Institutions Influence Preferences: Evidence from a Common Pool Resource Experiment

Carlos Rodríguez-Sickert
Universidad Católica de Chile
Instituto de Sociología

Ricardo Andrés Guzmán*
Universidad Católica de Chile
Instituto de Economía

Juan Camilo Cárdenas†
Universidad de los Andes
Facultad de Economía

December 16, 2005

Abstract

We model the dynamic effects of external enforcement on the exploitation of a common pool resource. Fitting our model to the results of 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.

*Corresponding author. Email: rnguzman@puc.cl.

†We are deeply indebted with Sam Bowles, Rajiv Sethi, Marcos Singer, Rodrigo Harrison and Rodrigo Troncoso. Their unconditional cooperation substantially improved this article.

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 may exert on social preferences. Often, that assumption is harmless. Other times, its consequences may range from unexpected to disastrous. English authorities learned this hard way. They decided to incentivize blood donations by paying donors. Instead of increasing, blood donations plummeted (Titmuss 1969).

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

In all the examples above, institutions designed to incentivize cooperation displaced moral behavior[1]. One is tempted to conclude such institutions will always have a negative effect on social preferences. But it needn't be so. Here is a counterexample.

The metro of Santiago (Chile) is among the cleanest in the world. Some of the trains are more than twenty five years old; still, they shine like new. The people of Santiago take pride on their metro, and everybody cooperates to keep it clean. The metro seems even cleaner when compared to the buses that roam the surface: the buses' seats are torn, their floors are littered, their interiors are covered with graffiti. A curious thing, since the passengers of the buses and the passengers of the metro are the very same people. If you ask someone why he doesn't dirty the metro, a typical answer would be: "because you shouldn't." A minute later you bump into the same person

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

carving his name on the inside of a bus.

The reason why the barbarians who ride the buses get civilized when they enter the metro lies on institutions. Dirty a bus and no one will care; doesn't matter if you are seating next to a policeman. On the other hand, dirtying the metro is prohibited (and socially scorned). If a security guard catches you writing on a wall, you will surely be evicted, and maybe have to pay a small fine. Chances of being caught, however, are close to zero. Very few guards watch the metro, and more often than not they are looking in the wrong direction. Thus, the prohibition is effective even though it is hardly enforced (and the authority is being rational by not enforcing it, since almost nobody infringes it!). Instead of hitting passengers in their pocketbooks, the "thou shall not dirty" commandment seems to act directly on the utility lobule of their brains[2]. The passengers internalize the social norm: they are moralized.

In this paper we model the dynamic effects of external enforcement on the exploitation of a common pool resource (CPR). We are specially interested on the moralizing effect of low fines (as in the metro of Santiago case). The ingredients of our 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.* The type of a player depends on the institutional environment.

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 much her peers will extract, and updates that estimate as she sees 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 a CPR game each player

[2]The cleanness of their metro puzzles the inhabitants of Santiago. Many believe that subliminal messages in the background music induce people to keep the metro clean (seriously). So far as we know, that theory is a hoax.

chooses privately how many tokens she will extract from a common pool. The material payoff of a player depends positively on the number of tokens she extracts and negatively on the aggregate level of extraction. Thus, individual and social interest conflict.) In some treatments the experimenter fined players he surprised extracting more than one token (he applied sanctions 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. We expected the effect of the high fine, but the deterrence power of the low fine 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 and then slowly unraveled back to its original 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 internalize the norm (*i.e.* they adopt a cooperative type) after the experimenter prescribes extracting one token. We also find that internalization is strongest when the norm is not enforced. That result is consistent with Gneezy and Rustichini's: people are less willing to embrace the norm when they feel they can buy their way out of it. Finally, our study reveals that a player who cooperates conditionally under no fine is likely to cooperate unconditionally when a fine is in force; probably because the fine relieves her from the urge to retaliate uncooperative players in the only way she can: by ceasing to cooperate herself[4].

