The stickiness of norms

Katherine Farrow, Rustam Romaniuc

The Stickiness of Norms

Katherine Farrow* and Rustam Romaniuc†

Abstract

In this paper we study the role of social context, as characterized by different informal norm-enforcement mechanisms, on the deterrence legacy of temporary external regulations. In a public goods 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 social environments an external regulation that is implemented for a limited period of time and then removed. We observe a significant negative post-intervention effect of this removal in the context of peer disapproval, but no such effect 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.

*EconomiX, University of Paris Nanterre, 200 Ave de la République, 92000 Nanterre, France. *Email address:* katherine.farrow@parisnanterre.fr

†Catholic University of Lille – ETHICS-EA7446 & 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 that is 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 depend on government for either their 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 influenced by informal norms.

During the 1990s, the relationship between formal rules and informal norms became topical within the field of law and economics. 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 an area of research referred to as the law-and-economics of norms (Feldman 2009) and by the attention given to the topic in prominent law journals.[1]

On a more fundamental level, this literature has focused on the relative effectiveness of two types of deterrence: peer-enforced norms versus publicly-enforced, codified laws. Some have argued that peer-enforcement can serve as an adequate deterrent, effectively establishing and maintaining cooperation (Ellickson 1991). Others claim that informal mechanisms are not, in fact, effective and that formal rules enforced by external authorities are requisite elements of a stable social order.[2]

[1] 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. 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).

[2] The former contingent point to a large array of historical examples as evidence of the feasibility of

The focus on the comparative advantages of one system versus the other, surprisingly, neglects the fact that formal rules may not be effective without the support of corresponding informal social norms (Boettke *et al.* 2008). Indeed, insofar as informal norms can be considered an inherent element of the fabric of society (Elster 1989; Bicchieri 2006), they necessarily precede formal rules and therefore serve an important legitimizing function.[3] This literature also suggests that informal norms tend to be stickier than formal rules (Carbonara *et al.* 2012). The well-established interactions between these two systems gives us reason to believe that formal rules may affect the functioning of informal norms not only when they are implemented, but also when they are lifted. The long-lasting effects of temporary, formal rules on informal norms has received relatively limited attention within the law and economics literature to date.

Gneezy and Rustichini's (2000) classic field study documenting the capacity for formal sanctions to crowd-out private enforcement serves as a natural motivation for the present study. Here we are particularly interested in the potential for these effects to persist even after a formal rule is no longer in force. We are interested in whether the mechanism by which an informal norm is enforced may have implications for its robustness to post-intervention crowding-out effects. Given the problematic nature of manipulating and measuring the impact of temporary formal rules on informal norms in the field,[4] we turn to the use of experimental economic methods to investigate post-

order without "the backing of state authority" (Benson 1991). Informal norms appeared to successfully maintain social order in primitive and medieval societies ((Benson 1991; Friedman 1979) and continue to do so in contemporary societies (Ellickson 1991; Bernstein 1992). These types of informal enforcement mechanisms are understood to rely on shame – a disutility one suffers when others identify him as offending an established norm of conduct (Elster 1998; Bowles and Gintis 2006; Masclet *et al.* 2003; Guala 2012).

[3]The idea that norms are sticky has been put forward by spontaneous order theorists, who emphasize that care should be taken in establishing new rules that are designed and enforced by public authorities (Boettke *et al.* 2008; Williamson 2009). This observation has also been advanced by the law-and-economics of norms literature (Feldman 2009), in which the argument has been made that formal rules act as focal points in a landscape of informal norms typically characterized by multiple equilibria (Sunstein 1996; Cooter 1998). To the extent that formal rules indicate what appropriate behavior is, they too can reinforce informal normative mechanisms such as peer pressure.

[4]It is for this reason that the crowding-out of social disapproval in Gneezy and Rustichini (2000) is

intervention dynamics between the two in a controlled way.

In what follows, we carry out a laboratory public goods experiment in which we add and subsequently remove an externally-enforced formal rule, in the form of a monetary sanction, in the context of two different informal norm enforcement mechanisms. In a first treatment, we give individuals the opportunity to send anonymous, non-costly disapproval points to each other based on contributions made in the previous round. The inclusion of this treatment was motivated by a robust finding in the literature demonstrating that informal norms are supported by low-cost expressions of social disapproval, such as ridicule and gossip (Ellickson 1991; Boehm 1999; Feinberg *et al.* 2012; Guala 2012).

