Negative Economic Shocks and the Compliance to Social Norms

Francesco Bogliacino
Dipartimento di Scienze Economiche,
University of Bergamo, Bergamo
francesco.bogliacino@unibg.it[*]

Camilo Gómez
CERGE-EI, Prague
Camilo.Gomez@cerge-ei.cz

Rafael Charris
Economic Science Institute
Chapman University, Orange
rcharris@chapman.edu

Felipe Montealegre
University of Bologna, Bologna
andres.montealegre@studio.unibo.it

October 2022

[Click here for the latest version of the paper]

[*] Corresponding Author. Francesco acknowledges financial support from Open Evidence via contracts 18810 and 29047 with Universidad Nacional de Colombia. A special thanks to Roberto Galbiati, Zachary Grossman, Marcela Ibanez, Erin Krupka, Cesar Mantilla, Pietro Ortoleva, Daniel Parra, Leonardo Pejsachowicz, Gerhard Riener, Nuria Tolsa-Caballero, Eugenio Verrina, for comments and suggestions. Douglas Rodriguez and Sebastian Ramirez helped during data collection. We are very grateful to the participants at various seminars and conferences: BEBES, Universidad de Los Andes, Universidad del Valle, Cournot Seminar at Strasbourg, AMSE Seminar in Aix-en-Provence, Development Seminar in Goettingen, Warsaw Economic Seminar, GREDEG Seminar in Sophia Antipolis, ESA Conference 2021, LACEA Brain 2022, BEEC 2022 and NOBEC 2022. The usual disclaimer applies.

Abstract

We study why suffering a Negative Economic Shock (NES), i.e. a significant loss, may trigger a change in behaviour. We conjecture that people trade off concern for money with a conditional preference to follow social norms and that suffering a shock makes extrinsic motivation more salient, leading to more norm violation. We study this question experimentally: After administering losses on the earnings from a Real Effort Task, we analyze choices in prosocial and antisocial settings. To derive our predictions, we elicit social norms separately from behaviour. We find robust evidence that shock increases deviations from norms.

Keywords: Negative Economic Shocks; Social Norms; Norm compliance; Anti Social Behavior; Cooperation; Trust; Trustworthiness.

JEL Codes: C91; C92; D90; D91.

1 Introduction

A Negative Economic Shock (an NES from here on) is a large financial loss on earnings or accumulated assets. Shocks can be due to psychosocial stressors (divorce, job loss, injury) or traumatic events (violence, disasters). Exogenous negative shocks have been studied to understand the impact of poverty (Mani et al., 2013; Haushofer and Fehr, 2014; Carvalho et al., 2016)[1]. More recently, victims of negative economic shocks have tended to support extreme candidates at the elections, which has furthered the interest into how NES shapes behaviour.

However, when it comes to social behaviour, most of the literature focuses on the role of shocks as a collective threat. Facing the risk of aggregate shocks, societies develop tight cultural traits such as social sanctioning and norm compliance (Gelfand et al., 2017; Heinrich, 2020; Prediger et al., 2014). In a similar vein, exposure to warfare helps induces parochial prosociality as a form of collective insurance (Bauer et al., 2016). Similar arguments are extended to natural disasters (Cassar et al., 2017; Botchway and Filippin, 2021). This reasoning misses a point though. Suffering an NES has an independent effect on whether subjects follow social norms because this experience alters the relative cost of norm compliance. Even conditional on the same threat, the very fact that the impact of shocks is heterogeneous induces variations in norm compliance.

This article tries to address this issue. We conjecture that decision-makers (DMs) trade-off money and compliance to social norms (Kimbrough and Vostroknutov, 2018; Krupka and Weber, 2013; Bicchieri, 2006; List et al., 2004). Social norms are rules of behaviour that are contingent and for which a subject has a preference to conform, conditional on the expectation that most of the reference group follows in kind, and think it ought to be done (Bicchieri, 2006). Following social norms is costly. Punishing transgressors, avoiding cheating, and abstaining from free-riding involve carrying out a cost. If the decision-maker trades off the concern for money and the conditional preference to follow the social norm, she will face an increasing marginal cost of norm compliance when experiencing an NES, leading to more norm violations.

We derive this prediction from a behavioural model. Assuming that norms enter the utility function and participants are heterogeneous in their psychological cost of compliance, we ana-

[1] the comparison of the rich and the poor is confounded by environmental and individual factors, whereas shocks are a plausible source of variation or can be manipulated in the lab

lyze optimal behaviour in several binary decision problems where a substantive norm applies, considering both anti-social and pro-social tasks. In all these settings, participants should decide whether to harm their counterparts. Sometimes this action is *prescribed* by the norm, as in punishment and retaliation. Sometimes this action is *proscribed* by the norm, as in cheating or cooperation. The model predicts that we should observe more norm violations after experiencing a shock. We design three experiments to assess this prediction. The critical design choice is to manipulate NES by inducing significant losses (80%) on the earnings from a Real Effort Task (Bogliacino and Montealegre, 2020). After this initial stage, participants interact in one (or multiple) tasks, where we measure the change in norm compliance. The settings include stealing, cheating, Joy of Destruction (JoD), and cooperation.

Since the predictions are conditional on social norms, we elicit the normative expectations that hold for each situation using Bicchieri and Xiao (2009)'s methodology. Participants provide their personal normative beliefs (PNBs) over the action space of the decision-maker and then are asked to guess the modal response to the PNBs (under a simple incentive scheme). Participants did not make an actual choice in these settings, to make sure that we elicit social norms separately from behavior (Krupka and Weber, 2013).

As predicted, subjects steal more and cheat more after suffering an NES. The increase in stealing is almost one-fourth of a standard deviation (calculated on the outcome variable in the control). In the die-under-the-cup task (Fischbacher and Föllmi-Heusi, 2013), where participants are paid according to the number that they *report* from the throw of a dice, they are 14% more likely to report four and five, the number with the highest payoff. The effect is equivalent to more than one-fourth of a sd. When we look at the JoD, the *decrease* in retaliation is as large as 50% of a sd, again supporting the prediction of the model. In the prisoner's dilemma, an NES increases defection without reaching significance at the conventional levels.

The main results match several stylized facts. The fact that NES may generate antisocial behaviour, particularly crimes against property, has not gone unnoticed. Compelling quasi-experimental evidence document a positive relationship between negative economic shocks and antisocial behaviour. For example, Dube and Vargas (2013) use the change in coffee prices to study variations in crime in communities that are highly dependent on income from the harvest. Cortés et al. (2016) use the collapse of the Ponzi scheme in Colombia to detect variation in a portfolio of criminal activities. Bignon et al. (2017) exploit the regional variation in the exposure to phylloxera in wine-producing regions in France to identify the increase in property crime. Dix-Carneiro et al. (2018) use the trade liberalization shock in Brazil to estimate the causal impact of the shock on criminal activity. Weather shocks have also led to an increase in property crimes (Mehlum et al., 2006). Cheating has been less studied. Aksoy and Palma (2019) look at cheating under "scarcity" - the shock around paycheck variation - but could not detect any significant variation. Bogliacino and Montealegre (2020) also look at the effect of NES in the die-under-the-cup task, finding no effect, but the presence of four tasks may have diluted the incentives. In a recent paper, the manipulation of NES correlates with disproportionate predatory behaviour (Blanco et al., 2021). Since the authors manipulate shocks and criteria of assignment of social status, their evidence suggests that circumstances favour antisocial instincts. Additionally, Boonmanunt et al. (2020) document that people under scarcity are less responsive to a social norm intervention, in line with our main argument.

Our theoretical framework also outperformed other theories. In the JoD, the norm of retaliation generates a trade off between compliance and income (money burning is costly). Although grounded on the same reasoning as in stealing and cheating, the model predicts *less* antisocial activity, and is consistent with our controlled evidence. Theories of crime like strain theory (Merton, 1938) cannot predict this finding. Strain theory states that the frustration caused by NES should increase all antisocial behavior (stealing *and* money burning). Psychological theories of NES cannot outperform our results either: NES induce cognitive load (Mani et al., 2013; Bogliacino and Montealegre, 2020) but there is no consensus on the relationship between cognitive load and social preferences (Alós-Ferrer and Garagnani, 2020). Similarly, the increase of risk aversion cannot account for these results (Haushofer and Fehr, 2014).

To document the robustness of our argument on shocks and norm compliance, we conducted a further experiment, where (a) we imposed a rule instead of relying on an elicited norm, (b) we separately controlled for the wealth effect, and (c) we used fully unexpected shocks. We found that the effect of an NES over norm compliance is robust and distinguishable from a pure wealth effect.