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. The way a low fine sustains cooperation is analogous to the way the yellow card keeps peace in a football field. Without the card, violence escalates after the first kick in the shin; it makes no difference if the kick was inten-

[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.

tional or an accident. The card gives players the sensation that bad behavior doesn't always go unpunished, so they don't need to make their own justice. Being close substitutes for reciprocation, low fines and yellow cards act as stabilizing mechanisms in a world of feeble social order.

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 $t \in \{1, ..., T\}$.

The material payoff of a player depends positively on the number of tokens she extracts and negatively on the aggregate level of extraction. Denote by $\pi(x_{it}, \bar{x}_{-it})$ the material payoff of player i, where $\bar{x}_{-it} = \frac{1}{N-1}\sum_{j \neq i} x_{jt}$. Function $\pi(x_{it}, \bar{x}_{-it})$ is increasing in x_{it} and decreasing in $\bar{x}_{-it}$. The sum of the payoffs of all players is maximized when they all extract the minimum amount.

Assume that the social norm is to extract one token, and that everybody is aware of it. At the end of each round, an external authority inspects each player with probability $p_t \in [0, 1]$. If the authority discovers some 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 to the sea). Thus, the expected material payoff of player i in round t is $\pi(x_{it}, \bar{x}_{-it}) - p_t f_t \cdot (x_{it} - 1)$.

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 models 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's [2005].)

We allow the type of a player to depend on institutions. Let $\theta_i(\omega_t) \in \{\text{S}, \text{UC}, \text{CC}\}$ be the type of player i when the institutional environment is ω_t. We shall postpone the definition of "institutional environment" until the next section. For now, bear in mind that the institutional environment may comprise such things as the enforcement of a norm by an external authority, and that it may change over time (which is why we add subindex t to ω).

Let $u(x_{it}, \bar{x}_{-it}, \theta_i[\omega_t])$ be the utility function of player i when she is of type $\theta_i(\omega_t)$. We define $u(x_{it}, \bar{x}_{-it}, \theta_i[\omega_t])$ as follows:

$$\begin{aligned}
u(x_{it}, \bar{x}_{-it}, \text{S}) &= \pi(x_{it}, \bar{x}_{-it}) - p_t f_t \cdot (x_{it} - 1) \\
u(x_{it}, \bar{x}_{-it}, \text{UC}) &= \pi(x_{it}, \bar{x}_{-it}) - p_t f_t \cdot (x_{it} - 1) + \beta_1 \pi_{\max} \frac{x_{it} - 1}{x_{\max} - 1} \\
u(x_{it}, \bar{x}_{-it}, \text{CC}) &= \pi(x_{it}, \bar{x}_{-it}) - p_t f_t \cdot (x_{it} - 1) \\
&\quad + \beta_1 \pi_{\max} \cdot \left(1 - \beta_2 \frac{\bar{x}_{-it} - 1}{x_{\max} - 1}\right) \frac{x_{it} - 1}{x_{\max} - 1},
\end{aligned}$$

where $\beta_1, \beta_2 > 0$, and $\pi_{\max}$ is the maximum material payoff a player may get in one round. This means that an unconditional cooperator that extracts $x_{\max}$ tokens experiences a 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. Let $\mathrm{q}(\theta \mid \omega)$ be the probability that a player will be of type θ given that institution is 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 logit function of her expected utilities:

$$\mathrm{P}_{it}(x) = \frac{\exp\{\lambda \cdot u(x, \varepsilon_{it}, \theta_i[\omega_t])\}}{\sum_{y=1}^{x_{\max}} \exp\{\lambda \cdot u(y, \varepsilon_{it}, \theta_i[\omega_t])\}},$$

where $\lambda > 0$ is her tendency to maximize.

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

process:

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

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 the experimental evidence.)

The steady state of $\bar{x}_t$, the mean extraction level of the group in round t, has one important property. If there are no conditional cooperators in a group, $\bar{x}_t$ 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].

