Negative Economic Shocks and the Compliance to Social Norms

Francesco Bogliacino
Escuela de Economía,
Universidad Nacional, Bogota
fbogliacino@unal.edu.co*

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

February 2022

[Click here for the latest version of the paper]

*Corresponding Author. Francesco acknowledges financial support from Open Evidence via the contracts 18810 and 29047 with Universidad Nacional de Colombia. A special thanks to Roberto Galbiati, Zachary Grossman, Marcela Ibanez, Cesar Mantilla, Pietro Ortoleva, Daniel Parra, Leonardo Pejsachowicz, Gerhard Riener, Eugenio Verrina, for comments and suggestions. 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, and ESA Conference 2021. The usual disclaimer applies.

Abstract

This paper is about why suffering a Negative Economic Shock, i.e. a large loss, may trigger a change in behavior. We conjecture that people trade off a concern for money with a conditional preference to follow social norms, and that suffering a shock makes the first 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 elicit decisions in set of pro-social and anti-social settings. To derive our predictions, we elicit social norms separately from behavior.

We find that a shock increases deviations from norms in antisocial settings — more subjects cheat, steal, and avoid retaliation, with changes that are economically large. This is in line with our prediction. The effect on trust and cooperation is instead more ambiguous.

As a robustness check, we conducted an additional experiment to study the difference between an intentional shock and a random shock in a trust game. We found that the two induce partially different effects and that victims of intentional losses are more sensitive to the in-group bias. This may explain why part of the literature studying shocks in natural settings found an increase in pro-social behavior, contrary to our prediction.

As a final robustness check, we found that the effect of NES over norm compliance is robust when rules are imposed, the framing pro and anti social is removed, the shock is not fully anticipated. Moreover, it can be distinguished by a pure wealth effect.

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 now on) is a loss of earnings or accumulated assets. Shocks occur in both developed and developing countries as a result of natural disasters, violence and conflicts, health and trauma, macroeconomic crises and recessions. This paper investigates why suffering an NES can lead to behavioral change. There is compelling *causal* evidence that crimes against property surge due to shocks (Bignon et al., 2017; Cortés et al., 2016; Dix-Carneiro et al., 2018; Mehlum et al., 2006). At the opposite, quite a range of field (and natural experiment) studies claim to have found more prosociality after an NES (Cassar et al., 2017; Castillo et al., 2020; Bauer et al., 2016). Why do we get such contrasting predictions?

In stealing, cheating, trusting, and the other decision problems studied by the above-mentioned literature, *norms make preference social* (Kimbrough and Vostroknutov, 2018). In other words, the observed behavior largely reflects the presence of norms (Basu, 1998, 2000; Cialdini et al., 1991; Fershtman et al., 2012; Levitt and List, 2007; List, 2007). Social norms recommend and prohibit. They are rules of behavior 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). In other words, social norms are scripts that guide behavior and save cognitive resources for the decision-maker (Bicchieri, 2006; Bicchieri and Dimant, 2018, 2019).

To understand how exposure to shocks can lead to behavioral change, we first study the effect of random and frame-less losses. Following social norms is costly: Punishing transgressors, avoid cheating, and abstaining from free riding involve carrying out a cost. Our conjecture is that 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.

To assess whether this is the case, we reason in the following way. Through a model where norms enter the utility function and participants are heterogeneous in their psychological cost of compliance, we analyze optimal behavior in a set of 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 the counterpart. 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 violation after experiencing a shock. To assess this prediction, we design three experiments (and use data from our previous work). The key design choice is to manipulate NES by inducing large 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 and we measure whether norm compliance increases or decreases. The settings in which we explore our predictions are six: stealing, cheating, and Joy of Destruction (JoD) for anti-social behavior; trust, trustworthiness and cooperation for other regarding behavior.

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 paid to correctly guess the modal response to the PNBs. This elicitation is performed on participants who did not make an actual choice, to make sure that we elicit social norms separately from behavior (Krupka and Weber, 2013).

These are the main results. 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 declare four and five, the number with the highest payoff. This 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 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* anti-social activity, and is consistent with our controlled evidence. This result cannot be predicted by theories of crime like strain theory (Merton, 1938), according to which the frustration caused by NES should drive more stealing *and* money burning.