We now move to present the theoretical predictions and the experimental evidence. Formal proofs are reported in an Appendix and the experimental protocols are available in the Supplementary Online Materials (SOM).

2 The model

In this section, we study the problem of a decision-maker (DM) facing a binary choice involving a social norm. We derive a set of predictions on the effect of an NES - modelled as a reduction in the DM's asset position - in a series of standard experimental tasks. The DM derives utility from income, including assets and the monetary payoff from her choices, but has a conditional preference to follow social norms. Acting in violation of a norm results in a psychological cost. DMs are indexed by their norm propensity θ, to capture heterogeneity in norm compliance. The preferences are similar to Krupka and Weber (2013), Kimbrough and Vostroknutov (2018), and Levitt and List (2007). Models of preferences with social image have a similar framework, but the social image is endogenous (Andreoni and Bernheim, 2009; Benabou and Tirole, 2006).

2.1 The optimal choice

Formally, a DM (i) should choose $d_i \in \{0,1\}$. If the setting is strategic, she will be interacting with j. By convention, $d = 1$ is the harmful action, defined as the action that causes a loss to the counterpart or prevents her from enjoying a gain. Preferences include two terms. The first is the utility of income: an additive separable utility function $u(e + w(d_i, d_j))$, where $w(\cdot)$ is the monetary payoff, and e is the initial endowment. The second term is $\mathbb{1}_{d_i \neq n} c(\theta_i)$, the psychological cost of deviating from a social norm n. The cost increases in θ_i, the propensity to comply. We have $\theta_i \in [0,1]$, with Cumulative Density Function $F(\cdot)$.

The problem can be written as follows

$$\max_{d_i \in \{0,1\}} u(e + w(d_i, d_j)) - \mathbb{1}_{d_i \neq n} c(\theta_i) \tag{1}$$

In some situations, like stealing, $d = 1$ transgresses a social norm (i.e. $n = 0$), in others, like punishment, it is prescribed by the norm ($n = 1$). If the norm is conditional, as in tit-for-tat, we will use the notation $n = d_j$.

An NES is modelled as $de < 0$.

The following assumptions hold:

Assumption 1. $u(\cdot) : \mathbb{R} \to \mathbb{R}$

$u'(\cdot) > 0$

$u''(\cdot) < 0$

Assumption 2. $c(\cdot) : [0, 1] \to \mathbb{R}$

$c'(\theta) > 0, \ c''(\theta) > 0$

Assumption 1 is the standard decreasing marginal utility of income. Assumption 2 formalizes the utility cost of norm violation and the dependence on the psychological parameter θ.

To understand the logic of the argument, consider a non strategic choice where a fairness norm is in place ($n = 0$) and $d = 1$ is a transgression. An agent of parameter θ chooses $d = 1$ if $u(e + w(1)) - u(e + w(0)) \geqslant c(\theta)$. The term $u(e + w(1)) - u(e + w(0))$ captures the benefit B of transgressing the norm, constant across agents. The cost is increasing in θ. In Figure 1, top panel, we plot the optimal choice as a function of θ: there is a threshold $\bar{\theta} = \theta_1$ below which agents will transgress, and above which they will comply.

What happens when a DM suffers an NES? Due to the concavity of the utility function, the marginal utility of transgression increases, leading to more norm violations. In the top panel of Figure 1, for the new benefit curve, more people choose to carry out $d = 1$, i.e. $\bar{\theta}$ moves to the right, from θ_1 to θ_2.

Consider also the opposite situation where $d = 1$ is costly and recommended by the norm (i.e. $n = 1$). An agent of parameter θ chooses $d = 1$ if: $u(e + w(1)) - u(e + w(0)) \geqslant -c(\theta)$. The left-hand side is the utility loss from punishment, and the right-hand side is the utility cost of norm violation. In presence of an NES, concavity implies that $\frac{\partial u(e+w(1)) - u(e+w(0))}{\partial e} > 0$, the utility loss from following the norm increases and less people will choose $d = 1$. This is illustrated in the bottom panel of Figure 1.

In settings with interaction, we need to introduce strategic uncertainty: the DM will now maximizes $E[u(e+w(d_i, d_j)) - \mathbb{1}_{d_i \neq n} c(\theta)]$. Define p to be the expected likelihood that d_j chooses 1. There are three cases, either $n = 0$, $n = 1$, or $n = p$ (tit-for-tat). We can write the expression in a compact form as $p(u(e+w(1,1))-u(e+w(0,1)))+(1-p)(u(e+w(1,0))-u(e+w(0,0)))) \geq (1-2n)c(\theta)$.

Consider when the norm is tit-for-tat. The DM chooses 1 if $p(u(e+w(1,1))-u(e+w(0,1)))+(1-p)(u(e+w(1,0))-u(e+w(0,0)))) \geq (1-2p)c(\theta)$. There are three terms: $u(e+w(1,1))-u(e+w(0,1))$ is the utility loss from retaliation, $u(e+w(1,0))-u(e+w(0,0))$ is the benefit of defection, and $(1-2p)c(\theta)$ is the (expected) psychological cost.

We will derive our predictions in the two extreme cases, $p=0$ and $p=1$. These predictions are testable once beliefs are elicited in an experiment. Conditioning on a degenerate belief also represents a plausible description of decision-making in one-shot interactions. In the Appendix, we will show that the conclusions are supported in equilibrium by a formal comparative statics result.

Under $p=1$, there is a cost of retaliation if $u(e+w(0,1))-u(e+w(1,1))>0$. When the latter condition holds, the DM chooses $d=1$ only if the cost of transgression is larger than the cost of retaliation. Since an NES raises the cost of retaliation, the share of DMs who chooses $d=1$ decreases. If retaliation is not costly, everybody will make the same choice, regardless of the shock.

Under $p=0$, there is a benefit from defection if $u(e+w(1,0))-u(e+w(0,0))>0$. The optimal choice is determined by whether $u(e+w(1,0))-u(e+w(0,0)) \geq c(\theta)$. Since an NES increases the benefit from defection, the share of DM who chooses $d=1$ increases. If defection is not profitable, everybody will comply, regardless of the shock.

The reasoning for $n=0$ and $n=1$ are special cases of the tit-for-tat.

2.2 Settings

We will consider four settings: cheating, stealing, Joy of Destruction (JoD), and cooperation in prisoner's dilemma (PD).

In the cheating and stealing task, the payoffs for the DM are $w(1)>w(0)$ and the norm is $n=0$.

The JoD is a simultaneous interaction where $d=1$ is costly and harmful. $d=1$ is called money

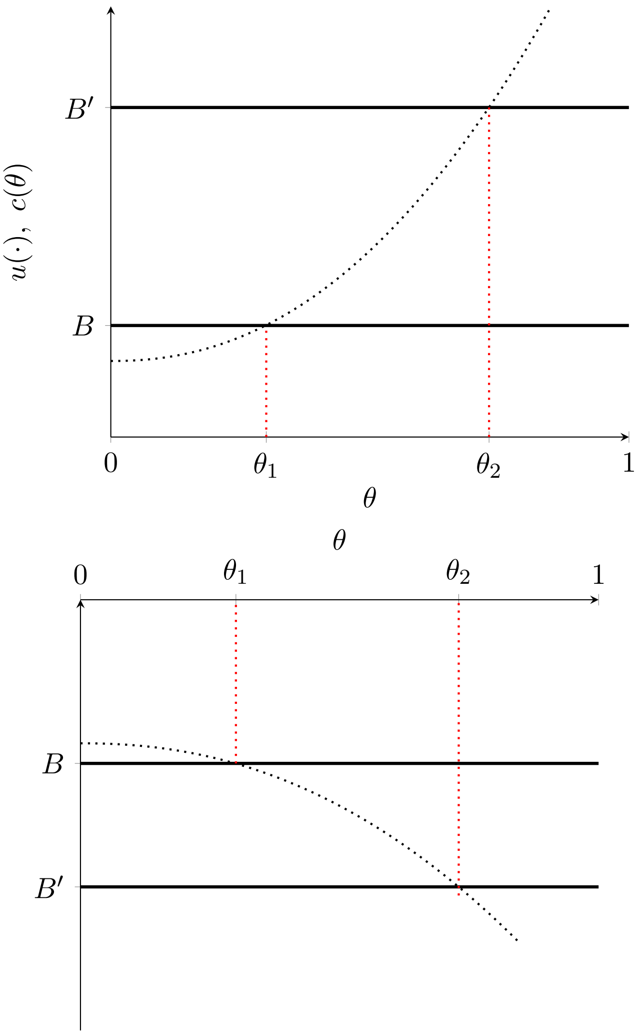

Figure 1: The optimal choice

Setting	Social Norm	Prediction
Stealing	Do Not Steal	$S(NES) > S(C)$
Cheating	Do Not Cheat	$C(NES) > C(C)$
Joy of Destruction	Retaliation	$D(NES) < D(C) \| P = 1$
PD	Tit-for-Tat	$C(NES) < C(C) \| P(C) = 1$

Table 1: Theoretical predictions from the norm compliance model.

burning. In the standard calibration (Abbink and Herrmann, 2011), the initial endowment is 10, the cost of burning is 1 and the damage inflicted is 5. More generally, it must hold that $w(0,0) > w(1,0) > w(0,1) > w(1,1)$. The social norm is to *retaliate* (Abbink and Herrmann, 2011).