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

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 tokens ($x_{\max} = 8$). The experimenter then informs players the aggregated level of extraction, but does not reveal individual levels. Player i's payoff on round t is given by

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

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

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 shall fine her privately. The experimenter may force the sanction mechanism on the players, or let them vote on it. In either case, he first explains to the players that having a fine is convenient to them because it discourages extracting more than one token, and when everybody extracts one token the material welfare of each player is maximized.

We identify four institutional environments:

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

HF A high fine is in force.

LF A low fine is in force.

RF A fine was proposed to the players in the past, but they rejected it.

We do not distinguish between fines imposed by the experimenter and fines approved by vote, because that distinction made no difference on the behavior of players. 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) institution.

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

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

The expected material payoff of player i on round t is therefore

$$\pi(x_{it}, \bar{x}_{-it}) - \frac{1}{5} f(\omega_t)(x_{it} - 1),$$

where $\frac{1}{5}$ is the probability she will be inspected.

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

Control (8 groups) Institution is NF during all twenty rounds.

High fine (14 groups) Institution is NF during the first ten rounds, and HF during the last ten rounds.

Low fine (26 groups) Institution is NF during the first ten rounds, and LF during the last ten rounds.

Rejected Fine (16 groups) Institution is NF during the first ten rounds, and RF during 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.

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 treatment. 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 that the control, cooperation increases on round 11. The social optimum, however, is never reached. Extraction falls even when the players vote against the fine.

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

Compare the results of the experiment with the predictions of Table 1. According to the predictions, initial extraction levels should be 60% higher than what they actually are. Under the high fine, extraction should drop to one. Instead, it stays over two. One expects a low fine to exert little deterrence. On the contrary, 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.

Figure 2 shows histograms of individual extraction levels under different treatments. Under both fine treatments extraction concentrates near 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 they didn't care for material payoffs. A complementary explanation for the flatness is that players are heterogenous along the moral dimension: some feel strongly that they should not take more than one token; others have no qualms and maximize their material payoffs by taking eight. Also note how the histograms that represent the rejected fine treatment get flatter on rounds 15 and 20, as cooperation deteriorates.

The unraveling process is better understood by examining, one by one, the groups that rejected a fine. Figure 3 shows four of such groups. Group 1 starts extracting a high amount, and there it stays. Groups 2, 3, and 4

start extracting 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 cooperating, but somewhere along the way they cease to cooperate abruptly (first group 2 and, a bit later, group 3). The smooth, concave line representing the rejected fine treatment in Figure 1 results from averaging up many groups like 2 and 3.

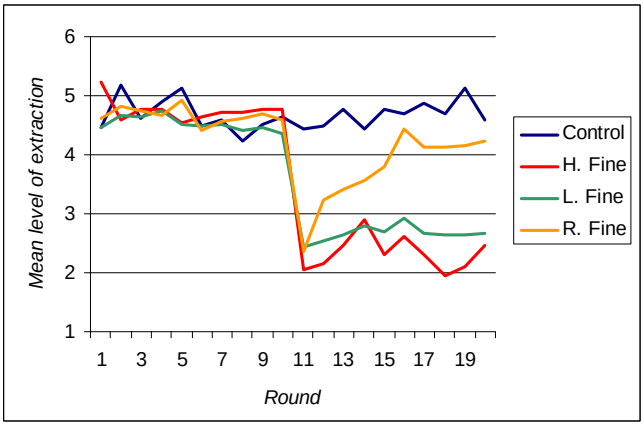

Figure 1: Experimental results, aggregate behavior.

Institution	Group mean extraction	Group deviation	Individual deviation
No fine	4.64	2.29	1.83
High Fine	2.33	1.94	1.04
Low Fine	2.66	2.06	1.15
Rejected Fine	3.74	2.29	1.79

Table 2: Summary statistics of the CPR experiment.