When asked to keep or share, in a binary version of the trust game (Berg et al., 1995), participants show less trustworthiness under NES by 8% of a sd, in the direction of the model, but clearly not statistically significant. In terms of trust, the exposure to an NES decreases trust conditional on not expecting the counterpart to share in return, again in line with our prediction, but the variation is small (9% of a sd), and not statistically significant. In the prisoner's dilemma, an NES slightly decreases cooperation, without reaching significance at the conventional levels.

Summing up, NES make extrinsic motivation more salient and may induce some people to deviate from social norms. We study this question experimentally: after formalizing a simple model, we conduct a set of experiments in which NES are administered randomly and participants make choices in a set of binary tasks, where norms apply. We find that NES increase deviations from norms in antisocial settings — more subjects cheat, steal, and avoid retaliation with economically relevant variations. This is in line with the predictions of the model. The effect in prosocial settings is instead inconclusive.

Towards the end of the paper, we study "framed" shocks. We compare random and intentional shocks based on the conjecture that shocks in natural disasters or conflicts are confounded by additional contextual elements. We conduct an additional experiment where participants interact in a non-binary trust game. However, a third treatment is added to the random shock and the control: some participants are randomly matched with someone who deliberately decided to rob them. This treatment is closer to the essence of the field studies and provides an abstract setting to study violence.

We first notice that allowing for an extensive margin (the trust game has more statistical power in this version) provides more conclusive evidence on the effect of random shocks, which decrease trustworthiness. Nevertheless, the key result is that the victims of intentional shocks are more prone to in-group bias. They are sensible and react upon the belief that the counterpart has been shocked, in a way that is not present either in random shock or the control. When we compute the treatment effects usually estimated in the field studies, we encounter evidence that is partially consistent with a positive shock effect, largely driven by the in-group bias.

Finally, we conclude this work by running a further experiment. This experiment adds three things: (a) we did not rely on an elicited norm but declared it, (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.

This paper makes four contributions to the existing literature. The fact that NES may generate anti-social behavior, in particular crimes against property, has not gone unnoticed in the literature. Using quasi-experimental research designs, there is compelling evidence of a positive relationship between negative economic shocks and anti-social behavior. 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 increase in property crimes (Mehlum et al., 2006). All this evidence is compatible with our results on stealing and motivates our research question.

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 is shown to correlate with disproportionate predatory behavior (Blanco et al., 2021). Since the authors manipulate shocks and criteria of assignment of social status, their evidence suggests that circumstances favor antisocial instincts. On JoD, the only related paper is Prediger et al. (2014), documenting positive money burning effect of long term exposure to scarcity. This contrasts with our prediction, which is conditional on the belief (the authors do not report a model with interaction). Their manipulation is not experimentally controlled, though, and takes place over a larger period, leaving space for the presence of confounds.

Studies on how social preferences evolve around exposure to violence, natural disasters, the pandemic, and other major shocks tend to argue that experiencing them increases pro-sociality (Adger et al., 2005; Cassar et al., 2017; Bauer et al., 2016; Bogliacino et al., 2020, 2021; Botchway and Filippin, 2021).[1] This evidence motivates our Experiment V, when we look at shocks occurring in the context of a traumatic event.

[1] Contrary to the main claim from the literature, Aycinena and Blanco (2021) show that exposure - but not the realization - of shocks in the context of Covid19 reduces trust.

Controlled evidence on the effect of shocks is provided by Bejarano et al. (2021, 2018). They found that shocks on the endowment in a trust game affect behavior but in a way that is entirely in line with assigning different endowments from the very start. Since they administer NES on the endowment in the stage game and not on the initial asset position separately from the main interaction (as in our setting), we think that inequality becomes very salient. In fact, their results are in line with a prediction from a variation of an inequality aversion model (Fehr and Schmidt, 1999). Our results complement their evidence.

Since norms are scripts that humans partially incorporate in 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 [2] 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 Boonmanunt et al. (2020) document that people under scarcity are less responsive to a social norm intervention.

The rest of the paper is organized as follows. Section 2 derives the theoretical predictions. Section 3 presents the elicitation of normative expectations and sections 4-7 present the experimental designs and results from four studies. Section 8 presents the experiment with intentional

[2] Carvalho et al. (2016) found no effect, but it should be highlighted that paycheck variations are temporary, expected and expected to be temporary.

and random shocks, Section 9 the experiment with an exogenous rule. A formal proof of Proposition 1 is in the Appendix. 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 social image have a similar framework, but 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 the counterpart a loss 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 $\mathbb{1}_{d_i \neq n} c(\theta)$, the psychological cost of deviating from a social norm n. The cost increases in θ, the propensity to comply. We have $\theta \in [0, 1]$, with Cumulative Density Function $F(\cdot)$.