The prisoner's dilemma is a symmetric simultaneous game where $w(1,0) > w(0,0) > w(1,1) > w(0,1)$. We assume that the relevant social norm is conditional cooperation (Gachter, 2007).

2.3 Theoretical Predictions

As discussed in Section 2.1, when there is a trade off between income and norm compliance, an NES makes people more attentive to income leading to more transgression. For a trade off to exist, following the norm should be costly in terms of payoff. This is the case for cheating, stealing and trustworthiness, where the cost of following the norm is the loss of income from $d = 1$.

For the JoD, the trade off exists when a DM expects the counterpart to burn, because burning is costly but the retaliation is prescribed by the norm. The prisoners' dilemma has a clear trade-offs: under tit-for-tat, conditional cooperation is costly because defection is profitable. An NES generates more norm violation.

The predictions are summarized in Table 1 below. A formal discussion is in the AppendixA.

2.4 Equilibrium and Comparative Statics: general results

Table 1 presents the predictions under $p = 0$ or $p = 1$. These are testable given the elicitation procedure used in the lab experiments and plausible as a description of how a DM interacts in a one shot decision. It is also a formal equilibrium prediction if θ_j belongs to i's information set.

Alternatively, Assumption 3 states that the distribution $F(\cdot)$ of the norm propensity parameter is common knowledge. We can show that the direction of the effect of the shock is maintained.

Assumption 3. *$F(\theta)$ is common knowledge.*

This is the definition of equilibrium:

Definition 1. *Given a symmetric simultaneous 2X2 game, with preferences $u(e + w(d_i, d_j)) - \mathbb{1}_{d_i \neq n} c(\theta_i)$, with randomly drawn players i, j, finite payoffs functions $w(d_i, d_j)$, an equilibrium with social norm n is a distribution of choices for the population such that each DM maximizes her utility and expectations are mutually consistent.*

We apply the refinement that the equilibrium be stable. Here, stability means that small perturbations induce incentives that drives behavior towards equilibrium.

The following proposition holds (the proof is in the Appendix 1).

Proposition 1. *Under assumptions 1, 2 and 3, the following comparative statics hold in equilibrium: a) in the JoD, $\frac{\partial P(d=1)}{\partial e} > 0$; a) in the PD, $\frac{\partial P(d=1)}{\partial e} < 0$.*

2.5 Extensions

We used concavity to prove the result. Under Assumption 1, an NES is just a wealth effect. The concavity of the utility function (i.e. risk aversion) is undisputed (Camerer, 1995; Starmer, 2000). However, if it is just concavity that drives the norm compliance effect, then also poverty would increase transgression. A similar proposition is not empirically testable (rich and poor differ across multiple dimensions), but would be coherent with the dominant interpretation of shocks as a plausible variation to study the causal effect of poverty (Mani et al., 2013; Haushofer and Fehr, 2014; Boonmanunt et al., 2020).

Loss aversion (Kahneman and Tversky, 1979) generates a norm transgression effect for an NES, separately from a wealth effect. When we condition on the belief, the problem of the DM can be reduced to one of the two cases where $d = 1$ is either costly but recommended or profitable but forbidden. As a result, we can prove the general argument without strategic interaction. Assume *a fortiori* that the utility function is linear (risk neutrality) but with loss aversion. The problem of the DM becomes:

$$\max_{d \in \{0,\ 1\}} e' + w(d) - v^l(\max\{0, e - e' - w(d)\}) - \mathbb{1}_{d_i \neq n} c(\theta) \tag{2}$$

with $v^l(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ and increasing, and e' is the current endowment, either equal to e, in the control, or lower than e in case of NES. In the formulation of the v^l function, e is the reference point.

In the control, the DM chooses 1 if $w(1) - w(0) \geq (1 - 2n)c(\theta)$, in presence of a (large enough) shock, and defining $\Delta e = e - e'$, if $w(1) - w(0) + v^l(\Delta e - w(0)) - v^l(\Delta e - w(1)) \geq (1 - 2n)c(\theta)$. This will lead to the same predictions as in Table 1, but without reducing an NES to a wealth effect. For instance, a positive shock would be void of consequences in this case, whereas under concavity the effect of shock would be symmetrical.

3 Eliciting social norms

The predictions are conditional on social norms. Following the definition by Bicchieri (2006), social norms should be supported by normative expectations. Normative expectations are second-order beliefs: what one expects others think should be done in a given contingency.

There are two main methods to elicit normative expectations: the coordination game by Krupka and Weber (2013) and the two steps elicitation method by Bicchieri and Xiao (2009). The former asks participants to rate the actions available to the DM in terms of moral appropriateness but pays them to match the modal response. As in any coordination game, salience drives participants' strategic choices (Mehta et al., 1994), and shared beliefs associated with norms become salient.

The two steps elicitation method by Bicchieri and Xiao (2009) recovers first-order and second-order belief. Subjects report their Personal Normative Beliefs (PNBs) for the action sets available to the DM, usually as a singleton. Then,they are paid to guess the response to the PNBs questions. As discussed and analyzed in Aycinena et al. (2021), KW and BX methods elicit the intensive and extensive margin respectively. Given our interest in carrying out an action, more than its intensity, we rely on the BX method.

We send an online invitation to a sample drawn from the subject pool at the Unbiased Lab (Universidad Nacional) to fill in an online incentivized survey (it is available in SOM, Section I). Data were gathered in February 2021.

Participants went through two parts. Part A elicited PNBs over the action space for the DM in each prosocial and antisocial task used in this article. The PNB is the "personal opinion on what is the appropriate and morally correct action of Individual A, selecting one of the following options". Each question included a description, the sample size, and the pool of participants. The sequence of questions came in random order. In total, participants evaluated six decisions, three antisocial and three prosocial. The decisions include the trust and trustworthiness choices in a trust game, which we will discuss in the second part of this paper.

After stating their PNBs, in part B, participants were asked to predict the modal action among the respondents in the original experiment (empirical expectations) and the modal response to the PNBs questions among the respondents in the current experiment (normative expectations). Empirical expectations are collected for completeness. In each question, the order of the available options was randomized. Each participant made twelve predictions, and one was randomly selected for payment at the end. A correct guess was paid 25000 COP. The show-up fee was 10000 COP. The average time of completion was 35 mins. We collected 109 observations. On average, participants earned 21000 COP (6 USD). Participants did not make decisions in the settings, and they did not participate to the experiments. This is to ensure elicitation of normative expectations separately from behavior (Krupka and Weber, 2013).[2]

These were the action sets of the decision-maker in each setting described to the participants. For the stealing task, the decisions included stealing and not stealing. For the die-under-the-cup task, the action set included truthful reporting, reporting the first three numbers unconditionally, reporting four or five unconditionally, reporting six unconditionally, misreporting the drawn number plus or minus a maximum of two to own advantage, misreporting the drawn number plus or minus a maximum of two to own disadvantage. For the JoD, the possible actions were burning unconditionally, abstaining unconditionally, choosing the same action as the counterpart (tit-for-tat), and choosing the opposite action of the counterpart. For the trust (cooperation)

[2] To avoid deception, we describe real experiments carried out in our lab. As a result, for the trust game, we use a dichotomous trust game whose data we do not report here and for the prisoner dilemma, we use as reference the experiment by Bogliacino et al. (2020).

