Document de travail du LEM / Discussion paper LEM 2018- 03

The Stickiness of Norms

Katherine FARROW
EconomiX, University of Paris Nanterre / katherine.farrow@parisnanterre.fr

Rustam ROMANIUC
LEM UMR 9221 / rustam.romaniuc@univ-catholille.com

http://lem.cnrs.fr/IMG/pdf/dp2018-03.pdf

http://lem.cnrs.fr/

The Stickiness of Norms

Katherine Farrow* and Rustam Romaniuc†

Abstract

In this paper we study the role of social context, as characterized by different internal norm-enforcement mechanisms, on the legacy of temporary external regulations. In a public good game, we create conditions in which a prosocial norm of cooperation is enforced via either anonymous peer punishment or face-saving concerns. In two test treatments, we introduce to these each of these social environments an external regulation that is implemented for a limited period of time and then removed. Results indicate that both peer disapproval and face-saving concerns are effective mechanisms for increasing cooperation and that these effects persist over time. Whereas we observe a significant negative post-intervention effect in the context of peer disapproval, no such effect exists in the context of face-saving concerns. Our findings reveal the importance of the type of norm-enforcement mechanism in determining the robustness of norm adherence in the long term.

*Montpellier Laboratory for Theoretical and Applied Economics, University of Montpellier; Present address: EconomiX, University of Paris Nanterre, 200 Ave de la République, 92000 Nanterre, France *Email address:* katherine.farrow@parisnanterre.fr

†Catholic University of Lille – Laboratory for Experimental Anthropology & LEM-CNRS, 60 bd Vauban, 59000 Lille, France. *Email address:* rustam.romaniuc@univ-catholille.com

1 Introduction

Law is traditionally defined as a set of formal rules, promulgated by legislatures, regulatory agencies, and courts, and backed by the threat of monetary punishment or imprisonment (Posner and Rasmusen 1999). However, rules of conduct can also be informal insofar as they do not dependent on government for either promulgation or enforcement. When norm-enforcement consists in the refusal to interact with the offender or in the expression of disapproval of one's actions, for example, behavior is considered to be constrained by informal norms.

During the 1990s, the relationship between formal rules and informal norms became topical among legal scholars, economists, and within the field of law and economics more broadly. As Ellickson notes, "in the mid-1990s norms became one of the hottest topics in the legal academy" (1998, p. 543). The quantity and the quality of published papers on this topic rose significantly, as evidenced by the development of the area of research referred to as the law-and-economics of norms (Feldman 2009) and by the attention allocated to this subject in prominent law journals. In the second half of the 1990s, there have been at least eight major symposium issues on the subject of laws or formal rules and norms.[1]

On a more fundamental level, one primary question dominated discourse in the area during this time: how do informal norms perform compared to formal rules in inducing socially desired behaviors? More specifically, the debate involved a discussion about enforcement mechanisms. Some argued that the informal enforcement of norms by peers can serve to establish and maintain cooperative individual interactions, or

[1]Symposium, Law, Economics, and Norms, 144 University of Pennsylvania Law Review (1996); Symposium, Law and Society & Law and Economics, 375 Wisconsin Law Review (1997); Symposium, The Nature and Sources, Formal and Informal, of Law, 82 Cornell Law Review (1997); Symposium, Social Norms, Social Meaning, and the Economic Analysis of Law, 27 Journal of Legal Studies (1998); Symposium, Corporate Law and Social Norms, 99 Columbia Law Review (1999); Symposium, The Legal Construction of Norms, 86 Virginia Law Review (2000); Symposium, Norms, Law, and Order in the City, 34 Law and Society Review (2000); Symposium, New and Critical Approaches to Law and Economics: Part II, Norms Theory, 79 Oregon Law Review (2000).

social order (Ellickson 1991). Others claimed that informal mechanisms are not effective and that formal rules enforced by external authorities are requisite elements of stable social order. The former contingent pointed to a large array of historical examples as evidence of the feasibility of order without "the backing of state authority" (Benson 1991). Informal norms appeared to successfully maintain social order in primitive (Benson 1991) and medieval societies (Friedman 1979) and continue to do so in contemporary societies (Ellickson 1991; Bernstein 1992). These types of informal, internal enforcement mechanism are understood to rely on shame – a disutility that occurs when others identify an individual as offending an established norm of conduct (Elster 1998; Bowles and Gintis 2006; Masclet *et al.* 2003; Guala 2012).

The advocates of formal rules, on the other hand, have claimed that, because the norms that support social order can be considered a public good, self-interested individuals would neither contribute to their creation nor follow or enforce them.[2] This argument implies that peer control cannot be relied on as an adequate mechanism to enforce cooperation. According to this perspective, individual behaviors are therefore best controlled via explicit, centrally designed formal rules (see, e.g, Sened 1997; Aviram 2004).

The focus on the comparative advantages of one system versus the other, surprisingly, neglects the fact that there exists a temporal consideration in the external enforcement of formal rules. Namely, formal rules cannot be effective without the support of corresponding informal social norms (Boettke *et al.* 2008). Indeed, since informal norms constitute an inherent element of the fabric of society (Elster 1989; Bicchieri 2006), they necessarily precede formal rules and therefore serve an important legitimizing function. This implies that formal rules should take into account existing informal peer-based mechanisms of enforcement that are already in place. This observation has been made by spontaneous order theorists, who urge for caution when

[2]Norms are even more of a public good than formal rules since no political party, public agency, or lobby group can claim credit for creating a norm.

establishing new rules designed and enforced by public authorities (Boettke *et al.* 2008; Williamson 2009). This consideration also relates to what Boettke *et al.* (2008) has referred to as the stickiness of informal norms: the effectiveness of formal rules is improved if one takes into account the temporal sequence that understands informal norms as precedents for formal rules. This observation has also been advanced by the law-and-economics of norms literature (Feldman 2009), which argues that formal rules act as focal points in a system of informal norms typically characterized by multiple equilibria (Sunstein 1996; Cooter 1998). Formal rules can therefore be considered to harness the power of informal norms via their expressive power, increasing the salience of socially acceptable behavior and serving as coordination mechanisms that identify which norms should be observed. If a law prohibiting littering is enacted, for example, individuals can expect not only to pay for non-compliance, but also to be the target of ostracism from other members of the community to a greater extent than if no such law existed. In the absence of any legal code regarding littering behavior, the expectation of peer ostracism may be reduced, suggesting that formal rules also serve to legitimize informal norms and thus reinforce peer pressure.