In some situations, $d = 1$ transgresses a social norm (i.e. $n = 0$), as in stealing, in others, it is prescribed by the norm ($n = 1$), as in punishment. 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 violation. 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 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 $n = p$ (the latter is 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 $p=1$, mimicking the reasoning of a DM in a one shot interaction and consistently with the belief elicitation carried out in the lab. 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 it happens, 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 $u(e+w(1,0))-u(e+w(0,0)) \geq c(\theta)$. 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 six different settings: three anti-social and three pro-social. The anti-social tasks include cheating, stealing, and the Joy of Destruction (JoD). The pro-social tasks include the two sequential decisions of the investment game (trust and trustworthiness) and the prisoner's dilemma.

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 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).

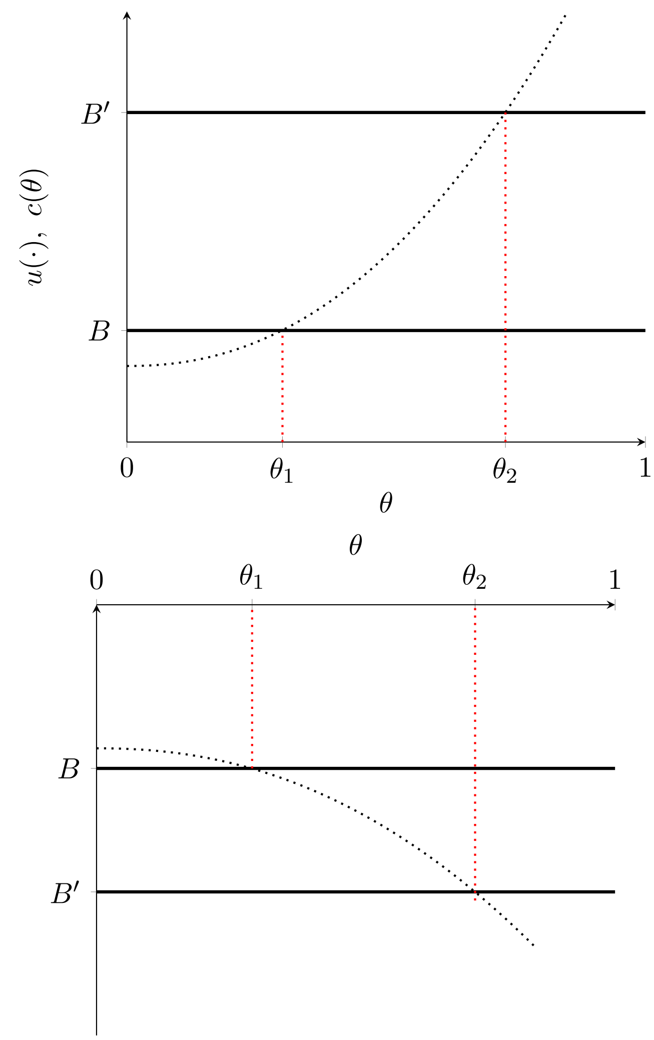


Figure 1: The optimal choice

For the trust game, we consider a sequential game where the first mover decides to pass or keep, and the second mover decides to share or keep. In payoff terms, for the first mover, $w(0,0) > w(1,1) = w(1,0) > w(0,1)$ holds and, for the second mover, $w(1) > w(0)$. To derive a prediction, the social norms for the trustor are either conditional trust or unconditional trust; the social norm for the trustee is to share.

The prisoner's dilemma is a symmetric simultaneous game where $w(1,0) > w(0,0) > w(1,1) > w(0,1)$. There are two social norms: conditional and unconditional cooperation.

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 money oriented leading to more transgression. For a trade off to exist, following the norm should be costly in terms of payoff. This is obviously the case for cheating, stealing and trustworthiness, where the cost of following the norm is the loss of income from $d = 1$.

To see why the trade off is needed, think at the trustor under the norm of tit-for-tat. The norm and the self interest recommend the same thing: trust if you expect the second mover to share and do not trust otherwise. In this case, the NES does not alter the behaviour in equilibrium. However, under the norm of unconditional trust, there is a trade off when the trustor expects the trustee to keep, since following the norm becomes costly (the transferred money will be lost). For those agents whose norm is unconditional trust and who expect lack of trustworthiness, an NES will increase norm transgression.