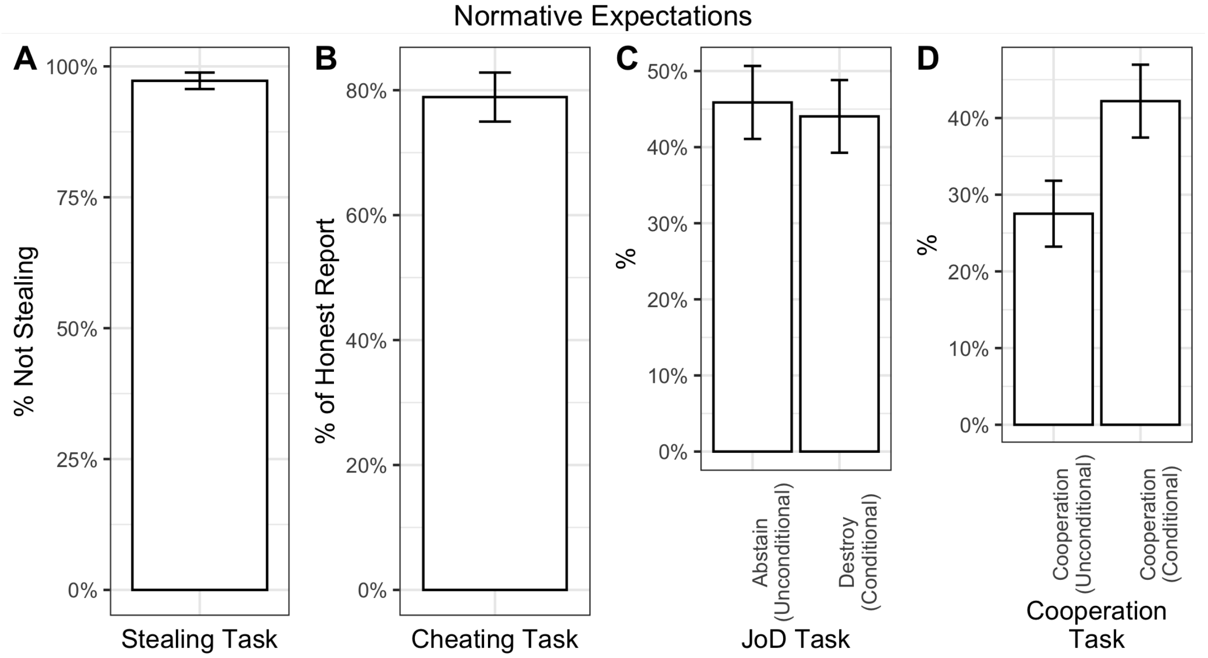

Figure 2: The elicited normative expectations

game, similarly to the JoD, the possible actions were trusting (cooperating) unconditionally, keeping (defecting) unconditionally, choosing the same action as the counterpart, and choosing the opposite action of the counterpart. For trustworthiness, the two available actions were sharing or keeping. In all cases, we use the same framing used in the original experiment to avoid furthering experimenter demand.

We show the elicited normative expectations in Figure 2. For the stealing task (Panel A), Do Not Steal was the predicted PNB by 97.25% of the participants. In panel B, truth-telling was predicted as the modal response to the PNB question for the cheating task by 78.90% of participants. In Panel C, for the case of JoD, the two modal normative expectations are non burning unconditionally (45.87%) and tit-for-tat (44.04%).[3] In Panel D, the modal normative expectation of cooperation is tit-for-tat (42.20%). The social norms for trust and trustworthiness will be discussed in Section 8.

Once illustrated the social norms that apply to the settings, we move to the assessment of the predictions. We will now present three different experiments.

[3] This suggests that for almost half the participants, our prediction is valid. Moreover, when the social norm is to abstain, shocks do not affect behavior (as we should expect zero burning) thus the general prediction is unaffected.

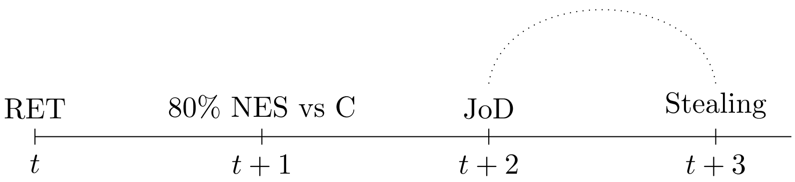

Figure 3: The timeline of Experiment I

4 Experiment I

4.1 Experimental Design and Procedures

Experiment I is a standard between-subject design, with a treatment and a control condition. In the treatment condition, participants suffered an NES. The NES was an 80% loss on the accumulated earning from a Real Effort Task (RET), experienced with a 50% probability. The probability was common information. The RET was the Niederle and Vesterlund (2007)'s task of summing sequences of two-digit numbers and took place over 4 minutes. The assignment to the experimental conditions occurred at the individual level, within each session.

After the treatment, the participants played the stealing task and the JoD (Abbink and Herrmann 2011) in random order. In the JoD, participants can burn half of the endowment of the counterpart at their own cost. The decision is simultaneous. The initial endowment is 10 ECUs and the cost of burning is 1 ECU. To avoid a positive endowment shock after the NES, the assignment of the 10 ECU preceded the RET. We elicit beliefs on whether the counterpart was affected by shock and whether the counterpart was going to burn. To reduce the likelihood of hedging (Blanco et al., 2010), incentives for beliefs were smaller (1 ECU if correct). In the Stealing task, participants can appropriate 80% of the earnings from the RET, from a participant to another experiment occurring simultaneously.

This is how incentives were determined. Participants received the show up fee and the gain from the RET immediately after the session's end. The money from one randomly selected task among the other two (and the beliefs) was paid one week later.

We could not allow stealing within the session, as this was instrumental to manipulate intentional shocks in another experiment (Section VIII). Additionally, two antisocial tasks with counterparts within the same session could generate compensatory behavior. To avoid asym-

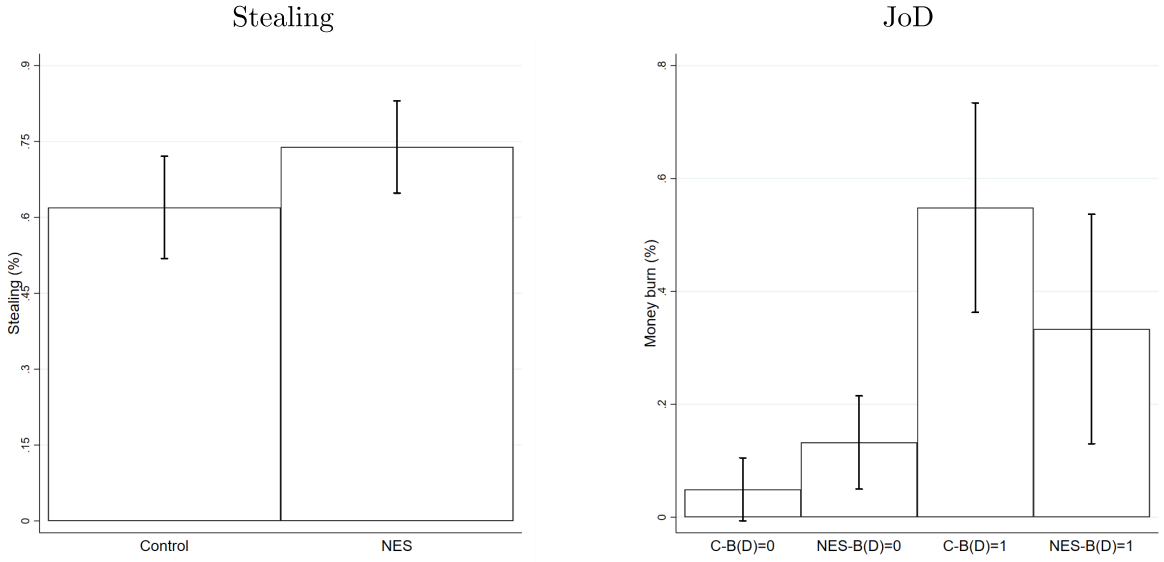

Figure 4: The impact of NES in the Stealing and JoD tasks

metry in the JoD and the Stealing incentives, we decided to pay both tasks with a delay.

The timeline of the experiment is reported in Figure 3. The procedures were as follows. After reading the general instructions aloud, we asked participants to follow the specific instructions on the computer screen for each task. Subjects could raise their hand at any time if they had any questions. A final questionnaire was handed out to the participants.

In total, we recruited 184 undergraduate students from the Unbiased subject pool. Invitations were randomized. Sessions took place in the lab, in presence, around October 2019.

Out of the 184 participants, 92 were in the NES condition and 92 in the control. The average session had 20 participants and there were nine sessions in total. The exchange rates were 1000 COP per ECU. On average each participant earned 17000 (±5200) COP (approximately USD 5). The experiment is programmed in oTree (Chen et al., 2016) and the English version of the protocol is available in the SOM, Section II.

4.2 Results

Participants, on average, solved 5.21 (±2.44) problems, and the performance is not different across experimental conditions ($\chi^2 = 9.18$, $p = 0.75$).