This paper uses a laboratory public good experiment to investigate how adding and subsequently removing an externally enforced formal rule – in the form of a monetary sanction – affects the functioning of an informal norm supported by two different norm-enforcement mechanisms.[3] In the first treatment, we give individuals the opportunity to send anonymous, non-costly disapproval points to each other based on the contributions made in the previous round. The inclusion of this treatment was motivated by a robust finding in the law and economics literature demonstrating that informal norms are supported by low-cost expressions of social disapproval, such as

[3]These questions should ideally be studied in a field setting. Manipulating and measuring the effect of different shaming strategies in the field, however, is problematic. It is for this reason that the crowding-out of social disapproval in Gneezy and Rustichini's (2000) seminal paper, for instance, is proposed as only one of a number of possible explanations for their results. Shaming in real-life settings may, for example, be influenced by factors such as the belief that those receiving expressions of disapproval could also retaliate (Nikiforakis 2008).

ridicule and gossip, rather than costly punishment (Ellickson 1991; Boehm 1999; Feinberg *et al.* 2012; Guala 2012). Ostrom *et al.* (1992) made the first attempts to design a laboratory experiment to study norm enforcement by peers. In the context of a common-pool-resource game, they show that people use "shaming" as a strategy to try to induce others to comply with what they consider to be appropriate conduct. Notably, this shaming strategy led to substantial improvement in cooperation levels. An experiment that allowed subjects to directly communicate their disapproval is Masclet *et al.* (2003). The authors found that simply providing subjects with this opportunity increases compliance with cooperation norms. They explain this result by the fact that social disapproval signals what is socially acceptable behavior and instills shame for deviating from the norm.[4]

In the second treatment, we implement another non-monetary mechanism that has been shown to invoke remorse in the deviant person. Following every round in this treatment, we display the pictures of all group members next to their individual contributions. Ho (1976) defines the concept of "face" as one's positive social value or respectability, the loss of which makes it more difficult to function in society, implying added costs of some sort. There exists a good deal of evidence that the loss of face is an important motivator of individual action (Bohnet and Frey 1999; Rege and Telle 2004; Coricelli *et al.* 2010; Coricelli *et al.* 2014). While the first mechanism we implement relies on the explicit expression of disapproval by one's peers, the latter rests on one's belief about how he/she is perceived by the others around him (Andreoni and Petrie, 2004; Bursztyn and Jensen 2017). Hu (1944) perceptively shows how the meaning of "face" as "the respect of the group for a man" (p. 45) is different from social disapproval. The author argues that individuals experience more intense negative feelings when confronted with the explicit disapproval of others relative to those felt when this disapproval is absent. Without witnesses, the negative feelings

[4]Subhasish (2013) and Nelissen and Mulder (2013) have confirmed this seminal result from Masclet, Noussair, Tucker, and Villeval.

associated with breaking a social convention could be considered to be feelings of guilt, whereas the presence of others arguably introduces the added, but distinct, emotion of shame. Our experimental design allows us to compare these two mechanisms in the context of a public good game. In the remainder of the paper, we refer to peer disapproval and face-saving as *internal enforcement mechanisms* and the formal sanction as an *external enforcement mechanism*.

In addition to being the first paper to investigate the advantages of peer disapproval compared to face-saving concerns in increasing compliance with a norm of cooperation, the novelty of our experimental protocol consists of measuring the resilience of these two internal mechanisms to the introduction and removal of an external enforcement mechanism, namely a mild monetary punishment. Many experimental papers in recent years have demonstrated that externally-enforced sanctions can reduce the effectiveness of informal norms (e.g. Gneezy and Rustichini 2000). Our laboratory experiment is the first to compare the long-lasting effects of monetary sanctions on norms that are enforced by two internal mechanisms: social disapproval vs. concern for one's social image.

Our results confirm, first, that both types of internal mechanisms increase cooperation and that this impact is persistent over time, especially with respect to peer-disapproval. Second, we find a striking difference in the effectiveness of these internal enforcement mechanisms once the external mechanism has been removed. Under conditions of peer disapproval, we observe a strong negative post-intervention crowding-out: cooperation falls to levels below those observed under baseline conditions. When face-saving concerns are salient, however, we observe no such effect: in fact, cooperation in the post-intervention periods remains above baseline levels. Thus, our data suggest that while both types of internal mechanisms appear to be complements with respect to an external mechanism, making social image concerns salient is ultimately a more suitable internal norm-enforcement mechanism for use in

conjunction with external mechanisms insofar as it appears to be robust to the negative post-intervention effect that is observed in the context of peer disapproval. In other words, our findings suggest that an informal norm enforced via a milder shaming strategy (i.e. face-saving concerns) is stickier than a norm enforced via a strategy involving explicit punishment (i.e. peer disapproval).

2 Experimental design

2.1 The experimental game

We study cooperation in the context of what has become the benchmark for experimental research on social dilemmas, the public good game. Subjects in our game are assigned to groups of four and endowed with $E_i = 20$ tokens and must choose how to allocate this amount between a public account (g_i) and a private account (c_i). Each token left in the private account generates a benefit equal to 1 Experimental Currency Unit (Ecu). In addition to the Ecus kept on the private account, each participant receives a fixed benefit $\alpha = 0.4$ Ecus from the total group contribution to the public account, $\sum_{j=1}^{4} g_j$. Parameters are set such that $0 < \alpha < 1 < n\alpha$. From $1 < n\alpha$, it follows that the utilitarian optimum and the efficient symmetric outcome is for all group members to contribute their entire endowments to the public account. However, under this specification, it remains in each individual's self-interest to contribute zero to the public account. Since the game is symmetric, the Nash equilibrium is therefore g_j. The payoff function under baseline conditions is given by:

$$\pi_i = 20 - g_i + 0.4 \sum_{j=1}^{4} g_j$$

We begin each treatment with ten periods of play under these conditions. This familiarizes subjects with the game and creates a challenging environment for

cooperation, as subjects become accustomed to levels of free-riding that typically characterize play in the public good game by the end of the first ten periods. The subjects were informed that a second and a third sequence of the game will follow but were only given the instructions corresponding to the first sequence of 10 rounds.