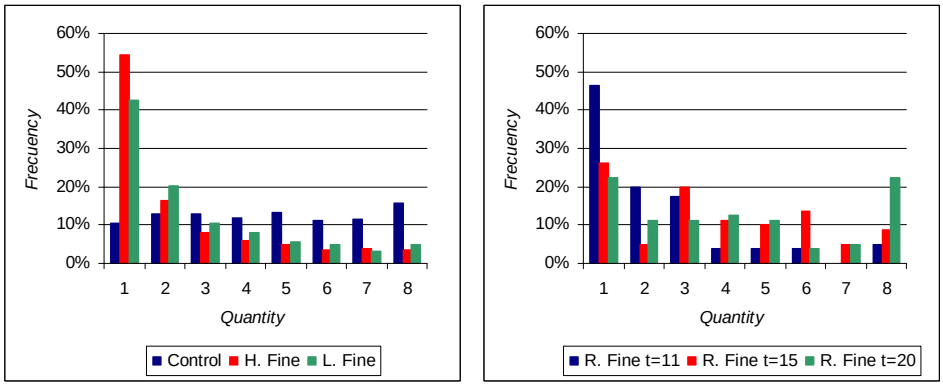

Figure 2: Experimental results, distribution of individual extraction levels.

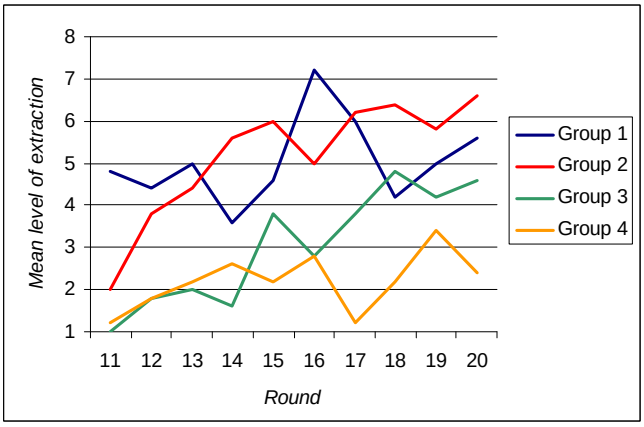

Figure 3: Experimental results, groups that voted against a fine.

4 Model estimation and simulation

We used maximum-likelihood to estimate the parameters of our model: λ, β_1, β_2, $\varepsilon(\cdot)$, ϕ, and $\mathrm{q}(\cdot \,|\, \cdot)$ (see Appendix A). 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 be of type θ when institution is ω.

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

1. If $\omega \in \{\mathrm{NF}, \mathrm{HF}, \mathrm{LF}\}$, $\varepsilon(\omega)$ coincides with a stable steady state of $\bar{x}_t$.
2. If $\omega = \mathrm{RF}$, $\varepsilon(\omega)$ is a convex combination of the stable steady states of $\bar{x}_t$.

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, $\mathrm{q}(\theta \,|\, \omega)$, under each institution. Finally, Table 5 displays the estimated initial expectations, $\varepsilon(\omega)$. 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 20% of the players are cooperative. When a fine (high or low) is in force, the percentage rises to 70%, and to 96% when the players reject a fine. The stronger effect of rejected fines is consistent with the findings of Gneezy and Rustichini (2000): people are less willing to embrace the norm when they feel they can buy their way out of it. Also, our results reveal that the enforcement of the norm induces more players to cooperate unconditionally: unconditional cooperators are 21% when a fine is rejected, and about 50% when a fine (high or low) is in force[6]. Our hypothesis is that fines relieve the cooperative player from the

[6]We bootstraped the model 51 times, taking each group history as an independent observation. In 49 of the 51 bootstrap estimations we found that $\mathrm{q}(\mathrm{S} \,|\, \mathrm{NF}) > \mathrm{q}(\mathrm{S} \,|\, \omega)$ for all $\omega \in \{\mathrm{HF}, \mathrm{LF}, \mathrm{RF}\}$. In *all* bootstrap estimations we found that $\mathrm{q}(\mathrm{S} \,|\, \mathrm{RF}) < \mathrm{q}(\mathrm{S} \,|\, \omega)$ for all $\omega \in \{\mathrm{NF}, \mathrm{HF}, \mathrm{LF}\}$, and also that $\mathrm{q}(\mathrm{UC} \,|\, \mathrm{RF}) > \mathrm{q}(\mathrm{UC} \,|\, \omega)$ for all $\omega \in \{\mathrm{HF}, \mathrm{LF}\}$.

urge to retaliate uncooperative ones in the only way she can: by ceasing to cooperate herself.