In Figure 4 (left panel), we report the average stealing rate by condition, with a 95% confidence interval. On average, stealing increases from 62% to 73.9% in presence of an NES. To assess the prediction, we run an OLS regression, controlling for order and compute a one sided test, since we are postulating a direction for the alternative hypothesis. Table 2, Column (1), reports the results. The effect is both economically relevant, around 25% of a standard deviation of the outcome in the control condition, and statistically significant ($F(1, 181) = 3.07$, $p = 0.04$ one-sided).

Table 2: OLS estimates of effect of NES on Stealing and JoD

	(1) Stealing	**(2) JoD**
NES	0.063*	
	(0.036)	
Order	0.113	0.073
	(0.071)	(0.052)
NES-B(D)=0		0.080
		(0.049)
C-B(D)=1		0.486***
		(0.096)
NES-B(D)=1		0.277***
		(0.102)
Constant	0.544***	0.011
	(0.066)	(0.037)
N	184	184
Test of Prediction	0.04	0.06

Robust standard errors shown in parenthesis. * p<0.10, ** p<0.05, *** p<0.01

The behavior in the JoD is shown in Figure 4 (right panel). Participants burn 4.91% of the time when they believe that the counterpart will not burn, and 54.38% when they expect others to do it. This stylized fact documents the social norm of retaliation plotted in Figure 2. However, under the shock, the likelihood to retaliate is only 33.33%, i.e. there is a 21.50% reduction, which is both economically relevant (51.85% of a sd computed for the control condition) and statistically significant ($F(1, 179) = 2.44$, $p = 0.05$ one-sided).

The supporting regression is reported in Table 2, Column (2). We run an OLS regressions with three dummies: NES and $B(D) = 0$, control and $B(D) = 1$, and NES and $B(D) = 1$. The omitted category is control and $B(D) = 0$. In the last row, we report the p-value of the main test.

To summarize, as predicted, we detected a positive effect of NES on stealing and a negative effect of NES on retaliation.

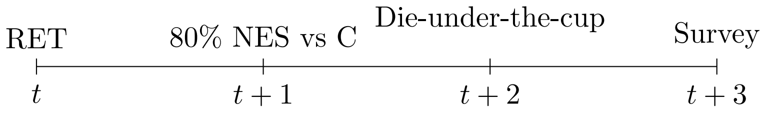

Figure 5: The timeline of Experiment II

5 Experiment II

5.1 Experimental Design and Procedures

Experiment II is a between-subject design, with two conditions and with the treatment assigned at the individual level. In the treatment, we manipulated a random shock of 80% of the accumulated earnings from a RET, with a 50% chance. Since this is an online experiment, because of the Covid-19 pandemic, we did not use the same task as in Experiment I, as we could not prevent participants from using a calculator. Instead, we chose a 4 minutes transcription task. The language used was Tagalog (the text was the Theory of Moral Sentiments; Smith 1759). We avoid more common languages to ensure that performance depends on effort and not on accumulated knowledge. Each fragment was 35 characters long. The software did not allow copying and pasting.

After the RET and the assignment to the experimental conditions, the main task was a "cheating game" based on Fischbacher and Föllmi-Heusi (2013)'s die-under-the-cup. In this task, participants rolled a dice privately and reported their results. Participants had access to an online dice, beyond the experimenters' control, but they could use any available dice. The payoff was calculated as 2000 COP times the reported number (from one to five) and zero for a reported six. After the second task, participants had to answer some demographic questions.

This experiment was conducted online. We sent random invitations to a sample from the Unbiased subject pool, excluding those that took part in previous experiments with NES. We sent out a link for participation, with included instructions. The timeline is depicted in Figure 5.

Participants received a show up fee, the earnings from the RET and the dice. In total, we recruited 158 participants. Data collection occurred in June 2020, on average each participant earned around 25000 COP (around 7 USD).

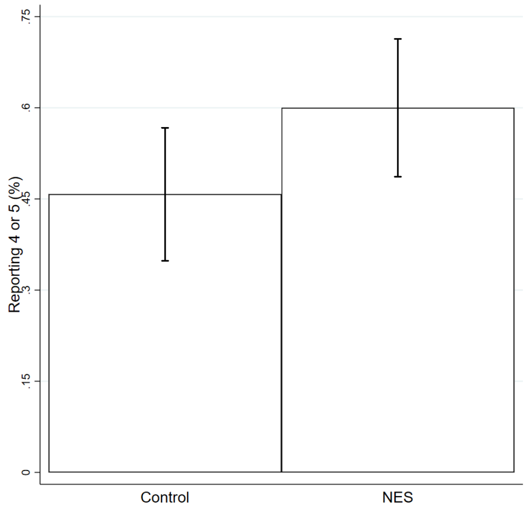

Figure 6: The impact of NES in the die-under-the-cup task

The experiments is programmed with oTree (Chen et al., 2016). The experimental protocol can be found in the SOM, Section III.

5.2 Results

On average, participants tried 11.91 ± 3.78 transcriptions, completing successfully 9.17 ± 4.15 of them. There is no difference between treatment and control ($\chi^2 = 23.13$, $p = 0.23$ and $\chi^2 = 26.20$, $p = 0.19$ respectively).

There was a considerable amount of cheating: we neatly reject the null hypothesis that the observed data comes from a fair dice ($\chi^2 = 37.11$, $p < 0.001$).

Since we do not observe the original draws, we cannot test for cheating directly, but we can measure how the likelihood of reporting the numbers with the highest payoff (four or five) differs between treatment and control, as in Bogliacino and Montealegre (2020).

Figure 6 shows the mean outcome, broken down by experimental condition. In the control, the likelihood of reporting 4 or 5 is 45.78%. It increases to 60% in presence of a NES. The difference is as large as 28.36% of a sd of the outcome in the control condition and is statistically significant ($t = -1.79$, $p = 0.03$ one sided, controlling for unequal variance).

1/2	C	NC
C	8, 8	0, 10
NC	10, 0	4, 4

Table 3: Experiment III: the stage game

6 Experiment III

6.1 Experimental Design and Procedures

In Experiment III, we test the effect of shocks on cooperation. This is an online experiment with a treatment and control between-subject design. In the first part, participants performed a RET (transcription task as in Experiment II). After knowing their performance, they either kept all the money (control) or suffered an 80% loss (NES). To avoid deception but allows for a surprise effect, in explaining the incentives for the RET, we warned that in the second part of the task, the total payment could change. The surprise effect is the first design change with respect to the previous two experiments.

Table 3 show the normal form of the PD. The strategies C and NC were labelled green and blue. The second innovation with respect to the standard design consisted in the introduction of the this procedure: a) participants declared their strategy conditional on the belief that the counterpart was playing C and NC; b) then they declare which scenario was more likely among the counterpart playing C, NC, or choosing randomly. Subjects knew that the answer was incentive compatible. Eliciting the belief contingent decisions represents the cleanest test of our prediction, and overcomes the endogeneity of the belief.

Since we asked four comprehension questions with feedback, we excluded those that made more than one mistake. The total number of observations is 297, of which 146 are in the control and 151 in the NES condition. Data collection took place in March-April 2022 (PreAnalysis Plan recorded as AsPredicted #89448).

We recruited participants via the REBEL Lab at Universidad del Rosario. Participants are invite using ORSEE (Greiner, 2015). The protocol is programmed in oTree Chen et al. 2016 and can be found in the SOM-Section IV.

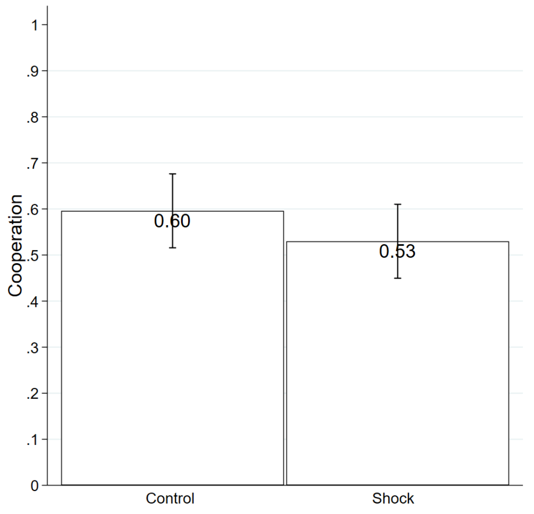

Figure 7: The impact of NES in the Prisoners' Dilemma

6.2 Results

On average, participants successfully completed 8.88±4.02 transcriptions. There is no difference between treatment and control ($\chi^2 = 25.18$, $p = 0.12$).