Our experimental manipulations consist of two variations to the standard public good game, which are designed to mimic an external enforcement mechanism – based on monetary punishment meted out by the experimenter – and two different internal enforcement mechanisms: one based on anonymous peer disapproval and the other based on social image. In the *Peer Disapproval* condition, participants are informed after the first sequence of the game (periods 1-10) that for the second sequence (periods 11-20), after every round, they will now be informed about the individual contribution of the other group members and they will have the opportunity to send points of disapproval to the other group members, from 0 to 10 points, where sending 0 indicates no disapproval and sending 10 indicates strong disapproval of another group member's contribution in that round.

In the *Saving Face* condition, at the end of the first sequence of the game, participants are informed that for the second sequence of the game their photograph will now appear next to their contribution amounts, which are made available to the rest of the members of their group after each round of play.[5] In the *Saving Face* and *Peer Disapproval* conditions, we inform subjects at the end of the second sequence of the experiment that the third sequence is exactly the same as the second one.

The *Saving Face* and *Peer Disapproval* conditions are implemented in two different environments that we will refer to as the *No Sanction* treatment and the *Sanction* treatment. In the *No Sanction* treatment, the second and third sequences are identical, i.e. we either implement *Saving Face* or *Peer Disapproval* in the second and third sequence of the game. In the *Sanction* treatment, we simultaneously introduce an

[5]It is worth noting that under the two conditions, *Peer Disapproval* and *Saving Face*, individual contributions are displayed.

external enforcement in the form of a monetary punishment in period 11 along with either peer disapproval or saving face. These elements are used in tandem through period 20. In periods 21-30 (third sequence of the game) we remove the external mechanism, leaving only the internal mechanism (peer disapproval or saving face) in effect for the remainder of the experiment. Thus, the only difference between the *Sanction* and *No Sanction* treatments is the presence or absence of an external enforcement – in the form of monetary punishment – in the second sequence of the game.

The monetary sanction itself is implemented by informing subjects that 0.3 Ecus will be subtracted from every Ecu not allocated to the public account and which therefore remains on subject's private account. The intensity and framing of the sanction were chosen so as to replicate two specific characteristics of institutional punishments that are currently utilized in many real-world policies. These types of punishments are typically mild (Engel 2014), and their punitive intent is clear. In order to implement a mildly costly punishment, we set the subtraction rule so as to ensure that donating zero remains the dominant strategy for money-maximizing individuals, which preserves the nature of the decision as a social dilemma, i.e. one that pits an individual's interest against the interest of the group. The payoff function under the sanction conditions is given by:

$$\pi_i = 20 - g_i + 0.4 \sum_{j=1}^{4} g_j - 0.3(20 - g_i)$$

where the last term represents the penalty proportional to the amount of tokens placed in the individual account. In the Sanction treatment, the return from each token left on the private account is reduced from 1 Ecu to 0.7 Ecus. Full contribution from every subject under this treatment yields $\pi_i = 32$ Ecus, and contributing zero and paying $s_i = 0.3$ for every token kept on the private account yields $\pi_i = 38$ Ecus for the free-rider. Thus, a money-maximizing individual does not contribute to the public account so long

as the sanction amount is less than the marginal per capita rate of return.

To emphasize the punitive nature of this incentive as a sanction, we frame the subtraction rule in order to make explicit the fact that Ecus are subtracted when individuals deviate from the desirable action that benefits the group. Specifically, the instructions read that 0.3 Ecus are subtracted from each Ecu that is not allocated to the public account (see the instructions in the Appendix 1). Generally, in public good experiments, it is assumed that members of the group share the understanding that the desirable action of each individual is one that favors the interest of the group, and that deviations from this action are undesirable (e.g. Andreoni and Gee 2012). Our treatment makes salient this contribution norm by emphasizing the wrongdoing. However, we avoid using words such as tax, punishment, or sanction in order to minimize experimenter demand effects (Zizzo 2010) and avoid the possibly varied connotations that participants may attach to this vocabulary.

To mimic centralized government enforcement, we make it clear to participants that the subtraction rule is applied by the central computer. The legitimacy of the enforcement figure has been shown to play an important role in public good experiments with punishment (Baldassarri and Grossman 2011). Thus, while in some experiments the punishment is meted out by a randomly chosen participant (e.g. Engel 2014), we elect to deliver punishment in the *Sanction* treatment through the central computer as the experimenter is most likely to be seen as a legitimate authority (Milgram 1963; Karakostas and Zizzo 2015).

2.2 Experimental procedures

The experiment consists of ten sessions of which four were conducted at the Laboratory for Experimental Economics in Montpellier (LEEM) and six were conducted at the Laboratory for Experimental Anthropology (Anthropo-Lab) at the Catholic University of Lille. The sessions were conducted by the same experimenter between March 2015

and March 2017.[6] A total of 196 subjects participated in our experiment. None of them had previously participated in a public good experiment. Subjects interacted through individual computer terminals using the software developed by the engineers at the LEEM. The exchange rate was 20 Ecus = 1 euro. Subjects earned an average of 20 euros, and payments were made privately at the end of the session. Sessions lasted for two hours, including the taking of the photos that would be used in the experiment, the instructions, and payments.

In the *Saving Face* condition, subjects were asked permission for their picture to be taken. They were informed that they could opt not to have their photograph taken, in which case they would be remunerated the show-up fee and allowed to leave. None of the participants refused the photograph. In order to preserve social distance between the experimenter and the subjects, subjects were informed that the assistant who took their picture was not involved in the subsequent experiment. Photographs were taken in a consistent manner for all subjects, who were instructed to maintain a neutral face.[7] Participants were then shown to the laboratory where the game was explained and two example scenarios reviewed.

