPR experiment.



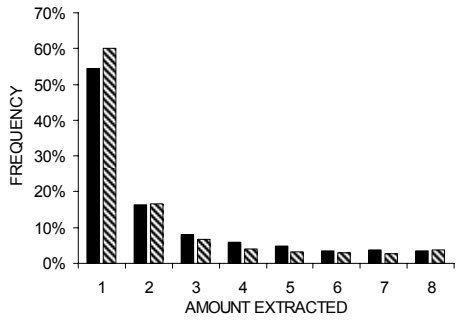
(a) No fine (b) High fine

(c) Low fine (d) Rejected fine

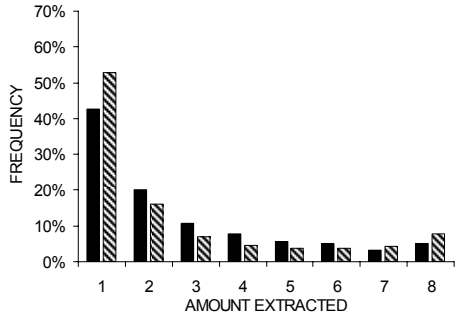
Figure 5 Mean levels of extraction, actual and simulated



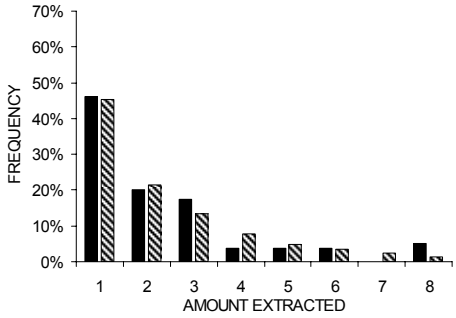
(a) No fine



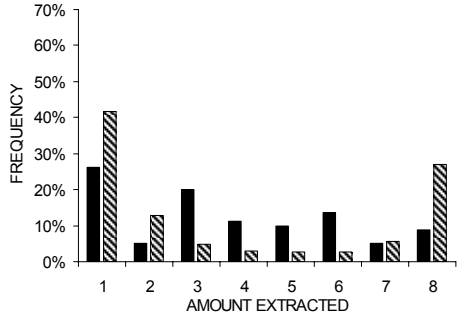
(b) High fine



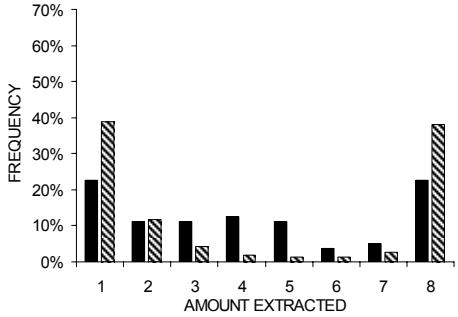
(c) Low fine



(d) Rejected fine, $t = 11$.



(e) Rejected fine, $t = 15$.



(f) Rejected fine, $t = 20$.

Figure 6 Distribution of individual levels of extraction, actual and simulated.

Next, we re-estimated our model subject to the restriction that preferences are not state-dependent (i.e., forcing $q(\theta \mid NF) = q(\theta \mid HF) = q(\theta \mid LF) = q(\theta \mid RF)$, for all $\theta \in \{S, UC, CC\}$).[9] Using a likelihood ratio test we were able to reject, at a 99% confidence level, the hypothesis that the distribution of types does not change across treatments. [10] We also simulated the restricted model, using the estimated parameters as inputs, and it was unable to accurately mimic the experimental evidence (see Figure 5).

Finally, we used our model and the restricted model to predict the amount extracted by each of the 320 experimental subjects in the last round of play. To predict the extraction of a particular player, we used the posterior probability of that player being of type $\theta \in \{S, UC, CC\}$, given the priors in $q(\theta \mid \omega)$ and the behavior of the player and of the other members of his group during the first 19 rounds of play. Table 7 displays the mean prediction errors for both models under each institution. Our model outperformed the restricted model in all scenarios. We conclude that, in our CPR experiment, institutions influenced the social preferences of players.