These are the main results. Under the belief that the counterpart would not cooperate, participants chose cooperation 9.58% of the time. The shock marginally decreased cooperation to 9.27%. The difference was not statistically significant ($F(1, 295) = 0.01$, $p = 0.46$ one sided). Coming to the main outcome of interest, under the belief that the counterpart would cooperate, participants chose cooperation 59.58% of the time. The shock increased defection and lowered cooperation to 52.98%. The difference is not statistically significant ($F(1, 295) = 1.32$, $p = 0.12$ one sided) and represents around 15% of a sd in the control.

To wrap up, there is a marginal reduction in cooperation due to NES, but the effect is not statistically significant at the conventional level.

7 Experiment IV: Evidence from Rule Following

Across three different experiments, NES induce more norm violations. The evidence is robust for anti-social tasks and less so for the prisoner's dilemma. To provide a more conclusive test, we designed an additional experiment. This experiment improves on the previous design in three ways: a) it separately controls for a wealth effect, b) it induces the rule to follow instead of relying on elicited social norm, c) it allows for an intensive margin to increase statistical power.

The main task is the Rule Following Task (Kimbrough and Vostroknutov, 2018). Participants faced a RET under four possible conditions: (1) each correct transcription is paid one point, the potential 80% loss is announced but not administered (Control); (2) each correct transcription is paid one point, the potential shock is announced and administered (Shock treatment); (3) each correct transcription is paid one point (High treatment) and there is no exposure to shock; (4) each correct transcription is paid 0.2 points (Low treatment) and there is no exposure to shock. Treatments (1) and (2) followed the procedures in Experiment III: the first part was described as having two phases, where the payment could change between phases one and two. In the second part, participants should guide a stick man across a path with five traffic lights. Endowed with 30 seconds, each second worth one point, they were paid for the seconds left when they reached the end of the path. The rule announced to participants but not enforced was to wait for the green at each traffic light. It took five seconds for a traffic light to change from red to green.

This is a between-subject incentivized survey. Random invitations were sent to those participants in our database that did not participate in any other experiment with shocks. The analysis was pre-registered (Aspredicted #80601).[4] We paid both tasks and a show-up fee.

These are the results. The performance in the RET was different across treatments, as expected given the differences in incentives ($\chi^2 = 74.81$, $p = 0.09$). The main outcome variable was the number of seconds spent crossing the path. On average, subjects spent 5.92 seconds (4.80 sd). In the control, the outcome was 7.51 (±6.80). In treatment (2) (the NES condition in Experiment I-III) the outcome was 4.95 (±2.73). A (one sided) t-test controlling for unequal variance returns $t = 2.46$ (one sided $p < 0.01$). In treatment (3), the outcome was 5.89 (±4.34), whereas in treatment (4) it was 5.43 (±4.28). A t-test controlling for unequal variance returns $t = 0.55$ (one sided $p = 0.29$). In other words, the wealth effect did not generate the same result as the shock. Results are plotted in Figure 8.

We run an OLS regression using the seconds as the outcome variable, with robust standard

[4] Notice that we expected to reach 300 participants, but due to the end of the term and the fact that we could not use participants from other experiments we had to close the data collection at 210 participants. However the variability is much lower than expected.

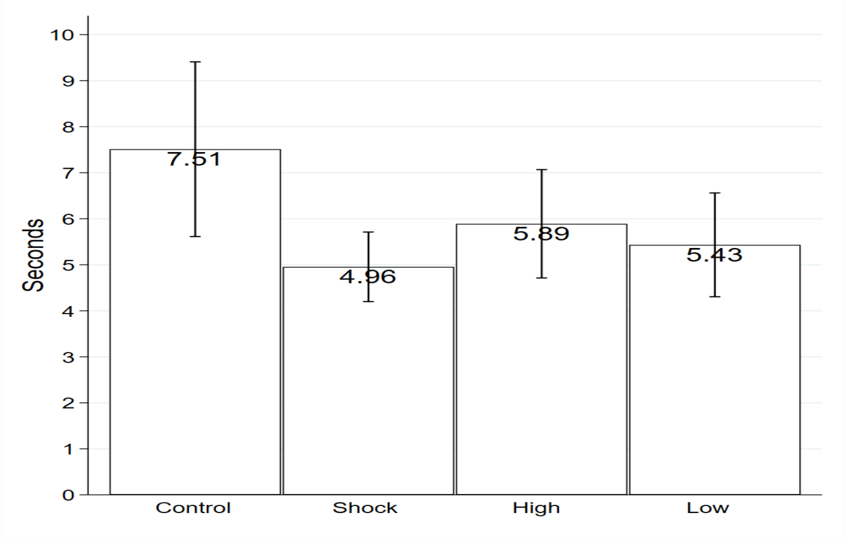

Figure 8: Results from the rule following task

errors. The difference between the shock effect (dummy for treatment 2) and the wealth effect (the difference in performance between treatment 3 and 4) is 2.94 seconds and is statistically significant ($t = -2.27$, $p = 0.02$).

8 Concluding remarks

This article shows how experiencing wealth losses can be a source of norm violation. Although studies on crime against property have documented this objective fact, the literature failed to grasp the underlying mechanism and its implications for our theories of strategic behavior.

Since norms are scripts that humans partially incorporate into their preferences (Gintis, 2007), it is not surprising that people manipulate or elude norms if allowed to do so (Bicchieri et al., 2021; Andreoni and Bernheim, 2009; Bicchieri, 2010). Dictator games are widely used in this literature to avoid confounds from strategic beliefs. Dana et al. (2007) introduce the concept of *moral wiggle room* to explain why when settings change, but the action space does not, subjects behave more egoistically. List (2007) documents a sizable behavioral change following minimal variation in the action space. Instead of relying on contextual changes, we provide evidence from indirect incentive effects.

The literature on shocks is now rapidly expanding. In experimental settings, the manipulation of losses or windfalls has been used to study poverty or scarcity, usually exploiting paycheck variation or natural experiments. This literature focused on the cognitive impact: Mani et al. (2013) found a negative effect in sugarcane farmers in India [5] while Bogliacino and Montealegre (2020) found a negative effect of NES on cognitive performance in the lab. Haushofer and Fehr (2014) claim that suffering NES (and in general, poverty) increases stress, which induces lower risk propensity (in the gain domain) and higher present bias, further worsening the cognitive performance in decision tasks. The impact on social norms has been overlooked, though, although.

This article has implications for the ongoing discussion on global challenges such as pandemics, global warming, and war threats. Theories of cultural evolution suggest that the western culture evolved through a peculiar *W.E.I.R.D.* [6] psychology and social norms, forged in response to the Marriage and Family Program of the Church (Heinrich, 2020). If vulnerability to shocks can drive the abandonment of norms and the establishment of new rules of behavior, the heterogeneity of such vulnerability has important implications for the evolution of cultural norms. But, as we stated in the final part, shocks should be understood according to their source and (possibly) individual versus collective nature. This article is but a beginning of this theory.

Finally, this article has implications for the discussion around welfare state protection or liberalization in general. Market *disciplines*, but downward adjustment of price and earnings may lead to unintended consequences (Dix-Carneiro et al., 2018). Recent evidence from the UK confirmed that this may be the case (d'Este and Harvey, 2022).

[5] Carvalho et al. (2016) found no effect, but paycheck variations are temporary, expected and expected to be temporary.

[6] Acronym of Western, Educated and from an Industrialized, Rich, and Developed country. It has been coined by Heinrich et al. (2010).

References

Abbink, Klaus and Benedikt Herrmann, "The moral costs of nastiness," *Economic Inquiry*, apr 2011, *49* (2), 631–633.

Aksoy, Billur and Marco A. Palma, "The effects of scarcity on cheating and in-group favoritism," *Journal of Economic Behavior and Organization*, 2019, *165*, 100–117.

Alós-Ferrer, Carlos and Michele Garagnani, "The cognitive foundations of cooperation," *Journal of Economic Behavior and Organization*, 2020, *175*, 71–85.

Andreoni, James and B Douglas Bernheim, "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, sep 2009, *77* (5), 1607–1636.

Aycinena, Diego, Francesco Bogliacino, and Erik Kimbrough, "Measuring Norms: Assessing the Bicchieri-Xiao method," *Working Paper*, 2021, (August).

Bauer, Michal, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts, "Can War Foster Cooperation?," *Journal of Economic Perspectives*, aug 2016, *30* (3), 249–274.

Benabou, R and J Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

Bicchieri, Cristina, *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge University Press, 2006.