At the outset of each session, subjects were informed that the central server would allocate them randomly to groups of four people. Each session consists of 30 periods, divided into three sequences of 10 periods. The total number of sequences in the session is common knowledge, as is the fact that at the end of the experiment only one sequence out of the three is chosen at random to determine the payment amount.

Table 1 provides detailed information about the described treatments, the number of sessions, subjects and groups for each treatment, and the segment in which each treatment was implemented.

[6] It is worth noting that the two laboratories follow the same rules to recruit subjects and to run the experiments. A between subjects comparison shows that there is no significant difference in average contributions in the first sequence of the game (which is identical across all our treatments) between groups in Montpellier and groups in Lille.

[7] We followed the procedure from Tognetti et al. (2013).

Table 1. Experimental treatments

Sessions	**Groups**	**Sequence 1** Periods 1-10	**Sequence 2** Periods 11-20	**Sequence 3** Periods 21-30
Sessions 1-2	9	Baseline	Peer Disapproval	Peer Disapproval
Sessions 3-4	10	Baseline	Peer Disapproval + Sanction	Peer Disapproval
Sessions 5-7	15	Baseline	Saving Face	Saving Face
Sessions 8-10	15	Baseline	Saving Face + Sanction	Saving Face

3 Hypotheses and Results

Our between subjects design allows us to study (i) the short-term effects on cooperative decisions from the two internal mechanisms (peer disapproval vs saving face) by comparing group level contributions in Sessions 1–2 to contributions in Sessions 5–7 in Sequence 2 of the game, (ii) the long-term effects of the two internal mechanisms by comparing group level contributions in the same sessions in Sequence 3, (iii) the resilience of these two internal mechanisms to the introduction and removal of an external enforcement mechanism by comparing group level contributions in Sessions 3–4 to contributions in Sessions 8–10 in Sequence 3.

Our hypotheses result from the mounting evidence, mentioned in the introduction, showing the positive impact of monetary and nonmonetary punishment on cooperation in social dilemmas (Fehr and Gachter 2000; Masclet *et al.* 2003; Coricelli *et al.* 2014). Hypotheses 1 and 2 formulate our expectations about cooperation levels when we implement the two internal mechanisms in Sequence 2 and 3.

Hypothesis 1 *The two internal enforcement mechanisms – peer disapproval and saving face – increase average group contributions compared to the Baseline.*

Hypothesis 2 *The feeling of shame in the absence of social disapproval is milder. Thus, in the long-run, we expect higher levels of contributions in Peer Disapproval than in the Saving-Face treatment.*

Our next hypothesis concerns the resilience of the two internal mechanisms to the removal of the external enforcement mechanism. More specifically, we hypothesize that the act of removing the sanction differentially impacts the mechanisms that support norm adherence in each social environment.

Hypothesis 3 *The introduction of an externally enforced sanction for free-riding legitimizes social disapproval and its removal signals that those who express disapproval are no longer backed by the authority that implemented the sanction. This reduces the effectiveness of peer disapproval but not of saving-face. One's image concern is not affected by the removal of external enforcement.*

Average contributions per treatment are shown in Table 2, and the evolution of contributions across the thirty periods of play are depicted in Figures 1 and 2. Contribution behavior in the pooled baseline treatments follows the typical pattern, with the average contribution starting at 7.82 tokens, or about 40% of the endowment, in period 1 and declining to 3.70 tokens, or around 19% of the original endowment, by period 10.

Table 2. Average contributions (*s.d.*) by treatment

Internal mechanism	**Treatment**	**Periods 1-10**	**Periods 11-20**	**Periods 21-30**
Disapproval	No sanction	Baseline 5.98 (1.25)	Disapproval 8.60 (1.14)	Disapproval 9.31 (1.54)
Disapproval	Sanction	Baseline 4.07 (1.37)	Disapproval + sanction 12.45 (3.06)	Disapproval 3.69 (2.47)
Saving face	No sanction	Baseline 5.45 (1.30)	Saving face 7.75 (1.13)	Saving face 7.40 (1.00)
Saving face	Sanction	Baseline 6.21 (1.32)	Saving face + sanction 13.79 (0.76)	Saving face 9.26 (1.47)

A series of multiplicity-adjusted Mann-Whitney tests fails to reject the null hypothesis that the mean contribution levels in the baseline periods across treatments are drawn from the same distribution. We furthermore note that contributions in Sequence 1 in all treatments follow the same pattern over time, and arrive at virtually identical average contribution levels in period 10. In what follows, we address each of the hypotheses presented above.

3.1 How do peer disapproval and saving face compare?

First, we are interested in the relative performance of each type of internal norm enforcement mechanism. Within-subject Wilcoxon signed rank tests indicate that peer disapproval significantly increases average contributions in periods 11-20 by 2.62 tokens relative to baseline levels in periods 1-10 ($z = 4.626$, $p < 0.001$). The saving-face mechanism also significantly raises average contribution levels in periods 11-20, by 2.30 tokens, relative to baseline conditions ($z = 4.536$, $p < 0.001$). Over time, this effect does not diminish over time, as both mechanisms succeed in maintaining these higher levels of cooperation in Sequence 3 relative to Sequence 1 ($z = 4.626$ and $p < 0.001$, and $z = 4.626$ and $p < 0.001$ for the disapproval and saving-face, respectively).[8]