In the second treatment, we implement another informal mechanism that has been shown to invoke remorse among those who deviate from normative behavior. Specifically, 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). 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. Thus, following every round in this treatment, we display a picture of each group member next to his contribution. While the first informal norm enforcement mechanism we implement relies on the explicit expression of disapproval by one's peers, the latter rests on one's belief about how he is perceived by the others around him (Andreoni and Petrie, 2004; Bursztyn and Jensen 2017). The novelty of this work is that we use controlled laboratory conditions to measure the impact of social context, as characterized by different norm-enforcement mechanisms (disapproval or face-saving), on the legacy of temporary monetary sanctions. In doing so, we aim to investigate the importance of norm-enforcement mechanisms in determining the robustness of norm adherence in the long run.

proposed as only one of a number of possible explanations for their results.

In each of our treatments, subjects progress through three different sequences in the experiment.[5] In the first sequence of all treatments, subjects play a standard public goods game for ten periods. In the second sequence (periods 11-20) of the control treatments, we either provide the opportunity to express social disapproval or display subjects' photographs next to their contributions at the end of each period. In the second sequence of the test treatments, we complement the informal mechanism (disapproval points or photographs) with an exogenously-imposed monetary sanction.[6] In the third sequence of the game, we remove the monetary sanction in the test treatments, leaving the informal mechanisms (disapproval points or photographs) in place.

We find a striking difference in the effectiveness of these informal enforcement mechanisms once the external mechanism has been removed. In a social context characterized by peer disapproval, we observe strong negative post-intervention crowding-out: in the treatment where subjects had been exposed to monetary sanctions, cooperation under peer-disapproval in the third sequence falls to levels below those observed under baseline conditions, i.e. when subjects had not been exposed to external monetary sanctions. This is not the case for the treatment leveraging face-saving concerns, where post-intervention cooperation levels remain higher than under baseline conditions. Insofar as this face-saving mechanism appears to be robust to the negative post-intervention effect that is observed in the context of peer disapproval, these results suggest that increasing the saliency of social image concerns may ultimately be a more suitable deterrent strategy than relying on anonymous peer disapproval, and especially so in conjunction with the use of formal, external enforcement mechanisms. In this way, our findings suggest that face-saving concerns lead to the creation of stickier norms than anonymous peer-disapproval.

[5] We use a partner-matching protocol in which the composition of groups remains fixed throughout the experiment.

[6] The sanction is mild in that the dominant strategy remains free-riding.

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 goods game. Subjects in our game are assigned to groups of four and endowed with $E_i = 20$ tokens. They 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 nonetheless 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 baseline conditions. This serves to familiarize subjects with the game and create a challenging environment in which cooperation can arise, as subjects are able to become accustomed to the high levels of free-riding that typically characterize play in the public goods game by the end of the first ten periods. Subjects are informed that the experiment will consist of three sequences, and that they will be provided instructions for the each sequence at the relevant time.

Our experimental manipulations consist of two variations to the standard public

goods game, which are designed to mimic an external enforcement mechanism, based on a monetary sanction meted out by the experimenter, and two types of informal enforcement mechanisms: one based on anonymous peer disapproval and the other based on social image. In the *Peer Disapproval* condition, participants are informed at the beginning of the second sequence (periods 11-20) that they will now be able to see the individual contributions of their group members after every round and will have the opportunity to send points of disapproval to the other members of their group. Subjects can send anywhere between 0 and 10 disapproval points, where 0 indicates no disapproval and 10 indicates strong disapproval of another group member's contribution in that round.

In the *Saving Face* condition, at the beginning of the second sequence of the game, participants are informed that their photograph will now appear next to their contribution amounts, which will be visible to the rest of the members of their group after each round of play. Information regarding individual contributions are therefore displayed under both of the *Peer Disapproval* and *Saving Face* conditions. In the *Saving Face* and *Peer Disapproval* treatments, we inform subjects at the beginning of the third sequence that it will be conducted in the same way as the second sequence.