Table 5 also shows the stable steady states of $\bar{x}_t$ 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 coordinate immediately around the long run value of $\bar{x}_t$: where equilibria are unique, there is little space for confusion. On the other hand, $\bar{x}_t$ 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[7]. The unraveling of cooperation is the transit from the high cooperation equilibrium to the low cooperation one.

Our findings solve the two puzzles in the experimental 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 players reject a fine, the internalization of the social norm "extract only one token" explains the increased cooperation; violations (accidental or not), coupled with reciprocal preferences, account for the unraveling. Low fines stabilize cooperation by preventing a spiral of negative reciprocation. Because the imposition of a low fine may moralize selfish players, the "fine enough or don't fine at all" policy prescription of Lin and Yang (2005) is qualified.

Parameter	Estimate
λ	0.0027 (0.0016)
β_1	4.0903 (1.1677)
β_2	3.2285 (0.6048)
ϕ	0.4828 (0.2130)

Table 3: Estimated parameters.

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

$\mathrm{q}(\theta \mid \omega)$	NF	HF	LF	RF
S	80% (18%)	30% (9%)	29% (8%)	4% (4%)
UC	8% (8%)	48% (11%)	52% (8%)	21% (11%)
CC	12% (11%)	22% (15%)	19% (12%)	75% (11%)

Table 4: Estimated distribution of types.

Institution	Initial expectation	Stable steady state(s)
No fine	4.80 (0.14)	4.80
High fine	2.47 (0.28)	2.47
Low fine	2.60 (0.30)	2.60
Rejected fine	2.42 (0.29)	1.89; 6.37

Table 5: Estimated initial expectations.

To test the descriptive accuracy of our model, we simulated each treatment 500 times, using the estimated parameters as inputs. Figure 4 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 5 compares the actual and simulated histograms of individual extraction. The results of the experiment and the output of the simulation are extremely similar. Our model provides good account of the player's behavior, at both the group and the individual level.

Finally, we re-estimated our model subject to the restriction that $\mathrm{q}(\theta \mid \mathrm{NF}) = \mathrm{q}(\theta \mid \mathrm{HF}) = \mathrm{q}(\theta \mid \mathrm{LF}) = \mathrm{q}(\theta \mid \mathrm{RF})$, for all $\theta \in \{\mathrm{NF}, \mathrm{RF}\}$. Using a likelihood ratio test we were able to reject, with 99% confidence, the hypothesis that the distribution of types does not change across treatments[8]. We also simulated the restricted model, using estimated parameters as inputs, and it was unable

[8]The log-likelihoods of the unrestricted and restricted models are $\mathcal{L}_U = -11854.08$ and $\mathcal{L}_R = -12250.78$. The likelihood ratio statistic is $2(\mathcal{L}_U - \mathcal{L}_R) = 793.40 > \chi^2_6(.99) = 16.81$, so we reject the hypothesis.

to accurately mimic the experimental evidence (particularly the unraveling). We conclude that, in our CPR experiment, institutions influenced the social preferences of players.

	Mean extraction		Group dev.		Individual dev.	
Institution	*Actual*	*Sim.*	*Actual*	*Sim.*	*Actual*	*Sim.*
No fine	4.64	4.73	2.29	2.45	1.83	1.95
High Fine	2.33	2.59	1.94	2.14	1.04	1.32
Low Fine	2.66	2.66	2.06	2.22	1.15	1.30
Rejected Fine	3.74	3.43	2.29	2.66	1.79	1.56