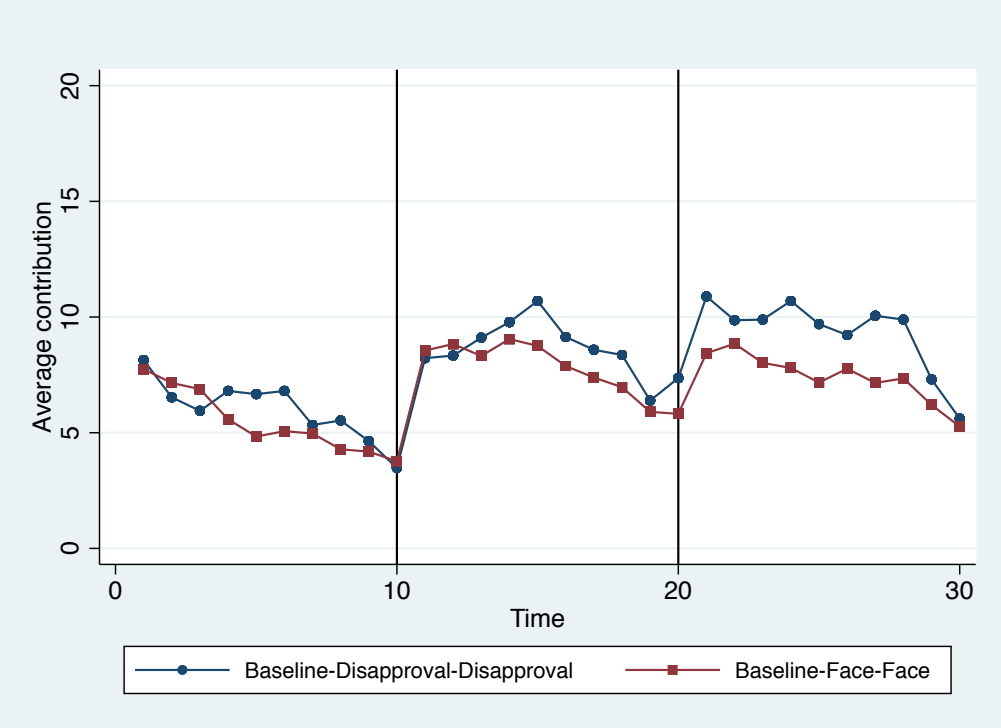
Result 1. *Both the peer disapproval and saving-face mechanisms have a positive effect on contributions relative to baseline levels, and this effect is persistent over time.*

Additionally, a between-subjects Mann-Whitney test indicates that contribution levels under peer disapproval are higher than those under the face mechanism in Sequence 3 ($z = 2.647$, $p < 0.001$) but not in Sequence 2.

Result 2. *In the long term, the peer disapproval mechanism yields higher contributions than the saving-face mechanism.*

[8]This set of 4 within-subject tests are evaluated using a Bonferroni-adjusted p-value of 0.0125.

Figure 1. Mean contributions under internal norm-enforcement mechanisms



3.2 Are both types of internal mechanisms subject to post-intervention crowding out?

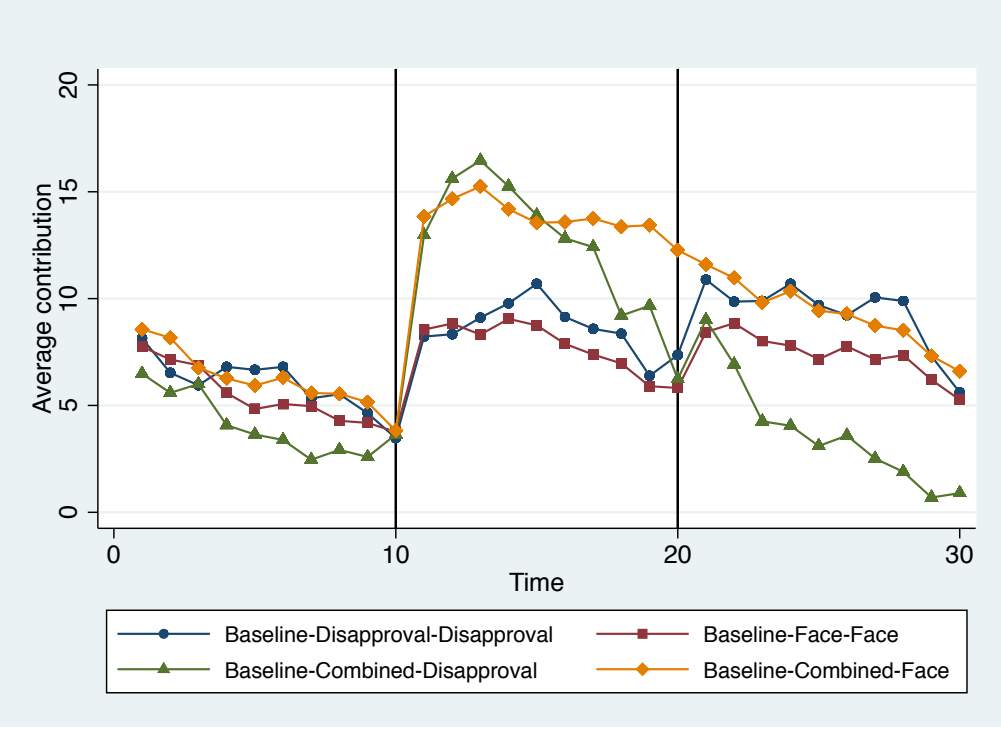
To investigate the presence of post-intervention crowding out effects, we compare contribution levels between subjects in the *Sanction* treatments with those in the *No Sanction* treatments over the final 10 periods of play (Sequence 3). By the time subjects have reached this point in the game, those in the *Sanction* treatment will have experienced an external enforcement mechanism that has been removed, while those in the *No Sanction* treatment will not have been exposed to such a sanction.

Figure 2 shows that when the sanction is removed we observe a significant decline in contribution levels in the context of peer disapproval. A Mann-Whitney test rejects the null hypothesis that contribution levels across treatments are drawn from the same underlying distribution in the post-intervention period ($z = 3.554$, $p < 0.001$), suggesting that anonymous peer disapproval is indeed vulnerable to a strong

negative post-intervention effect resulting from the removal of an external enforcement mechanism. In contrast, we observe no such negative spillover in the context of the saving-face mechanism. In fact, a Mann-Whitney test indicates that this mechanism manages to maintain an even higher level of cooperation after the sanction is removed relative to the scenario in which subjects have not been exposed to an externally enforced sanction ($z = 2.57$, $p = 0.010$).