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 two trade-offs: a) under tit-for-tat, conditional cooperation is costly because defection is profitable; b) under unconditional cooperation, cooperating against a free rider is costly. In both cases, an NES generates more norm violation.

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

Setting	Social Norm	Prediction
Stealing	Do Not Steal	$S(NES) > S(C)$
Cheating	Do Not Cheat	$C(NES) > C(C)$
Money Burning	Retaliation	$D(NES) < D(C)\|P = 1$
Trustworthiness	Share	$Tw(NES) < Tw(C)$
Trust	Trusting	$T(NES) < T(C)\|P(Tw) = 0$
PD	Tit-for-Tat	$C(NES) < C(C)\|P(C) = 1$
PD	Cooperate	$C(NES) < C(C)\|P(C) = 0$

Table 1: Theoretical predictions from the norm compliance model.

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 ifθ_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 $u(e + w(d_i, d_j)) - c(d_i, \theta)$, 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, $\frac{\partial P(d=1)}{\partial e} < 0$.*

For the trust game, equilibrium should be refined to incorporate backward induction. The trustor anticipates that $F(\bar{\theta}_{SM})$ will chose $d = 1$, thus a DM chooses $d = 1$ in equilibrium iif $F(\bar{\theta}_{SM})(u(e + w(1, 1)) - u(e + w(0, 1))) + (1 - F(\bar{\theta}_{SM}))(u(e + w(1, 0)) - u(e + w(0, 0))) \geq (1 - 2n)c(\theta)$. Notice that for $n = d_j$, the problem has a bang-bang feature: if $\bar{\theta}_{SM}$ is sufficiently low, then the only equilibrium is $\bar{\theta}_{FM} = 0$, and the other way around. For $n = 0$, there is an interior equilibrium where $\bar{\theta}_{FM} > 0$ and using concavity we can prove that $\frac{\partial \bar{\theta}_{FM}}{\partial e} < 0$, or in other words, that $\frac{\partial P(d_{FM}=1)}{\partial e} < 0$.

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) and undesirable, 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:

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

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 not produce any effect 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 expectations, including normative expectations. Normative expectations are second order beliefs: what one expect others think it ought to 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, focality 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) asks subjects to report their (unincentivized) Personal Normative Beliefs (PNBs) for the action sets available to the DM, usually as a singleton. Then, in a second phase, they are asked to guess the response to the PNBs questions. As discussed and analyzed in Aycinena et al. (2021), KW and BX methods elicit respectively the intensive and extensive margin. Given that in our model the norm recommends an action, we rely on the latter.

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 pro-social and anti-social task used in this article. The response is elicited as a singleton (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 anti-social and three pro-social.

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). In each question, the order of the available options was randomized. In total, 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 participated to the experiments. This is to ensure elicitation of normative expectations separately from behavior (Krupka and Weber, 2013).

These were the action sets of the decision-maker in each setting, as they were 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) 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.

Empirical expectations are gathered for completeness.

We show the elicited normative expectations in Figure 2. For the stealing task, 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% od 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] For trust and cooperation, a more prevalent normative expectation of tit-fot-tat (66.97% and 42.20%) coexists with unconditional trust (23.85%) and

[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.

unconditional cooperation (21.10%). For trustworthiness, 85.32% of participants thought that sharing is the modal response to the PNBs question.

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

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. The endowment was assigned before starting the RET, to avoid giving a positive endowment shock after the NES. We also collected incentivized beliefs on whether the counterpart was affected by shock, and whether the counterpart was going to destroy. To reduce salience and the likelihood of hedging (Blanco et al., 2010), beliefs were paid with small incentives (1 ECU if correct). The Stealing task allowed participants to appropriate 80% of the earnings - obtained in a RET - from another participant in another session from another experiment.

This is how incentives were determined. Participants received the show up fee and the gain from the RET immediately after the end of the session. The money from the other two tasks (and the beliefs) were paid one week later.