_ , "Norms, preferences, and conditional behavior," *Politics, Philosophy & Economics*, 2010, *9* (3), 297–313.

_ **and Erte Xiao**, "Do the Right Thing: But Only if Others Do So," *Journal of Behavioral Decision Making*, 2009, *22* (October 2008), 191–208.

_ **, Eugen Dimant, and Erte Xiao**, "Deviant or wrong? The effects of norm information on the efficacy of punishment," *Journal of Economic Behavior & Organization*, 2021, *188*, 209–235.

Bignon, Vincent, Eve Caroli, and Roberto Galbiati, "Stealing to Survive? Crime and Income Shocks in Nineteenth Century France," *Economic Journal*, feb 2017, *127* (599), 19–49.

Blanco, Mariana, D Houser, and J Vargas, "How to make a criminal," *mimeo*, 2021.

_ **, Dirk Engelmann, Alexander K. Koch, and Hans Theo Normann**, "Belief elicitation in experiments: Is there a hedging problem?," *Experimental Economics*, 2010, *13* (4), 412–438.

Bogliacino, Francesco and Felipe Montealegre, "Do negative economic shocks affect cognitive function, adherence to social norms and loss aversion?," *Journal of the Economic Science Association*, jun 2020, *6* (1), 57–67.

_ **, Camilo E Gomez, and Gianluca Grimalda**, "Crime-related Exposure to Violence and Social Preferences: Experimental Evidence from Bogota," Jul 2020.

Boonmanunt, Suparee, Agne Kajackaite, and Stephan Meier, "Does poverty negate the impact of social norms on cheating?," *Games and Economic Behavior*, 2020, *124*, 569–578.

Botchway, Ebo and Antonio Filippin, “Cooperation as Adaptation to Persistent Risk of Natural Disasters : Evidence from a Natural Experiment,” 2021.

Camerer, Colin, “Individual Decision Making,” in John H Kagel and Alvin E Roth, eds., *The Handbook of Experimental Economics*, Princeton University Press, nov 1995, pp. 587–704.

Carvalho, Leandro, Stephan Meier, and Stephanie Wand, “Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday,” *American Economic Review*, 2016, *42* (2), 407–420.

Cassar, Alessandra, Andrew Healy, and Carl von Kessler, “Trust, Risk, and Time Preferences After a Natural Disaster: Experimental Evidence from Thailand,” *World Development*, jun 2017, *94*, 90–105.

Chen, Daniel L., Martin Schonger, and Chris Wickens, “oTree-An open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance*, mar 2016, *9*, 88–97.

Cortés, Darwin, Julieth Santamaría, and Juan F Vargas, “Economic shocks and crime: Evidence from the crash of Ponzi schemes,” *Journal of Economic Behavior and Organization*, 2016, *131*, 263–275.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, “Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness,” *Economic Theory*, 2007, *33* (1), 67–80.

Dix-Carneiro, Rafael, Rodrigo R. Soares, and Gabriel Ulyssea, “Economic Shocks and Crime: Evidence from the Brazilian Trade Liberalization,” *American Economic Journal: Applied Economics*, oct 2018, *10* (4), 158–195.

Dube, Oeindrila and Juan F. Vargas, “Commodity price shocks and civil conflict: Evidence from Colombia,” *Review of Economic Studies*, oct 2013, *80* (4), 1384–1421.

d’Este, Rocco and Alex Harvey, “The Unintended Consequences of Welfare Reforms: Universal Credit, Financial Insecurity, and Crime,” *Journal of Law, Economics, and Organization*, 2022, *Online First*, 521–565.

Fischbacher, Urs and Franziska Föllmi-Heusi, “Lies in disguise-an experimental study on cheating,” *Journal of the European Economic Association*, 2013, *11* (3), 525–547.

Gachter, S., “Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications,” in BS Frey and A Stutzer, eds., *Economics and psychology: A promising new cross-disciplinary field*, MIT Press, 2007, pp. 19–50.

Gelfand, Michele J, Jesse R Harrington, and Joshua Conrad Josuah Conrad Jackson, “The Strength of Social Norms Across Human Groups,” *Perspectives on Psychological Science*, sep 2017, *12* (5), 800–809.

Gintis, Herbert, “A framework for the unification of the behavioral sciences,” *Behavioral and Brain Sciences*, 2007, *30* (1), 1–16.

Greiner, Ben, “Subject pool recruitment procedures: organizing experiments with ORSEE,” *Journal of the Economic Science Association*, 2015, *1* (1), 114–125.

Haushofer, Johannes and Ernst Fehr, “On the psychology of poverty,” *Science*, 2014, *344* (6186), 862–867.

Heinrich, Joseph, *The WEIRDest people in the World*, Farra, Straus and Giroux, 2020.

_ **, Steven J Heine, and Ara Norenzayan**, "Weird people," *Behavioral and Brain Sciences*, 2010, *33* (2-3), 61–135.

Kahneman, Daniel and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 1979, *47* (2), 293–298.

Kimbrough, Erik O. and Alexander Vostroknutov, "A portable method of eliciting respect for social norms," *Economics Letters*, jul 2018, *168*, 147–150.

Krupka, Erin L and Roberto A Weber, "Identifying social norms using coordination games: Why does dictator game sharing vary?," *Journal of the European Economic Association*, jun 2013, *11* (3), 495–524.

Levitt, Stephen J and John A. List, "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

List, John A, "On the Interpretation of Giving in Dictator Games," *Journal of Political Economy*, 2007, *115* (3), 482–493.

List, John A., Robert P. Berrens, Alok K. Bohara, and Joe Kerkvliet, "Examining the role of social isolation on stated preferences," *American Economic Review*, 2004, *94* (3), 741–752.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao, "Poverty impedes cognitive function," *Science*, 2013, *341* (6149), 976–980.

Mehlum, Halvor, Edward Miguel, and Ragnar Torvik, "Poverty and crime in 19th century Germany," *Journal of Urban Economics*, may 2006, *59* (3), 370–388.

Mehta, Judith, Chris Starmer, and Robert Sugden, "The Nature of Salience: An Experimental Investigation of Pure Coordination Games," *American Economic Review*, 1994, *84* (3), 658–673.

Merton, Robert K, "Social Structure and Anomie," *American Sociological Review*, oct 1938, *3* (5), 672.

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," aug 2007.

Prediger, Sebastian, Björn Vollan, and Benedikt Herrmann, "Resource scarcity and antisocial behavior," *Journal of Public Economics*, 2014, *119*, 1–9.

Smith, Adam, *The Theory of Moral Sentiments*, London: Penguin, 1759.

Starmer, Chris, "Developments in nonexpected-utility theory: The hunt for a descriptive theory of choice under risk," *Journal of Economic Literature*, 2000, *38* (2), 332–382.

A Theoretical Predictions

For the cheating and stealing tasks, in equilibrium there will be a $\bar{\theta}$, defined by $u(e+w(1))-u(e+w(0))=c(\bar{\theta})$ such that a share $F(\bar{\theta})$ will choose $d=1$. Define $B(e)=u(e+w(1))-u(e+w(0))$, by Assumption 1, $B'(e)<0$, implying that an NES shifts $\bar{\theta}$ to the right.

This is our first prediction:

Prediction 1. *In the cheating and stealing tasks:*

- $\frac{\partial P(d=1)}{\partial e}<0$

The JoD game introduces strategic considerations. The social norm is $n=d_j$. The payoffs are $w(0,0)>w(1,0)>w(0,1)>w(1,1)$. Define p to be the expected likelihood of $d_j=1$. The agent chooses $d=1$ if $pu(e+w(1,1))+(1-p)(u(e+w(1,0))-c(\theta))\geqslant p(u(e+w(0,1))-c(\theta))+(1-p)u(e+w(0,0))$.

If $p=0$ then $u(e+w(1,0))-u(e+w(0,0))<c(\theta)$, which implies $d=0$ and no effect of NES. If $p=1$, the DM will choose d=1 if $c(\theta)\geq(u(e+w(0,1))-u(e+w(1,1))$, i.e. if the cost of transgression is larger than the cost of retaliation. The latter is increasing in the endowment by Assumption 1, implying a rightward shift of $\bar{\theta}$ as a result of a NES.

This is our second testable prediction, which applies to the JoD:

Prediction 2. *In the JoD task:*

- $\frac{\partial P(d=1|p=1)}{\partial e}>0$

Consider now the pro-social tasks, starting from the the trust game. Recall first that the decision is dychotomous for both the first (FM) and second mover (SM), and, second, that $d=1$ is the harmful action. In other words, for both FM and SM, $d=1$ is to keep.