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

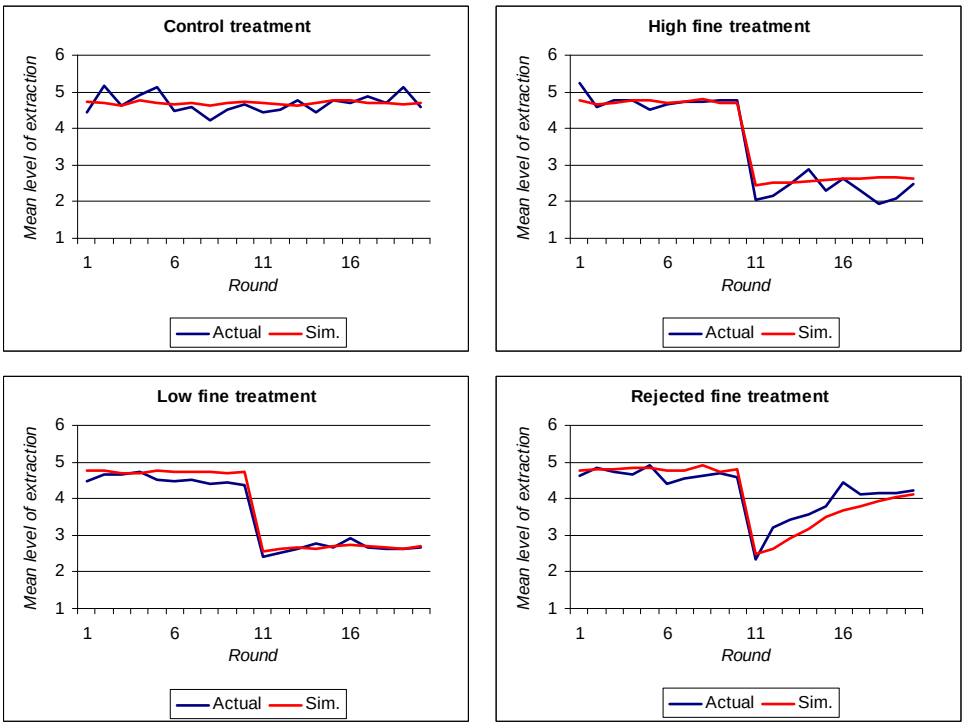

Figure 4: Mean levels of extraction, actual and simulated.