In the *Saving Face + Sanction* and *Peer Disapproval + Sanction* treatments, we introduce an external enforcement in the form of a monetary punishment in period 11 along with either peer disapproval or saving face, and this characterizes game play until period 20. In the third sequence of the game (periods 21-30) we remove the external mechanism, leaving only the informal mechanism (either peer disapproval or saving face) in effect for the remainder of the experiment. Thus, the only difference between the respective *Sanction* and *No Sanction* test and control treatments in each social context is the presence or absence of a monetary sanction 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. 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. Namely, 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. Given that the sanction amount is less than the marginal per capita rate of return, a self-interested individual will not, theoretically, contribute to the public account, which is also the case for the baseline condition.

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). In public goods experiments, it is generally assumed that subjects understand that the desirable behavior is behavior that favors the interest of the group, and accordingly, that deviations from this behavior are undesirable (e.g. Andreoni and Gee 2012). Our treatment makes salient this contribution norm by emphasizing the wrongdoing that is

implied in not contributing. We avoid referring to the penalty using words such as tax, punishment, or sanction, however, in order to minimize experimenter demand effects (Zizzo 2010) and avoid the possibly varied connotations that participants may attach to these words.

To mimic the centralized nature of a government-like sanction, 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 goods 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* treatments 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.[7] A total of 196 subjects participated in our experiment. None of them had previously participated in a public goods experiment. Subjects interacted through individual computer terminals using the LE2M software programmed by engineers at LEEM and Anthropo-Lab. 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 were used in the experiment, the reading of the instructions, and distribution of payments. Table 1 provides detailed

[7]It is worth noting that the two laboratories follow the same recruiting and experimental procedures. 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.

information about the number of groups in each treatment and characteristics of the treatments.

Table 1. Experimental treatments

Treatment	**Groups**	**Sequence 1** Periods 1-10	**Sequence 2** Periods 11-20	**Sequence 3** Periods 21-30
Peer disapproval	9	Baseline	Peer Disapproval	Peer Disapproval
Peer disapproval, Sanction	10	Baseline	Peer Disapproval + Sanction	Peer Disapproval
Saving face	15	Baseline	Saving Face	Saving Face
Saving face, Sanction	15	Baseline	Saving Face + Sanction	Saving Face

In the *Saving Face* treatments, 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 to have their photograph taken. In order to preserve social distance between the experimenter and the subjects, the assistant who took subjects' pictures 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.[8] Participants were then shown to the laboratory where the game was explained and two example scenarios were reviewed.

At the outset of each session, subjects were informed that the central server would randomly assign them to groups of four people, and that each session would consist of 30 periods divided into three sequences of 10 periods. The total number of sequences in the session was therefore common knowledge, as was the fact that at the end of the experiment only one sequence out of the three sequences would be chosen at random for payment.

[8]We followed the procedure described in Tognetti et al. (2013).

3 Theoretical background

Applied to the public goods game, traditional rational choice theory assumes that social or internalized norms will have no impact on contribution behavior. Under this framework, monetary punishments are considered to change behavior only when they are optimal, that is, when option X is made more attractive relative to option Y in monetary terms. We evaluate our findings based on the benchmark predicted by this theory, namely, zero contributions to the public good across all treatments and throughout the different sequences in our experiment.

In the standard public goods game played in the first sequence of ten periods, it can easily be seen that the dominant strategy is for all subjects to keep all 20 Ecus in their private account and contribute nothing to the public account. In equilibrium, this yields a gain of 80 Ecus at the group level. Alternatively, if all group members contributed their entire endowments to the public account, the individual gain would amount to 32 Ecus, and total group earnings would amount to 128 Ecus. However, as is well-known, a rational money-maximizing agent would pursue the benefit to be had by deviating from this strategy in favor of complete free-riding, hoping for a private gain of 44 Ecus. Since the game is symmetric, this strategy is assumed to be adopted by everyone, leading each subject to end up with their initial endowment of 20 Ecus.

Under peer disapproval and face-saving conditions, the subgame perfect equilibrium remains the same as in the first sequence of the game. For the control treatments involving informal norm enforcement mechanisms, this is because peer disapproval and face-saving concerns are irrelevant to the traditional rational agent. For the test treatments involving a formal enforcement mechanism, the monetary sanction is also non-deterrent, as explained above.