The analysis of the SM is straightforward. Since the social norm is $n=0$, the DM will keep if $B(e)=u(e+w(1))-u(e+w(0))\geqslant c(\theta)$, and since $w(1)>w(0)$, and $B'(e)<0$, this is equivalent to the analysis of the stealing and cheating tasks. Define $\bar{\theta}_{SM}$, the value of θ for the second mover, which is indifferent between sharing or keeping. $F_{SM}(\bar{\theta}_{SM})$ is the share of untrustworthy SMs.

The problem for the first mover under tit-for-tat is $p(u(e+w(1,1))-u(e+w(0,1)))+(1-p)(u(e+w(1,0))-u(e+w(0,0))))\geq(1-2p)c(\theta)$. Analyzing separately for $p=0$ and $p=1$, we can notice that $u(e+w(1,1))-u(e+w(0,1))>0$ and $u(e+w(1,0))-u(e+w(0,0))<0$. In other words, there is no cost of retaliation and no advantage of defection. Conditional on $p=0$ ($p=1$), the incentives and the norm prescribe to trust (not to trust). As a result, in this case, there is no effect of shock.

However in this sequential game, also the $n=0$ norm applies.

The FM decides to keep if $B(e)=p(u(e+w(1,1))-u(e+w(0,1)))+(1-p)(u(e+w(1,0))-u(e+w(0,0))))\geq c(\theta)$. By simple algebra $p=1\rightarrow B(e)>0$ and $p=0\rightarrow B(e)<0$. This implies that, conditional on $p=1$, since $B'(e)<0$, the $\bar{\theta}$ shifts to the right, while, conditional on $p=0$ there is no effect of shock, because of the TG assumption.

Summarizing, for the TG, the predictions are:

Prediction 3. *In the trust game:*

- $\frac{\partial P(d_{SM}=1)}{\partial e}<0$

- $\frac{\partial P(d_{FM}=1|p=1)}{\partial e} < 0$ *under* $n = 0$

Finally, we analyze the prisoner's dilemma game (PD). In this case $d = 1$ is No Cooperation.

We first derive the prediction for the case in which the social norm is to be a conditional cooperator ($n = d_j$). In this case, Player i will confess if $p(u(e + w(1, 1)) - u(e + w(0, 1))) + (1 - p)(u(e + w(1, 0)) - u(e + w(0, 0))) \geqslant (1 - 2p)c(\theta)$. Define $B(e) = p(u(e + w(1, 1)) - u(e + w(0, 1))) + (1 - p)(u(e + w(1, 0)) - u(e + w(0, 0)))$.

If $p = 0$, then $B(e) > 0$ and $B'(e) < 0$. In other words, an NES decreases cooperation. On the other hand, if $p = 1$, then there is no effect of shock because $B(e) > -c(\theta)$. That is, choosing $d = 1$ always gives a benefit grater than the cost.

For the norm of unconditional cooperation, $n = 0$, $p(u(e+w(1, 1))-u(e+w(0, 1)))+(1-p)(u(e+w(1, 0)) - u(e + w(0, 0))) \geqslant c(\theta)$ and under both $p = 0$ and $p = 1$, $B(e) = (u(e + w(1, 1)) - u(e + w(0, 1))) > 0$ and $B'(e) < 0$. This implies that $\frac{\partial P(d=1|p=1)}{\partial e} < 0$ and $\frac{\partial P(d=1|p=0)}{\partial e} < 0$.

Prediction 4. *In the prisoner's dilemma game:*

- $\frac{\partial P(d=1|p=0)}{\partial e} < 0$ *under the norm* $n = d_j$ *and* $n = 0$
- $\frac{\partial P(d=1|p=1)}{\partial e} < 0$ *under the norm* $n = 0$

B Proof of Proposition 1

Consider first the Prisoner's Dilemma. Notice that in an equilibrium, a DM chooses $d = 1$ iif $p(u(e+w(1, 1))-u(e+w(0, 1)))+(1-p)(u(e+w(1, 0))-u(e+w(0, 0)))) \geq (1-2n)c(\theta)$. Given the payoff of the PD, $p(u(e+w(1, 1)) - u(e+w(0, 1))) + (1-p)(u(e+w(1, 0)) - u(e+w(0, 0)))) > 0$. If $n = 0$ (norm of unconditional cooperation), $\exists\bar{\theta}$ such that $\forall\theta \in [0, \bar{\theta}]$, $d = 1$. In equilibrium, it must be that $p = F(\bar{\theta})$, thus $F(\bar{\theta})(u(e + w(1, 1)) - u(e + w(0, 1))) + (1 - F(\bar{\theta}))(u(e + w(1, 0)) - u(e + w(0, 0)))) - c(\bar{\theta}) = 0$. Define the equilibrium indifference condition for $\bar{\theta}$ as $\Phi(\bar{\theta}) = F(\bar{\theta})(u(e + w(1, 1)) - u(e + w(0, 1))) + (1 - F(\bar{\theta}))(u(e + w(1, 0)) - u(e + w(0, 0)))) - c(\bar{\theta}) = 0$.

Using Assumption 2, a single crossing property holds between the cost ($c(\theta)$) and benefit $F(\bar{\theta})(u(e+w(1, 1)) - u(e+w(0, 1))) + (1 - F(\bar{\theta}))(u(e+w(1, 0)) - u(e+w(0, 0)))))$ of deviation, and the benefit crosses the cost curve from above, i.e. $\frac{\partial\Phi(\bar{\theta})}{\partial\theta} < 0$. By Assumption 1, $\frac{\partial\Phi(\bar{\theta})}{\partial e} < 0$. Implicitly differentiating the equilibrium indifference conditions, gives $\frac{\partial\bar{\theta}}{\partial e} = -\frac{\frac{\partial\Phi(\bar{\theta})}{\partial e}}{\frac{\partial\Phi(\bar{\theta})}{\partial\theta}} < 0$, i.e a NES increases norm violation.

If the norm is $n = d_j$, the cost curve $(1 - 2F(\theta))c(\theta)$ has a zero in 0 and in 1/2, and it is first increasing then decreasing. This implies that there is more than one equilibrium, but only one is stable. In the stable equilibrium, $\frac{\partial\Phi(\bar{\theta})}{\partial\theta} < 0$ and the same comparative statics holds.

For the JoD, in equilibrium, a DM chooses $d = 1$ iif $p(u(e+w(1, 1))-u(e+w(0, 1)))+(1-p)(u(e+w(1, 0)) - u(e + w(0, 0)))) \geq (1 - 2p)c(\theta)$, where we use the social norm of retaliation. Since $p(u(e+w(1, 1))-u(e+w(0, 1)))+(1-p)(u(e+w(1, 0))-u(e+w(0, 0)))) < 0$, only high θ retaliate, i.e. by definition of equilibrium $p = 1 - F(\bar{\theta})$. The equilibrium indifference conditions becomes $(1-F(\bar{\theta}))(u(e+w(1, 1))-u(e+w(0, 1)))+F(\bar{\theta})(u(e+w(1, 0))-u(e+w(0, 0))))-(2F(\bar{\theta}))-1)c(\theta) = 0$. Notice that $\frac{\partial\Phi(\theta)}{\partial\theta} = -F'(\theta)\frac{\partial p(u(e+w(1,1))-u(e+w(0,1)))+(1-p)(u(e+w(1,0))-u(e+w(0,0))))-(1-2p)c(\theta)}{\partial p}$. For stability, we need $\frac{\partial p(u(e+w(1,1))-u(e+w(0,1)))+(1-p)(u(e+w(1,0))-u(e+w(0,0))))-(1-2p)c(\theta)}{\partial p} < 0$, thus $(1 - F(\bar{\theta}))(u(e + w(1, 1)) - u(e + w(0, 1))) + F(\bar{\theta})(u(e + w(1, 0)) - u(e + w(0, 0))))$ to cross

$(2F(\bar{\theta})) - 1)c(\theta)$ from below, i.e. $\frac{\partial \Phi(\bar{\theta})}{\partial \theta} > 0$. By Assumptions 1 and 2, $\frac{\partial \Phi(\bar{\theta})}{\partial e} > 0$, since $(1 - F(\bar{\theta}))(u(e + w(1,1)) - u(e + w(0,1))) + F(\bar{\theta})(u(e + w(1,0)) - u(e + w(0,0)))) < 0$, $\frac{\partial \bar{\theta}}{\partial e} = -\frac{\frac{\partial \Phi(\bar{\theta})}{\partial e}}{\frac{\partial \Phi(\bar{\theta})}{\partial \theta}} < 0$, i.e a NES increases norm violation and reduces the share of DM choosing $d = 1$.