The reason for the delayed payment was that we could not allow stealing within the session, as this was instrumental to manipulate intentional shocks in another experiment. Additionally, we

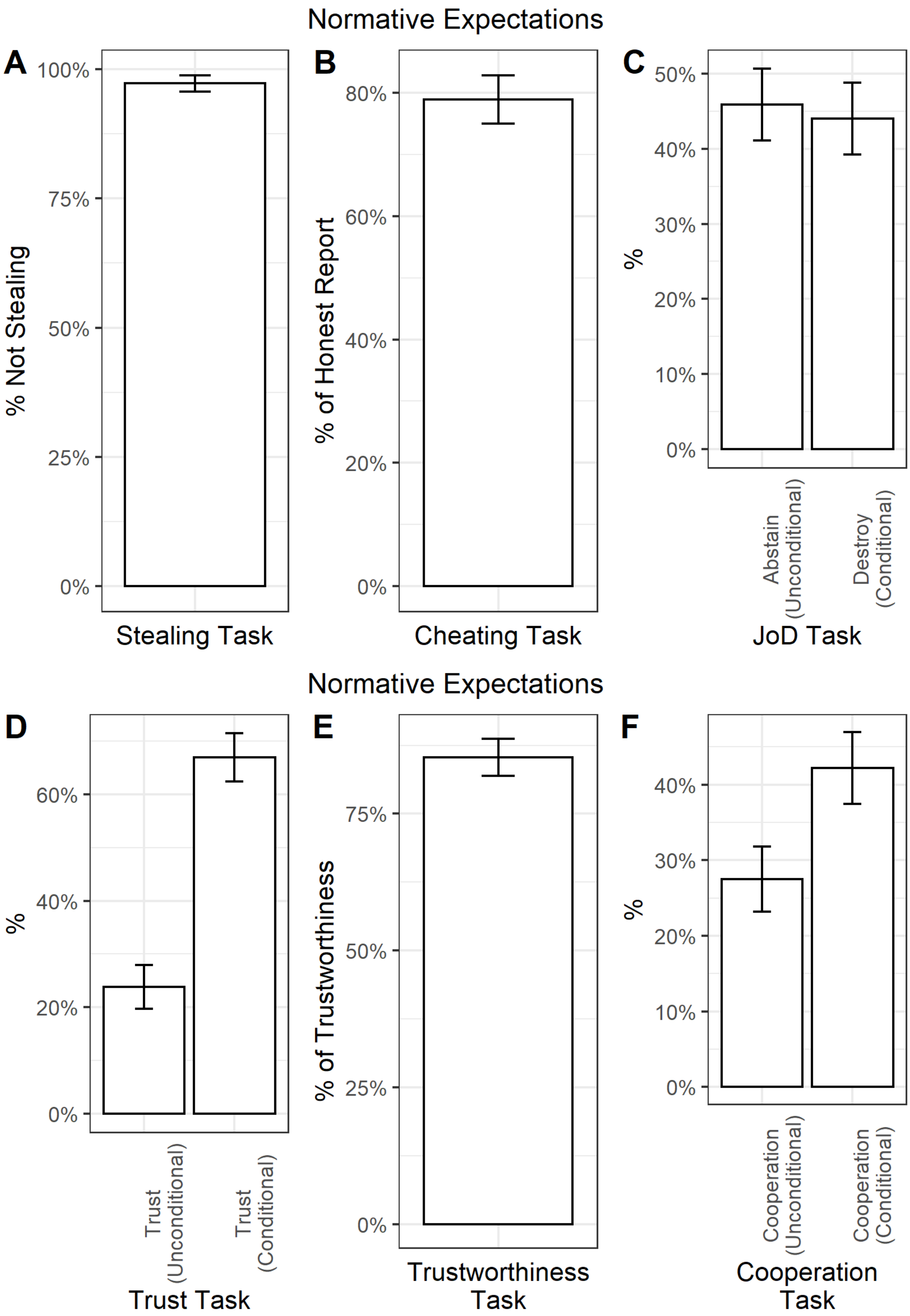


Figure 2: The elicited normative expectations

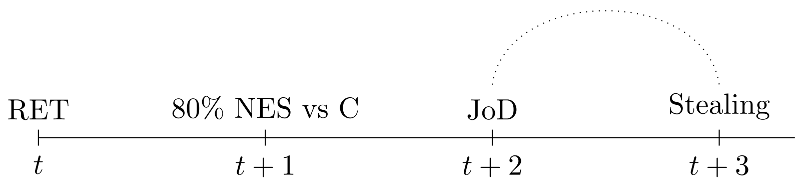


Figure 3: The timeline of Experiment I

suspected that two anti-social tasks with counterparts within the same session could generate compensatory behavior. To avoid asymmetry in the incentives for the JoD and the Stealing, we decided to pay both tasks with a delay.

The timeline of the experiment is reported in Figure 3. 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. They were told to raise their hand at any time if they had any questions. A final questionnaire was handed out to the participants.