In contrast to these predictions, however, repeated public goods experiments employing parameters similar to ours have shown that subjects tend to contribute

about 40% of their endowment in the first period and then reduce their contributions to reach virtually the game theoretic prediction by the final periods (Gachter 2014). Empirical evidence leads us to expect subjects in the standard public goods game to behave differently than the game theoretic predictions presented above. Further, any impact of peer-disapproval, photographs, and monetary sanctions must be attributed to behavioral factors such as the desire to avoid social disapproval[9], individuals' image concerns,[10] and the reinforcement or reduction of these two informal enforcement mechanisms by the exposure of subjects to exogenously imposed monetary sanctions.[11]

The lasting impacts of removing exogenously imposed monetary sanctions have received limited attention within the law and economics literature. In the absence of extant theoretical and empirical work on how temporary monetary sanctions may affect the success of peer disapproval and image concerns in maintaining social order, we consider the different motivations behind each of type of informal enforcement mechanism in order to formulate predictions regarding the way these motivations may be impacted by temporary exposure to exogenously-imposed monetary sanctions.

Peer disapproval relies on the explicit expression of disapproval by one's peers. In Masclet et al. (2003), targets of peer-disapproval change their behavior when the disapproval is perceived as legitimate. That is, the effectiveness of peer disapproval

[9]Ostrom et al. (1992) made the first attempts to design a laboratory experiment in order 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. An experiment that allowed subjects to directly communicate the extent of their disapproval is Masclet *et al.* (2003). Subhasish (2013) and Nelissen and Mulder (2013) followed and confirmed the seminal result from Masclet *et al.* that the possibility of receiving peer disapproval increases compliance with cooperation norms.

[10]There exists a good deal of evidence that social image is an important motivator of individual action. Rege and Telle (2004) find that linking players' identities with their contribution amounts significantly raises contributions in a public goods game. Bohnet and Frey (1999) compare play in public goods games and in dictator games under anonymity, one-way identification, two-way identification, and communication. They find that one-way identification doubles 'solidarity' in both public goods games and dictator games.

[11]Andreoni and Gee (2012), Xiao and Houser (2011), Stagnaro et al. (2016), and Pysakhovich and Rand (2015) experimentally examine the conditions under which monetary sanctions that are meted out by a third-party can increase the effectiveness of private enforcement. Gneezy and Rustichini (2000) use a field experiment, and Funk (2007) uses observational data, to show that the removal of exogenously imposed sanctions can reduce the effectiveness of private enforcement.

depends on some factor external to the individual expressing it. Exogenously-imposed sanctions make appropriate behavior salient, thus legitimating the disapproval expressed by community members for deviating from this behavior. Consequently, the removal of exogenously-imposed sanctions changes the normative character of contributing to the public good and, accordingly, of expressions of disapproval. This is also how Gneezy and Rustichini (2000) interpret some of their results. Specifically, they suggest that parents who were late to pick up their children from a daycare center may not have felt shame after the center began imposing a small fine for doing so.

Contrary to disapproval, image motivations rest on one's belief about how he/she is perceived by the others around him/her (Bursztyn and Jensen 2017). DellaVigna et al. (2012) found that when asked for donations face-to-face, people give money they would not otherwise have given. The authors suggest that people comply so as to avoid creating a negative social image of themselves in front of a stranger who they will most likely never see again. Notably, social image concerns can be relevant even in the absence of any form of explicit communication (e.g. disapproval points). This feature of the face-saving mechanism may serve to encourage continued adherence to the social norm and prevent contributions from falling dramatically after sanctions are removed. For these reasons, we expect the removal of sanctions to reduce the legitimacy of explicit social disapproval while leaving the threat of losing face one's quite intact.

4 Results

The presentation of the results is divided into three parts. Because the study of crowding-in(out) effects resulting from the removal of monetary sanctions requires that there would be informal norms to be crowded-in(out) in the first place, we begin with a manipulation check by examining the extent to which the two informal norm-enforcement mechanisms, i.e. peer disapproval and face-saving concerns, are successful in creating and maintaining

norms of cooperation over time. Second, we investigate whether implementing and subsequently removing an exogenously-imposed monetary sanction in an environment characterized by preexisting informal norms (created through either peer-disapproval or the face-saving mechanism) improves, reduces, or leaves unaffected group cooperation. Third, we conduct regression analyses in order to investigate the factors that influence individual contribution decisions across treatments in more detail.