This suggests that, whereas peer disapproval appears to be vulnerable to negative behavioral spillover resulting from an external sanction, entailing a drop in average contributions of 8.76 tokens, the saving-face mechanism is able to attenuate this effect entirely. In fact, when face-saving concerns are salient, we observe a *positive* post-intervention effect of a temporary sanction by which average contributions are 1.86 tokens higher in the long term than they are in the *No Sanction* treatment.

Figure 2. Mean contributions under internal and external norm-enforcement mechanisms

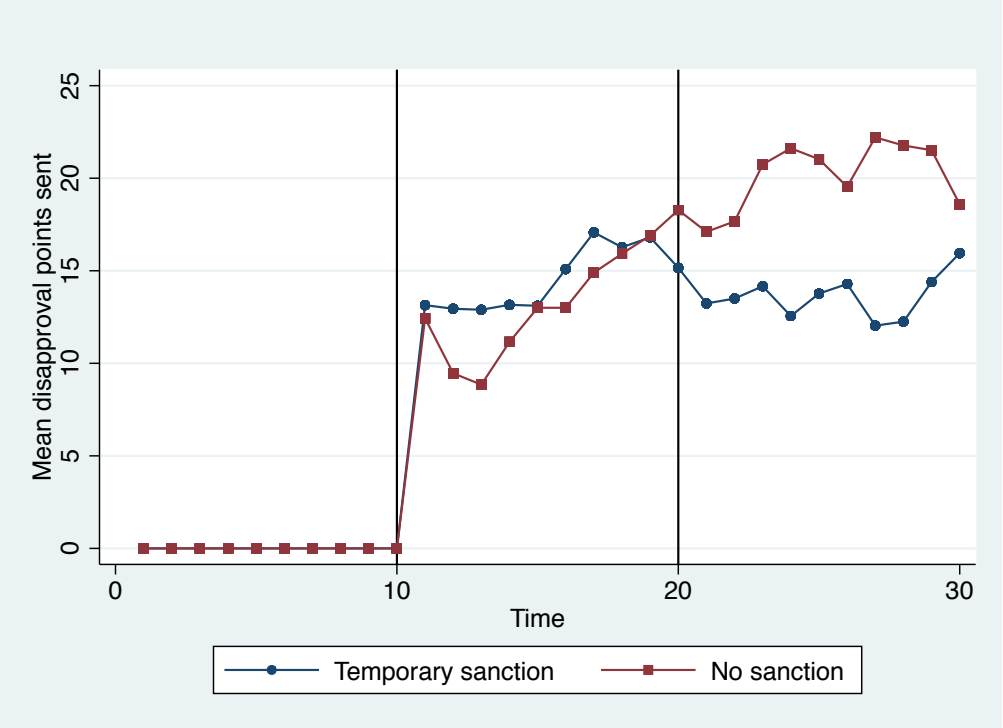


Result 3. *The removal of the externally-enforced sanction generates a negative behavioral spillover in the post-intervention period under peer disapproval. In contrast, the removal of an external sanction generates a positive behavioral spillover under saving face.*

As stated in Hypothesis 3, we suspect that the expressive function of removing the sanction has implications for the legitimacy of those who punish in the post-intervention period. An analysis of the number of disapproval points sent in Treatments with and without the sanction sheds light on the source of the negative post-intervention effect and provides support for our Hypothesis 3. Figure 3 depicts the average number of disapproval points sent within groups across the two treatments. In the *No Sanction* treatment, we observe a relatively constant average level of disapproval points sent from Sequence 2 to Sequence 3. In the *Sanction* treatment, however, the amount of disapproval points sent in the post-intervention period is significantly higher than those sent in the final ten periods of the *No Sanction* treatment (Mann-Whitney U test: z = -3.78, $p < 0.0002$).[9] Thus, the evidence we find here provides support for the mechanism we hypothesize behind Result 3.

[9]This result is developed in Romaniuc *et al.* (2016) who also examine disapproval points sent in these treatments using multivariate analysis. An OLS regression reveals that, among those who contribute less than average, the lagged number of disapproval points received is a significant predictor of contribution behavior when no sanction has been implemented, but that in the post-intervention period following the removal of a sanction, this parameter is no longer significant. This is evidence that removing a formal sanction can have the effect of desensitizing people to receiving peer punishment.

Figure 3. Mean number of disapproval points sent per group



3.3 Analysis of individual contributions

To estimate the relative importance of a variety of factors in determining contribution amounts in each period, we conduct multivariate analyses, which lend further support to our main results. Following Ashley *et al.* (2010), we estimate a random-effects dynamic panel regression censored between 0 and 20, the lower and upper bounds of the contribution range. The period number is included to account for any time trend, and the contribution that subjects made in the very first period of baseline play is included as a proxy for player type (i.e. degree of prosocial orientation). A lagged contribution amount variable is also included to take into account any autocorrelation in contributions between periods. The degree to which individuals over- or under-contributed relative to the average contribution in the group in the previous round are also included to control for the tendency towards conformity and the aversion to being a 'sucker' (Bougherara et

al. 2009). Face is a dummy variable that equals one if contribution decisions were made under the 'face' treatment in Sequence 2. Face2 is a dummy variable that equals one for decisions made under the Saving-Face in Sequence 3. 'disapproval + sanction' and 'face + sanction' are dummy variables that equal one for decisions made under these conditions in Sequence 2, 'post-sanction disapproval' and 'post-sanction face' equal one for decisions made in the post-intervention periods in Sequence 3 in the context of the peer disapproval and saving face mechanisms, respectively. We also include a dummy variable for gender, equal to one if the subject is male.