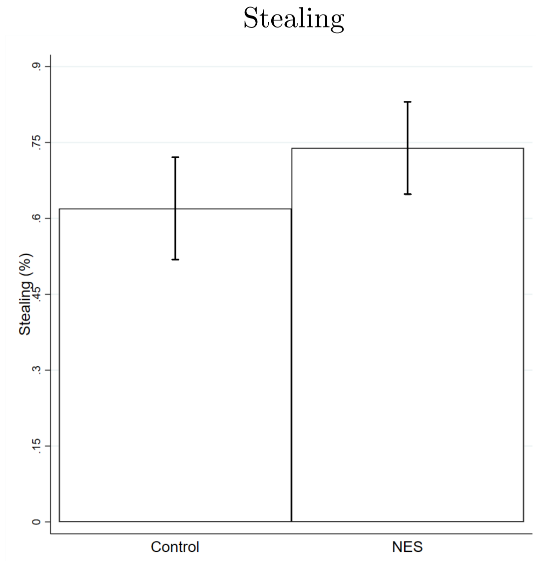
In total, we recruited 184 undergraduate students from the Unbiased subject pool. Invitations were randomized. Sessions took place in the lab, in presence, around October 2019.

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 9 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 9 (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 - that was randomized in the design - and compute a one sided test, since we are postulating a direction for the alternative hypothesis. Results are reported in Table 2, Column (1). The effect is both economically relevant, around 25% of a standard deviation of the outcome in the control condition, and statistically significant



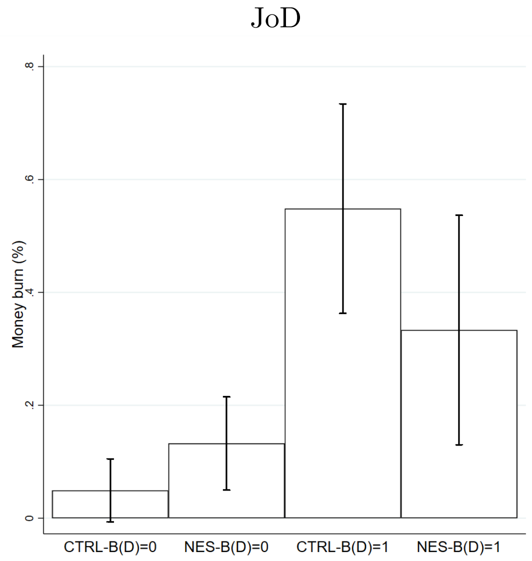


Figure 4: The impact of NES in the Stealing and JoD tasks

($F(1, 181) = 3.07$, $p = 0.04$ one sided).

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

	(1) Stealing	**(2) JoD**
NES	0.120*	
	(0.068)	
Order	0.126*	0.073
	(0.070)	(0.052)
NES-B(D)=0		0.080
		(0.049)
CTRL-B(D)=1		0.486***
		(0.096)
NES-B(D)=1		0.277***
		(0.102)
Constant	0.545***	0.011
	(0.067)	(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 reported in Figure 9 (right panel). Participants burn only 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.

5 Experiment II

5.1 Experimental Design and Procedures

Experiment II is built upon Experiment I. It is a between subject design, with two conditions and with the treatment assigned at the individual level, within session. 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 the Tagalog (the text was the Theory of Moral Sentiments; Smith 1759). We avoid more common languages to make sure that performance depended on effort and not on accumulated knowledge. Each fragment was 35 characters long. The software did not allow copy and paste.

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 were asked to roll a dice privately and to report their results. Participants had access to an online dice, beyond the control of the experimenters, 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.

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

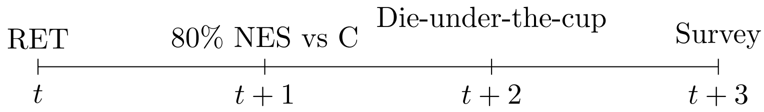


Figure 5: The timeline of Experiment II

Participants received a show up fee, the earnings from the RET, to guarantee salience of the shock, and those from the main task. In total, we recruited 158 participants. Data collection occurred in June 2020, on average each participant earned 24835 COP (around 7 USD).

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).

As expected, there is 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 the likelihood to report 4 or 5, the numbers with the highest payoff, 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 to report 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).