4.1 Do peer-disapproval and face-saving mechanisms establish norms of cooperation?

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

Treatment	**Periods 1-10**	**Periods 11-20**	**Periods 21-30**
Peer disapproval	Baseline 5.98 (1.25)	Disapproval 8.60 (1.14)	Disapproval 9.31 (1.54)
Peer disapproval, Sanction	Baseline 4.07 (1.37)	Disapproval + sanction 12.45 (3.06)	Disapproval 3.69 (2.47)
Saving face	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.[12] We furthermore note that contributions in

[12]The fact that participants had their pictures taken before the experiment began means that we may expect baseline behavior to differ across disapproval and face-saving treatments. When pooling baseline contributions across sanctioning mechanisms (i.e. according to whether pictures were taken or not), no

Sequence 1 in all treatments follow the same pattern over time, and arrive at virtually identical average contribution levels in period 10.

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 ($z = 4.536$, $p < 0.001$). 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).[13]

Manipulation check. *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.*

4.2 Are both types of informal mechanisms subject to post-intervention crowding out?

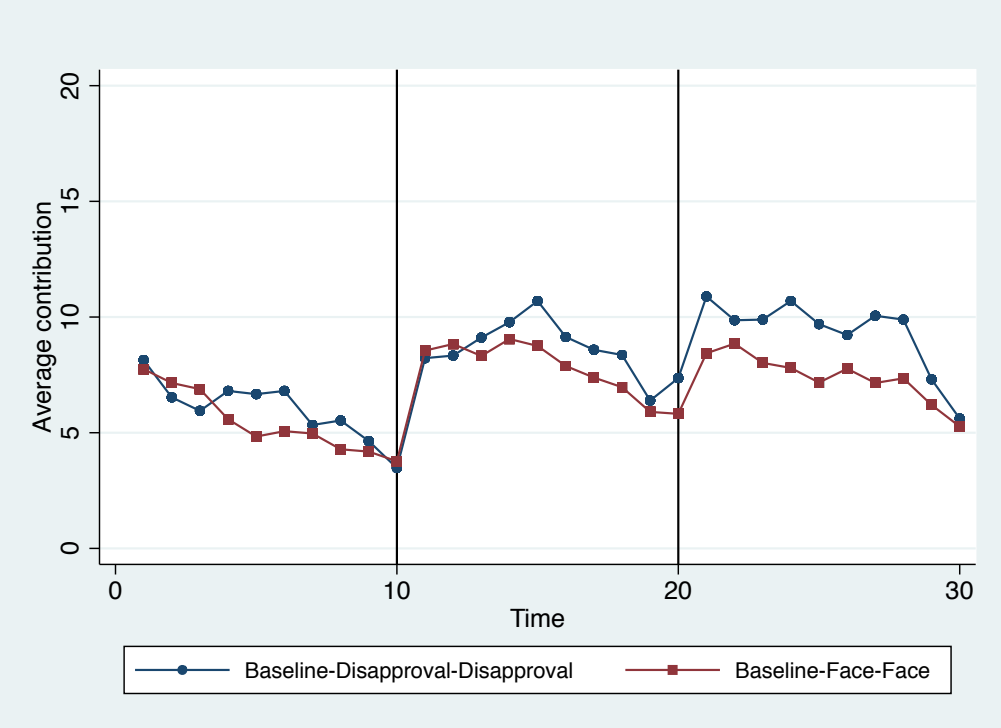
To investigate the presence of post-intervention crowding out effects, we carry out a between-subject comparison of contribution levels in the *Sanction* treatments with those in the *No Sanction* treatments over the final 10 periods of play in Sequence 3. By the time that subjects have reached this point in the game, those in the *Sanction* treatment will have been subject to 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, in the context of peer disapproval, we observe a significant decline in contribution levels after the sanction has been removed. A Mann-Whitney

significant difference exists (Mann-Whitney test: p = 0.15, with average contributions of 5.03 vs. 5.83, respectively).