	Institution			
	No fine	High fine	Low Fine	Rejected fine
Our model	0.75	0.71	0.68	0.86
Restricted model	0.81	0.78	0.79	0.92

Table 7 Mean errors of prediction for our model and for a model without state-dependent preferences.

[9] Estimated parameters for the restricted model: $\lambda = 0.003$, $\beta_1 = 4.5$, $\beta_2 = 4.25$, $\phi = 0.5$; $q(S \mid \theta) = 11\%$ and $q(UC \mid \theta) = 29\%$, for all θ; $\varepsilon(NF) = 5.7$, $\varepsilon(HF) = 1.7$, $\varepsilon(LF) = 1.8$, $\varepsilon(RF) = 2.2$.

[10] The log-likelihoods of the unrestricted and restricted models are $L_U = -11467{,}14$ and $L_R = -12202.57$. The likelihood ratio statistic is $2(L_U - L_R) = 1470.86 > \chi^2_6(.99) = 16.81$, so we reject the hypothesis.

5. CONCLUDING REMARKS

Authorities may influence social preferences when they prescribe and enforce social norms. We found, in a CPR experiment, that the external imposition of a norm affected preferences in two ways.

First, by moralizing players. A speech by the experimenter sufficed to induce players to cooperate. How? By sowing in them the seed of guilt. Aristotle argued in his Nichomachean Ethics that effective laws worked by inculcating habits in citizens, that is, by moralizing them.[11] Our results remind us that his argument is still relevant today.

Second, our model revealed that the enforcement of the norm affected the nature of moral sentiments. If the norm was enforced, players tended to comply with it irrespective of how others behaved. But if enforcement was absent, players conditioned their compliance on the good behavior of their peers.

Our results indicate that the extent of moralization is not the same when a norm is externally enforced as when it is not. In our experiment, more players became cooperative in the absence of enforcement. Why? We hypothesize that two effects operated simultaneously. First, a prescriptive effect, which always tends to increase cooperation. Second, a guilt relief effect that appears when norms are enforced, and which tends to decrease cooperation. Unfortunately, the guilt relief effect never appears alone in our experiment, so we can only infer its existence and measure it indirectly.| Further research is needed to untangle the guilt relief effect from the prescriptive effect.

Our results also bring attention to the dynamic effects of enforcement. Conditional cooperation makes compliance fragile: a single rotten apple may spoil the whole box (and the addition of many good apples cannot restore it). In our experiment, a small fine sufficed to stabilize cooperation by making more players cooperate unconditionally, and that prevented the spread of moral degradation. Consider the

[11] The word moral stems from Latin moralis, meaning custom.

implications for governmental corruption. Corrupt officers are hard to detect, so the expected punishment is often small compared to the potential gains from corruption. The occasional jailing of corrupt officers may nonetheless stabilize moral behavior, if it prevents them from thinking: “everybody else is doing it, so why can't we?”

Further research is needed to determine when the enforcement of a norm will shield moral behavior from resentment or from “bad examples.” For instance: sanctions were weakly enforced in our experiment, but they were fair. If some commoners were immune to punishment, punishment might cease to quench feelings of revenge; it would no longer serve to stabilize cooperation. Similarly, even if few people are beyond the reach of the law, the law may lose its effectiveness.

The way a low fine sustains cooperation may be analogous to the way the yellow card keeps the peace on a football field. Without the card, violence escalates after the first kick to the shin; it makes no difference if the kick was intentional or accidental. Perhaps the card gives football players the sensation that bad behavior does not always go unpunished, and that suppresses their impulse to seek their own justice. Being close substitutes for reciprocation, low fines and yellow cards may sometimes stabilize norm compliance in a world of feeble social order.

REFERENCES

Andreoni, James. 1995. Cooperation in public-goods experiments: kindness or confusion? *American Economic Review*, 85(4): 891-904.

Aristotle. 1962. *Nichomachean Ethics*. Indianapolis: Bobbs-Merrill.

Bowles, Samuel. 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, 36(1): 75-111.

Bowles, Samuel. 2005. Social preferences and public goods: are good laws a substitute for good citizens? Mimeo.

Bowles, Samuel, and Herbert Gintis. 2002. Social capital and community governance. *Economic Journal*, 112(483): 419-36.

Camerer, Colin, and Teck-Hua Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4): 827-74.