6 Experiment III

6.1 Experimental Design and Procedures

In experiment III, participants performed a RET, then they are assigned to the experimental conditions (80% NES or control, with 50% likelihood) and then performed a binary trust game

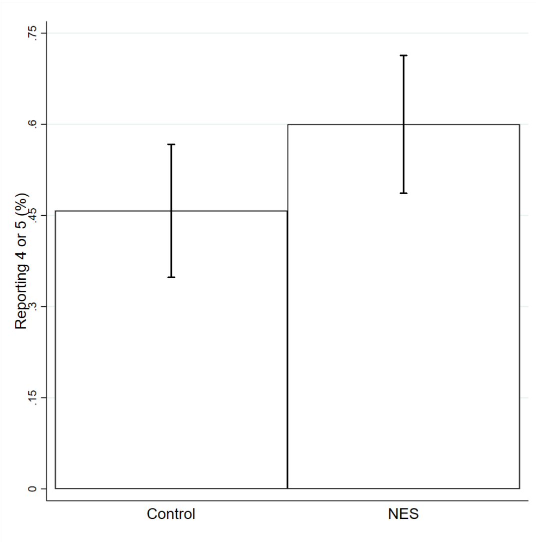


Figure 6: The impact of NES in the die-under-the-cup task

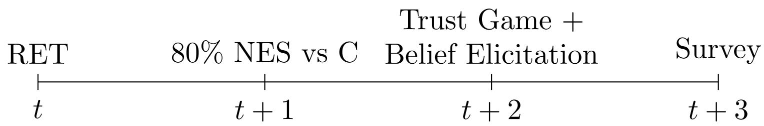


Figure 7: The timeline of Experiment III

(Berg et al., 1995) with role reversal and incentivized belief elicitation, following the timeline illustrated in Figure 7.

Assignment to treatment is between subjects, within sessions. The RET is identical to Experiment II.

The binary trust game is illustrated in Figure 8. Trustor and trustee were endowed with two ECUs. Trustor moved first and decided to either send her endowment to the trustee or keep it.

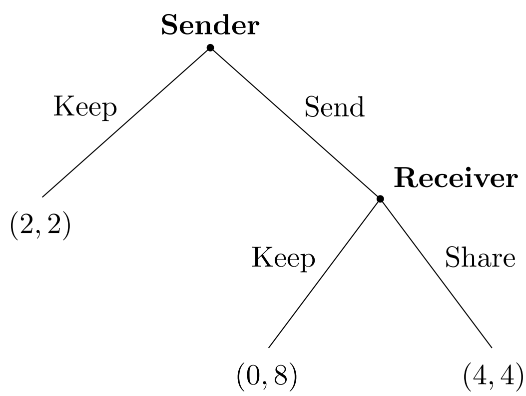


Figure 8: The Trust Game

If she sent, the amount was tripled. The trustee decided whether to share or keep. If he shared, both ended up with four ECUs, if he kept, he ended up with eight ECUs leaving the trustor with zero. If the trustor kept, both ended with the initial endowment.

Participants played both roles in random order. Before switching role, the participants have to declare their beliefs over the counterpart, paid with one ECU if correct.

The incentives included the RET (after the realization of the shock), one random decision between the trustor and the trustee, and a belief.

This experiment was conducted in online sessions on the Zoom platform. Random invitations were sent to the Unbiased subject pool, excluding participants from previous experiments with shocks. General instructions were read aloud and the participants were instructed to write in the chat to the research assistant for any questions. After the general introduction, participants were asked to follow the on screen instructions. A standard post experimental questionnaire was handed out at the end.

We run seven sessions in November-December 2020, with 150 participants, 78 in treatment and 72 in control. Participants earn on average $25'606$ COP, around USD 7.

The experiment was programmed in oTree (Chen et al., 2016). The protocol can be found in the SOM, Section IV.

6.2 Results

Participants solved, on average, 8.50 ± 3.78 transcriptions, and the performance does not differ between experimental conditions ($\chi^2 = 20.80$, $p = 0.23$).

On average, trustors sent 68.42% of the time in the control if they expected their counterpart to keep. This likelihood increased by 23 points, to 91.52%, when the subject believed that the trustee would share. This evidence is consistent with the presence of the two norms, the tit-for-tat and the unconditional trustor. When exposed to shocks, participants became less likely to trust, although the reduction was minimal (-3.42 pp, around 7% of a sd) and not statistically significant ($F(1, 146) = 0.05$, $p = .0.41$ one sided).