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

Figure 1. Mean contributions under informal norm-enforcement mechanisms

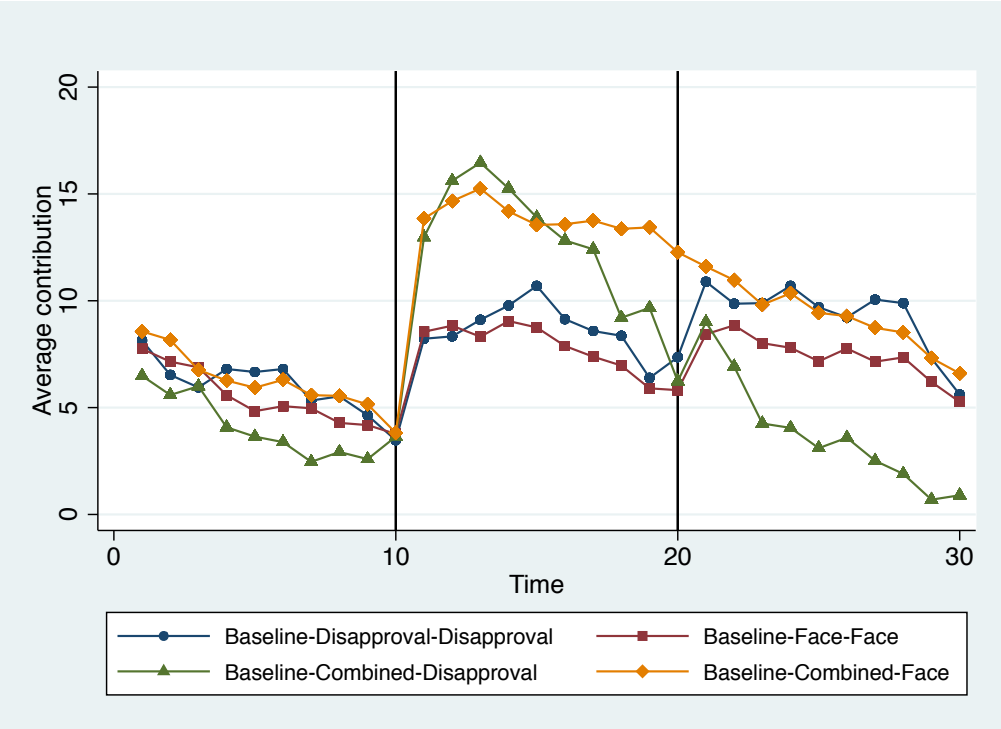


test rejects the null hypothesis that contribution levels across treatments are drawn from the same underlying distribution in the post-intervention period ($z = 3.554$, $p < 0.001$), suggesting that anonymous peer disapproval is indeed vulnerable to a strong negative post-intervention effect resulting 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 contributions at an even higher level 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. Furthermore, 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 informal and external norm-enforcement mechanisms