In the pooled regression, we observe the typical negative time trend that characterizes behavior in public good games. In line with Ashley et al. (2010), we also find that the initial contribution made by a subject in period 1 of baseline play significantly predicts their play throughout the rest of the game. The lagged contribution variable is also positive and significant, indicating a positive correlation between contribution amounts made in the previous and current periods. We observe that both types of internal norm enforcement mechanisms raise contributions to a similar degree relative to baseline levels (our reference period) in both the short term. Comparing 'disapproval2' and 'face2', we note that peer disapproval appears to be more effective than saving face in the long term. When combined with an external enforcement mechanism, the saving face mechanism appears to be more effective than anonymous peer disapproval. We observe that in a context of anonymous peer disapproval, the parameter estimate associated with the post-sanction variable is negative and significant, indicating a marked post-intervention crowding out of prosocial motivations that yields average contribution amounts significantly lower than baseline conditions, our reference condition.[10] Under conditions designed to leverage face-saving concerns, in contrast, the parameter associated with the post-sanction

[10]An analysis of disapproval points reveals that this decrease is not the result of a decrease in disapproval points sent. Indeed, disapproval points in the post-intervention period are sent with even greater frequency than in previous periods. Instead people no longer seem sensitive to receipt of disapproval. See Romaniuc et al. (2016) for further discussion.

covariate is *positive* and statistically significant, suggesting that, not only does this type of internal enforcement mechanism appear to be robust to post-intervention crowding out, but that it is also able to maintain contribution amounts at a level significantly higher than baseline conditions. It thus appears that removing en external enforcement mechanism has no detrimental effect on contributions in the continued presence of face-saving concerns, and that it can yield even higher contributions than under baseline conditions.

Table 3. Censored panel regression: contributions to the group account

Variable	*Parameter estimates (s.e.)*
period	-0.551***
	(0.304)
contribution in period 1	0.304***
	(0.048)
contribution in period t-1	1.04***
	(0.027)
under-contributed in t-1	-0.511***
	(0.038)
over-contributed in t-1	-0.722***
	(0.040)
disapproval	1.691**
	(0.514)
disapproval2	1.225*
	(0.527)
disapproval + sanction	4.855***
	(0.559)
post-sanction disapproval	-2.671***
	(0.576)
face	1.319**
	(0.401)
face2	0.907*
	(0.402)
face + sanction	5.432***
	(0.471)
post-sanction face	1.093*
	(0.453)
constant	-1.318*
	0.563)
Proportion censored at 0%:	0.276
Proportion uncensored:	0.572
Proportion censored at 100%:	0.152
N =	5255
Log likelihood =	-12572.161

Saving face emerges as the superior internal mechanism. First, as evidenced by the greater magnitude of the 'face + sanction' relative to the 'disapproval + sanction' parameters, the saving face mechanism appears to have a greater positive impact on contributions when implemented in tandem with an external sanction compared to peer disapproval. Second, our regression results confirm our previous tests, indicating the presence of a strong negative post-intervention effect in the context of peer disapproval, and a *positive* post-intervention effect in the context of the saving face mechanism.

4 Discussion

In this paper, we investigate the interplay between a formal, external norm enforcement mechanism, in the form of a monetary sanction, and two different types of internal enforcement mechanisms: anonymous peer disapproval and face-saving concerns. We find that while cooperation suffers from a negative behavioral spillover following the removal of an external enforcement mechanism under conditions of peer disapproval, no such crowding out occurs under face-saving conditions.

One interpretation of these results could be that the persistence of the expressive function of law depends on having the requisite conditions to support its continued enforcement, without which its expressive message may no longer be credible. As an internal enforcement mechanism, anonymous peer punishment does not appear to provide the social conditions necessary to support continued compliance to the norm. In contrast, we find that face-saving concerns appear to fulfill these conditions, not only managing to mitigate the negative spillover observed under conditions of anonymous peer punishment, but even maintaining cooperation at levels slightly higher than the no-sanction scenario. It should be noted that it fulfills these conditions despite the fact that no punishment is actually distributed among group members. Instead, the effectiveness of this type of enforcement mechanism is thought

to rest on the perceived threat of damage to one's 'face,' or social image. This suggests that policymakers could do well to seek ways to make behavior in social dilemmas observable, as doing so appears to create a strong social incentive to cooperate even once an external enforcement mechanism has been removed.

In economics, the social reality in which economic behavior takes place is increasingly recognized as an important element of decision context. In these social contexts, norms dictate what is acceptable and unacceptable behavior that can entail subsequent rewards or punishments. This work contributes further evidence of the importance of social forces in shaping the landscape of the incentives that actors face. Our results moreover suggest that the social environment can be an important factor in determining the degree to which sanctions are successful in the short term, as well as their legacy once they are no longer in place. In this way, we demonstrate that social context – notably the norm-enforcement mechanisms available – is a crucial determinant of the stickiness of beneficial norms over time. Given that external enforcement mechanisms serve to coordinate expectations around certain norms of conduct and internal enforcement mechanisms often serve as added incentives for compliance, pursuing a better understanding of the interplay between the two seems to be a highly important direction for continuing research.

Acknowledgements : This work benefited from funding provided by the Montpellier Laboratory for Theoretical and Applied Economics and the Catholic University of Lille.

References

Andreoni, J. and L. K. Gee (2012). "Gun for hire: Delegated enforcement and peer punishment in public goods provision". In: *Journal of Public Economics* 96.11-12, pp. 1036–1046.

Andreoni, J. and R. Petrie (2004). "Public goods experiments without confidentiality: a glimpse into fund-raising". In: *Journal of Public Economics* 88.7-8, pp. 1605–1623.

Ashley, R., S. Ball, and C. Eckel (2010). "Motives for Giving: A Reanalysis of Two Classic Public Goods Experiments". In: *Southern Economic Journal* 77.1, pp. 15–26.

Aviram, A. (2004). "A paradox of spontaneous formation: The evolution of private legal systems". In: *Yale Law & Policy Review* 22, pp. 1–68.

Baldassarri, D. and G. Grossman (2011). "Centralized sanctioning and legitimate authority promote cooperation in humans". In: *Proceedings of the National Academy of Sciences of the United States of America* 108.27, pp. 11023–11027.

Benson, B.L. (1991). "An Evolutionary Contractarian View of Primitive Law: The Institutions and Incentives Arising Under Customary American Indian Law". In: *Review of Austrian Economics* 5.1, pp. 41–65.