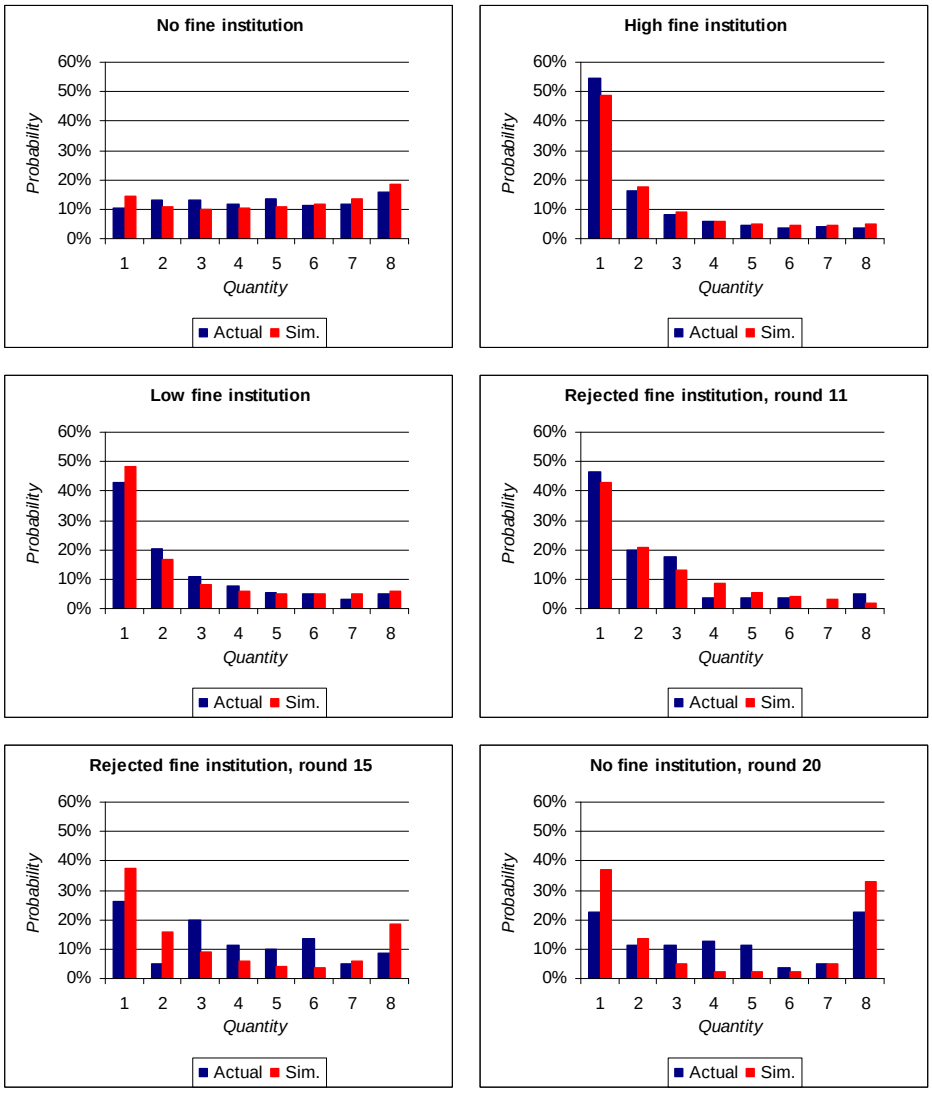

Figure 5: Distribution of individual levels of extraction, actual and simulated.

5 Concluding Remarks

Authority may influence social preferences by prescribing and enforcing norms. We found, in a CPR experiment, that the external imposition of a norm affected preferences in two ways.

First, by moralizing players. A mere speech by the experimenter sufficed to induce players to cooperate. How? By sowing in them the seed of guilt. Aristotle (1962) argued in his Nichomachean Ethics that effective laws worked by inculcating habits in citizens, that is, by moralizing them[9]. Our results remind us that his argument is still relevant in present times.

Second, our model revealed 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 also bring attention to the dynamic effects of enforcement. Conditional cooperation makes compliance fragile: a single rotten apple may spoil the whole box (and many good apples cannot turn a spoiled box good). In the experiment, a small fine sufficed to stabilize cooperation because it made more players cooperate unconditionally, preventing the spread of moral degradation. Consider the implications on governmental corruption. Corrupt officers are hard to detect, so the "expected punishment" will often be small compared with the potential loot. The occasional jailing of corrupt officers may nonetheless stabilize moral behavior. As in the metro of Santiago example, weak enforcement may prevent people from falling in the "Everybody else is doing it, so why can't we" trap.

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

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

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.

Gneezy, Uri and Aldo Rustichini. 2000. Pay enough or don't pay at all. *Quarterly Journal of Economics*, 115(3): 791-810.

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.*

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.

A The log-likelihood function

Here we build, step by step, the log-likelihood function.

- $\boldsymbol{\pi} = (\lambda, \beta_1, \beta_2, \mathrm{q}[\cdot \mid \cdot], \varepsilon[\cdot], \phi)$ is the vector of parameters we want to estimate.

- τ^g is the treatment applied to group g, and ω_t^g is the institution that group g faces on round t. Thus,

$$\omega_t^g = \begin{cases} \text{NF} & \text{if } t \leq 10 \vee (t \geq 11 \wedge \tau^g = \text{Control}) \\ \text{HF} & \text{if } t \geq 11 \wedge \tau^g = \text{High fine} \\ \text{LF} & \text{if } t \geq 11 \wedge \tau^g = \text{Low fine} \\ \text{RF} & \text{if } t \geq 11 \wedge \tau^g = \text{Rejected fine.} \end{cases}$$