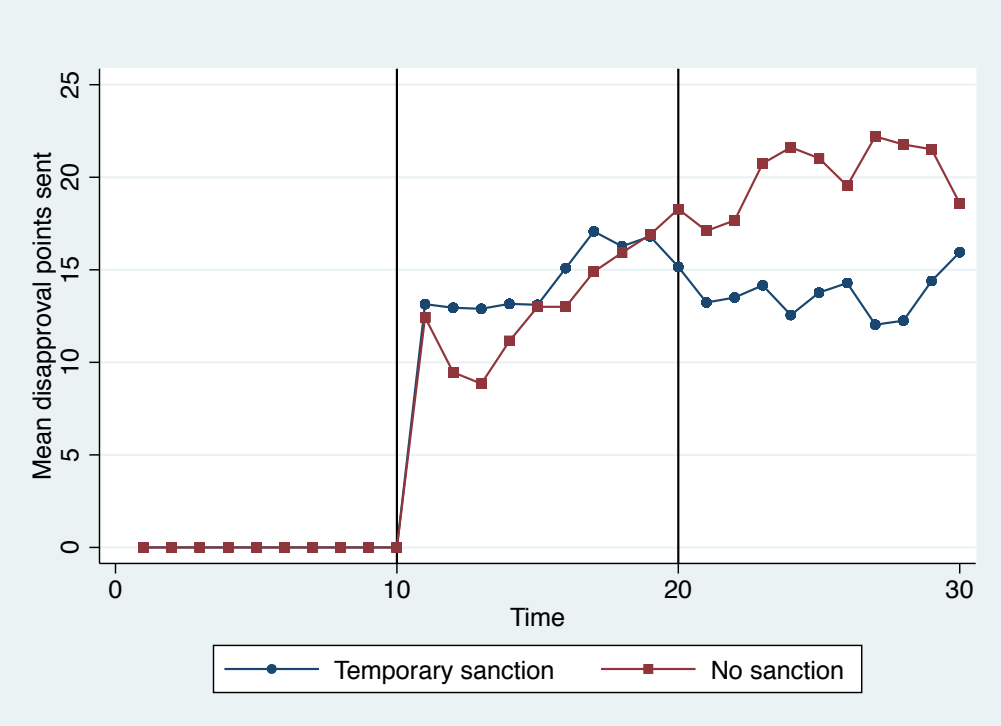


Result. *The removal of an externally-enforced sanction generates a negative behavioral spillover in the post-intervention period under peer disapproval. In contrast, the removal of an external sanction generates a positive behavioral spillover under saving face.*

As noted in Romaniuc et al. (2016), the expressive function of removing the sanction may have implications for the legitimacy of those who punish in the post-intervention period. An analysis of the number of disapproval points sent across *Sanction* and *No Sanction* peer disapproval treatments can shed light on the source of the negative post-intervention effect we observe. Figure 3 depicts the average number of disapproval points sent within groups across these 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, $p < 0.0002$).[14]

Figure 3. Mean number of disapproval points sent per group



4.3 Analysis of individual contributions

To estimate the relative importance of a variety of factors in determining contribution amounts in each period, we conduct panel regressions, which lend further support to our main results reported above. A significant Hausman test ($p < 0.001$) leads us to specify a fixed-effects panel regression with standard errors clustered at the group level. In order to account for temporal behavioral dynamics throughout the game, we

[14]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.

incorporate several lagged variables.[15] Since the two informal mechanisms may have unanticipated implications for the dynamics of play in the game, we estimate separate models based on each sample.[16] The model is specified as follows:

$$Y_{it} = \beta_{1it} X_{it} + \alpha_i + u_{it}$$

Where i identifies the participant and t identifies the time period. Y_{it} is the dependent variable, contributions made to the group account, X_{it} is the vector of time-varying independent variables, elaborated on below, α_i are the player-specific fixed effects (intercepts), and u_{it} is the error term. The first independent variable we include is a within-part 'period' variable indicating the period number (1-10), which is intended to capture a time trend. Dummy variables indicating whether 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). Given the inclusion of these lagged variables, observations are confined to periods 2 on. 'Sequence 2' is a dummy variable that equals one if contribution decisions were made in periods 11-20, and 'Sequence 3' is a dummy variable that equals one for decisions made in periods 21-30. These are interacted with 'period' in order to evaluate how time trends may themselves change under conditions in different stages of the game. 'Sequence 2 sanction' and 'Sequence 3 post-sanction' are dummy variables that equal one for decisions made in periods 11-20 and 21-30 of the *Sanction* treatments. Given this specification, the reference group in each model refers to contributions made in periods 1-10 under baseline conditions.

[15] To investigate the extent to which endogeneity may be an issue in our models, we also estimate a censored, mixed-effects GMM specification, which addresses potential endogeneity by allowing for the specification of latent variables. The results obtained from these models are qualitatively similar to the simpler and more easily interpretable specifications we report here.

[16] For example, taking participants' photographs prior to game play in the saving face treatments could conceivably alter their behavior in baseline periods even if their photographs are not used during these periods.

Table 3. Panel regressions: contributions to the group account

Variable	*Parameter estimates (s.e.)*	
	Peer disapproval	Saving face
Reference condition:	Baseline	Baseline
period (within-part)	-0.111	-0.120
	(0.071)	(0.070)
contribution in period t-1	0.640***	0.663***
	(0.060)	(0.042)
under-contributed in t-1	-0.192**	-0.387***
	(0.063)	(0.048)
over-contributed in t-1	-0.519***	-0.413***
	(0.081)	(0.057)
Sequence 2	2.528**	2.475***
	(0.830)	(0.553)
Sequence 2 * period	-0.176	-0.228*
	(0.117)	(0.088)
Sequence 3	2.675*	1.684**
	(1.324)	(0.601)
Sequence 3 * period	-0.246	-0.140
	(0.146)	(0.097)
Sequence 2 sanction	7.713***	5.561***
	(1.00)	(0.959)
Sequence 2 sanction * period	-0.738***	-0.377**
	(0.112)	(0.111)
Sequence 3 post-sanction	1.630	1.342
	(1.03)	(0.717)
Sequence 3 post-sanction * period	-0.335***	-0.060
	(1.10)	(0.076)
constant	2.834***	2.336**
	(0.709)	(0.719)
N =	2175	3480
ρ:	0.215	0.190
F =	110.01	104.47
R^2 =	0.496	0.574