Bernold, E. et al. (2015). "Social framing and cooperation: The roles and interaction of preferences and beliefs". In: *Available at: http://dx.doi.org/10.2139/ssrn.2557927*.

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.

Boettke, P. J., C. J. Coyne, and P. T. Leeson (2008). "Institutional stickiness and the new development economics". In: *American Journal of Economics and Sociology* 67.2, pp. 331–358.

Bohnet, I. and B. S. Frey (1999). "The sound of silence in prisoner's dilemma and dictator games". In: *Journal of Economic Behavior & Organization* 38.1, pp. 43–57.

Bowles, S. and H. Gintis (2006). "Prosocial emotions". In: *Economy as an Evolving Complex System, III*, pp. 339–366.

Bowles, S. and S. H. Hwang (2008). "Social preferences and public economics: Mechanism design when social preferences depend on incentives". In: *Journal of Public Economics* 92.8-9, pp. 1811–1820.

Bursztyn, L. and R. Jensen (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure". In: *Annual Review of Economics, Vol 9* 9, pp. 131–153.

Carpenter, J. and C. K. Myers (2010). "Why volunteer? Evidence on the role of altruism, image, and incentives". In: *Journal of Public Economics* 94.11-12, pp. 911–920.

Cooley, C.H. and H.J. Schubert (1998). *On Self and Social Organization*. University of Chicago Press.

Cooter, R. (1998). "Expressive law and economics". In: *Journal of Legal Studies* 27.2, pp. 585–608.

Coricelli, G., E. Rusconi, and M. C. Villeval (2014). "Tax evasion and emotions: An empirical test of re-integrative shaming theory". In: *Journal of Economic Psychology* 40, pp. 49–61.

Dufwenberg, M., S. Gachter, and H. Hennig-Schmidt (2011). "The Framing of Games and the Psychology of Play". In: *Games and Economic Behavior* 73.2, pp. 459–478.

Ellickson, R. C. (1991). *Order without law: How neighbors settle disputes*. Cambridge: Harvard University Press.

— (1998). "Law and economics discovers social norms". In: *Journal of Legal Studies* 27.2, pp. 537–552.

Elster, J. (1989). "Social Norms and Economic-Theory". In: *Journal of Economic Perspectives* 3.4, pp. 99–117.

Engel, C. (2014). "Social preferences can make imperfect sanctions work: Evidence from a public good experiment". In: *Journal of Economic Behavior & Organization* 108, pp. 343–353.

Fehr, E. and S. Gachter (2000). "Cooperation and punishment in public goods experiments". In: *American Economic Review* 90.4, pp. 980–994.

Feinberg, M., J. T. Cheng, and R. Willer (2012). "Gossip as an effective and low-cost form of punishment". In: *Behavioral and Brain Sciences* 35.1.

Feldman, Y. (2009). "The Expressive Function of Trade Secret Law: Legality, Cost, Intrinsic Motivation, and Consensus". In: *Journal of Empirical Legal Studies* 6.1, pp. 177–212.

Friedman, D. (1979). "Private creation and enforcement of law – A historical case". In: *Journal of Legal Studies*, pp. 399–415.

Gachter, S. and E. Renner (2010). "The effects of (incentivized) belief elicitation in public goods experiments". In: *Experimental Economics* 13.3, pp. 364–377.

Gneezy, U. and A. Rustichini (2000). "A fine is a price". In: *Journal of Legal Studies* 29.1, pp. 1–17.

Guala, F. (2012). "Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate". In: *Behavioral and Brain Sciences* 35.1.

Ho, D. Y. F. (1976). "Concept of Face". In: *American Journal of Sociology* 81.4, pp. 867–884.

Karakostas, A. and D. J. Zizzo (2016). "Compliance and the power of authority". In: *Journal of Economic Behavior & Organization* 124, pp. 67–80.

Masclet, D. et al. (2003). "Monetary and nonmonetary punishment in the voluntary contributions mechanism". In: *American Economic Review* 93.1, pp. 366–380.

Milgram (1963). "Behavioral study of obedience". In: *Journal of Abnormal Social Psychology* 67, pp. 371–378.

Nelissen, R. M. A. and L. B. Mulder (2013). "What makes a sanction "stick"? The effects of financial and social sanctions on norm compliance". In: *Social Influence* 8.1, pp. 70–80.

Nikiforakis, N. (2008). "Punishment and counter-punishment in public good games: Can we really govern ourselves?" In: *Journal of Public Economics* 92.1-2, pp. 91–112.

Ostrom, E., J. Walker, and R. Gardner (1992). "Covenants with and without a Sword - Self-Governance Is Possible". In: *American Political Science Review* 86.2, pp. 404–417.

Posner, R. A. and E. B. Rasmusen (1999). "Creating and enforcing norms, with special reference to sanctions". In: *International Review of Law and Economics* 19.3, pp. 369–382.

Rege, M. and K. Telle (2004). "The impact of social approval and framing on cooperation in public good situations". In: *Journal of Public Economics* 88.7-8, pp. 1625–1644.

Romaniuc, R. et al. (2016). "The perils of government enforcement". In: *Public Choice* 166.1-2, pp. 161–182.

Sened, I. (1997). *The Political Institution of Private Property.* Cambridge: Cambridge University Press.

Subhasish, D. (2013). "Non-Monetary Incentives and Opportunistic Behavior: Evidence from a Laboratory Public Good Game". In: *Economic Inquiry* 51.2, pp. 1374–1388.

Sunstein, C. (1996). "On the expressive function of law". In: *University of Pennsylvania Law Review* 144, pp. 2021–2053.

Tognetti, A. et al. (2013). "Is cooperativeness readable in static facial features? An inter-cultural approach". In: *Evolution and Human Behavior* 34.6, pp. 427–432.

Williamson, C. R. (2009). "Informal institutions rule: institutional arrangements and economic performance". In: *Public Choice* 139.3-4, pp. 371–387.

Zizzo, D. J. (2010). "Experimenter demand effects in economic experiments". In: *Experimental Economics* 13.1, pp. 75–98.