- ε_{it}^g is the number of tokens that player i of group g expects others to extract on round t. Define ε_{it}^g as follows:

$$\varepsilon_{it}^g = \begin{cases} \varepsilon\left(\omega_t^g\right) & \text{if } t = 1 \vee (\tau \neq \text{Control} \wedge t = 11) \\ \phi \varepsilon_{i(t-1)}^g + (1-\phi)\, \bar{x}_{-i(t-1)}^g & \text{otherwise,} \end{cases}$$

  We make two assumptions regarding $\varepsilon(\omega)$, the initial expectation of players under institution ω:

  1. If $\omega \in \{\text{NF}, \text{HF}, \text{LF}\}$, we assume $\varepsilon\left(\omega\right)$ coincides with a stable steady state of $\bar{x}_t$. That is, this condition must hold:

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

     and the derivative of the right hand side with respect to $\varepsilon(\omega)$ must be negative.

  2. If $\omega = \text{RF}$, we assume $\varepsilon\left(\omega\right)$ is a convex combination of the stable steady states of $\bar{x}_t$.

- $\mathrm{P}_{it}^g\left(\theta\right)$ is the probability that player i of group g chooses x_{it}^g on round t, given that her type is θ:

$$\mathrm{P}_{it}^g\left(\theta\right) = \frac{\exp\left\{\lambda \cdot u\left(x, \varepsilon_{it}^g, \theta\right)\right\}}{\sum_{y=1}^{x_{\max}} \exp\left\{\lambda \cdot u\left(y, \varepsilon_{it}^g, \theta\right)\right\}}.$$

- $\boldsymbol{\theta} = (\theta_1, ..., \theta_5)$ is a combination of players' types. The probability of observing combination $\boldsymbol{\theta}$ when institution is ω is:

$$\mathrm{q}\left(\boldsymbol{\theta} \,|\, \omega\right) = \prod_{i=1}^{5} \mathrm{q}\left(\theta_i \,|\, \omega\right).$$

- The history of group g is the matrix of its members' individual extractions over all rounds:

$$\begin{bmatrix} x^g_{1,1} & \dots & x^g_{1,20} \\ \vdots & \ddots & \vdots \\ x^g_{5,1} & \dots & x^g_{5,20} \end{bmatrix}.$$

- $\mathrm{h}^g\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)$ is the probability of observing the history of group g, given that the combination of types during the first ten rounds is $\boldsymbol{\theta}$ and during the last ten rounds is $\boldsymbol{\theta}'$.

$$\mathrm{h}^g\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \prod_{i=1}^{5} \left\{ \prod_{t=1}^{10} \mathrm{P}^g_{it}\left(\theta_i\right) \prod_{t=11}^{20} \mathrm{P}^g_{it}\left(\theta'_i\right) \right\}.$$

  Note that probabilities $\mathrm{P}^g_{it}\left(\cdot\right)$ are linked through expectations ε^g_{it}.

- H^g is the unconditional probability of observing the history of group g:

$$\mathrm{H}^g = \begin{cases} \sum\limits_{\boldsymbol{\theta}} \mathrm{h}^g\left(\boldsymbol{\theta}, \boldsymbol{\theta}\right) \mathrm{q}\left(\boldsymbol{\theta} \,|\, \mathrm{NF}\right) & \text{if } \tau_g = \text{Control} \\ \sum\limits_{\boldsymbol{\theta}, \boldsymbol{\theta}'} \mathrm{h}^g\left(\boldsymbol{\theta}, \boldsymbol{\theta}' \,|\, \boldsymbol{\pi}\right) \mathrm{q}\left(\boldsymbol{\theta} \,|\, \mathrm{NF}\right) \mathrm{q}\left(\boldsymbol{\theta}' \,|\, \omega^g_{11}\right) & \text{otherwise.} \end{cases}$$

  Probability H^g depends on parameter vector $\boldsymbol{\pi}$, so we write it $\mathrm{H}^g\left(\boldsymbol{\pi}\right)$

- The log-likelihood function is:

$$\mathcal{L}\left(\boldsymbol{\pi}\right) = \sum_{g=1}^{64} \log \mathrm{Q}^g\left(\boldsymbol{\pi}\right).$$