*, **, and *** indicate p-values of less than 0.05, 0.01, and 0.001, respectively.

In the baseline periods, the negative time trend is weakly insignificant ($p = 0.133$ and $p = 0.096$, respectively). The lagged contribution variable is positive and significant, indicating a positive correlation between contribution amounts made in the previous and current periods. We observe that both types of informal norm enforcement mechanisms raise contributions to a similar degree relative to baseline levels in the short term, and that peer disapproval does not seem to suffer from a negative time trend (that is, beyond the trend exhibited in baseline periods) in the short term. Regarding the parameter estimates for Sequence 3 across models, anonymous peer disapproval appears to be more effective than saving face in the long term, and neither exhibit an exaggerated negative time trend relative to baseline periods. When combined with an external enforcement mechanism (Sequence 2 sanction), the disapproval mechanism appears to be more effective than saving face (7.713 vs. 5.561, respectively). However, it is associated with a significant negative time trend that is nearly twice as large as that observed under saving face conditions (-0.738 vs. 0.377, respectively).

Turning to the post-sanction periods, the parameter estimate associated with the post-sanction dummy variable not significant in either model, indicating that, after controlling for the included covariates, contributions are not significantly higher in these periods than in baseline periods.[17] The parameter estimate under face-saving conditions is, however, only slightly insignificant ($p = 0.071$), pointing to a tendency for contributions to remain higher in the post-sanction periods relative to baseline periods. Furthermore, with respect to time trend in the post-sanction periods, we observe a highly significant negative trend in the context of anonymous peer disapproval, and no such trend in the context of saving face ($p = 0.438$). It thus appears that, whereas removing an external enforcement mechanism leads to a rapid

[17] 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, it seems that people are no longer sensitive to receipt of disapproval. See Romaniuc et al. (2016) for further discussion.

decline in contributions in the context peer disapproval, it has no detrimental effect on contributions in the continued presence of face-saving concerns. Thus, our regression results confirm our previous tests, indicating the presence of a strong negative post-intervention effect in the context of peer disapproval, and no such effect in the context of the saving face mechanism.

5 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 informal 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 post-intervention crowding-out occurs under face-saving conditions. Since our experimental design focuses on demonstrating evidence of effect rather than evidence of mechanism' (Berriet-Solliec et al., 2014), it prevents us from identifying a causal mechanism responsible for these findings. However, the novelty, magnitude, and relevance of these results nonetheless represent an important contribution to the law and economics literature and point to fruitful directions for future research.

Specifically, these findings suggest that the persistence of the expressive function of law depends on having sufficient conditions to support its continued enforcement, and that without these conditions in place, its expressive message may no longer be credible. As an informal enforcement mechanism, anonymous peer punishment does not appear to provide the social conditions necessary to support continued compliance with a norm of cooperation. 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 in the no-sanction scenario. We note that this mechanism 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 (Grimalda et al. 2016). In these social contexts, norms dictate what is acceptable and unacceptable behavior, and shape expectations regarding anticipated rewards or punishments. This work provides further evidence of the importance of social forces in shaping the landscape of the incentives that actors face. Our results moreover suggest that the nature of the social environment can be an important factor in determining the degree to which formal rules are successful in the short term, as well as the persistent impacts of these rules even after they have been removed. 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 and their robustness to changing institutional contexts. Given that formal rules serve to coordinate expectations around certain norms of conduct and that informal norms can impact the effectiveness of these rules and 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 French Environment and

Energy Management Agency (ADEME). We thank Gianluca Grimalda, Lisette Ibanez and Fabrice Le Lec for their comments at various stages of this work. We are grateful to Dimitri Dubois and Flovic Gosselin for programming the experiment in LE2M.