Cárdenas, Juan Camilo. 2004. Groups, commons and regulations: experiments with villagers and students in Colombia. *Psychology, Rationality and Economic Behavior: Challenging Standard Assumptions*. Bina Agarwal and Alessandro Vercelli, editors. International Economics Association.

Cárdenas, Juan Camilo, John Stranlund and Cleve Willis. 2000. Local environmental and institutional crowding-out. *World Development*, 28(10): 1719-1733.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. A theory of sequential reciprocity. *Games and Economic Behavior*, 47: 268-298.

Falk, Armin, and Michael Kosfeld. 2004. Distrust - the hidden cost of control. IZA Working Paper 1203.

Fehr, Ernst, and Simon Gaechter. 2000. Cooperation and punishment in public good experiments. *American Economic Review*, 90(4): 980-994.

Fischbacher, Urs, Simon Gaechter and Ernst Fehr. 2004. Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters*, 71(3): 397-404.

Gneezy, Uri, and Aldo Rustichini. 2000. A fine is a price. *Journal of Legal Studies*, 29(1):1--17.

Isaac, Mark, Kenneth McCue and Charles Plott. 1984. Public goods provision in an experimental environment. *Journal of Public Economics*, 26(1): 51-74

Janssen, Marco, and T.K. Ahn. 2003. *Adaptation vs. Anticipation in a Public Good Game*. Mimeo.

Kim, Oliver, and Mark Walker. 1984. The free rider problem: experimental evidence. *Public Choice*, 43(1): 3-24.

Lin, Chung-Cheng, and C.C. Yang. 2005. Fine enough or don't fine at all. *Journal of Economic Behavior and Organization*, forthcoming.

McAdams, Richard, and Janice Nadlery. 2005. Testing the focal point theory of legal compliance: expressive influence in an experimental hawk/dove game. *Journal of Empirical Legal Studies*, forthcoming.

Ostrom, Ellinor, Roy Gardner and James Walker. 1994. *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Michigan.

Rabin, Matthew. 1993. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5): 1281-1302.

Titmuss, Richard. 1969. *The Gift Relationship: From Human Blood to Social Policy*. Routledge, London.

APPENDIX: ESTIMATION PROCEDURE

Here we describe how we estimated the parameters of our model and its version without state-dependent preferences. First, we grilled the parameter space and restricted the search to that grill (see Table 8).

Parameter	Grill				
λ	0.0005	0.0010	…	0.0175	0.0200
β_1, β_2	0.25	0.50	…	9.75	10.00
ϕ	0.00	0.25	0.5	0.75	1.00
$\mathrm{q}(\theta \mid \omega)$	0.00	0.01	…	0.99	1.00
$\varepsilon(\omega)$	1.0	1.1	…	7.9	8.0

Table 8 Parameters grill.

Second, we imposed three restrictions to the candidate estimates of the parameter vector $\Theta = \left[\lambda, \beta_1, \beta_2, \phi, \mathrm{q}(\cdot \mid \cdot), \varepsilon(\cdot \mid \cdot)\right]$. These restrictions are:

1. If $\omega \in \{\mathrm{NF,HF,LF}\}$, $\varepsilon(\omega)$ coincides with a stable steady state of $\overline{x}_t = \sum_{i=1}^{N} x_{it}$ under institution ω.

2. If $\omega = \mathrm{RF}$, $\varepsilon(\omega)$ is a convex combination of the stable steady states of $\overline{x}_t$.

3. The candidate parameter vector should be able to reproduce the unraveling of cooperation under the rejected fine institution. To verify this restriction, we used each candidate parameter vector to simulate the rejected fine institution 500 times. We then checked that the simulated mean levels of extraction over ten rounds of play were within the 99% confidence intervals calculated from the experimental data (the confidence intervals are displayed in Figure 5d).

In the case of the model without state-dependent preferences, we added an additional restriction: $q(\theta \mid NF) = q(\theta \mid HF) = q(\theta \mid LF) = q(\theta \mid RF)$, for all $\theta \in \{S, UC, CC\}$.

Finally we evaluated the log-likelihood function of the model with every candidate parameter vector that satisfied the aforementioned restrictions, and selected the parameter vector that produced the higher value.