When playing as trustees, subjects shared 75.64% of the time in the control, and 72.22% in the

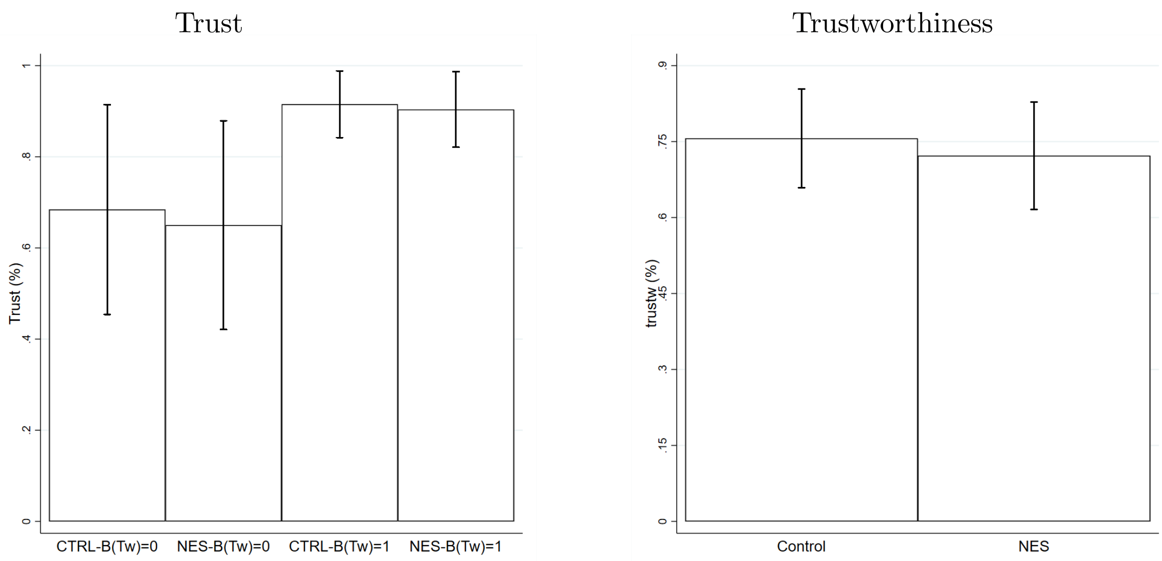


Figure 9: The impact of NES in the binary Trust Game

treatment. The reduction was qualitatively in the direction of the prediction, but quantitatively small (7.91% of a sd) and not statistically significant ($F(1,148) = 0.22$, $p = 0.32$ one sided).

The supporting OLS regressions are reported in Table 3. In Column (1), for trustworthiness, we included the treatment dummy as independent variable. In Column (2), for trust, we included three dummies in the regression ($NES{-}B(Tw) = 0$, $NES{-}B(Tw) = 1$, and $CTRL{-}B(Tw) = 1$), to estimate the impact with respect to the omitted category ($CTRL - B(Tw) = 0$). In the last row, we report the one sided p-value for the F test of the prediction.

Overall, the results are coherent with the prediction, but are inconclusive: due to NES, there is a minimal reduction of trustworthiness and a minimal reduction of trust conditional on the belief of non trustworthiness, but in both cases the estimated coefficients are not statistically significant. However, for the trust decision, this result reflect the smaller share of participants following the social norm of unconditional trust, as shown in Figure ??.

7 Experiment IV

7.1 Data description

Finally, we use the data from Study II in Bogliacino et al. (2020). The reader could find there a detailed description of the procedures and the experimental protocol (programmed in oTree;

Table 3: OLS estimates of effect of NES on Trust and Trustworthiness

	(1) Trust	**(2) Trustworthiness**
NES	-0.034	
	(0.072)	
NES-B(Tw)=0		-0.034
		(0.153)
CTRL-B(Tw)=1		0.231**
		(0.114)
NES-B(Tw)=1		0.220*
		(0.116)
Constant	0.756***	0.684***
	(0.049)	(0.108)
N	150	150
Test of prediction	0.32	0.41

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

Chen et al. 2016).

Participants are endowed with 20 tokens and could choose between two strategies, Cooperate (C) and Non Cooperate (NC). The game has a loss framing, with the payoffs reported in Table 4, right panel. This was instrumental to allow the initial endowment to be shocked. The strategies C and NC of the Table 4 were labelled green and blue. Since the participants were asked to make choices contingent of the counterpart's district of residence, we present the average over the 19 choices.

The experiment follows a 2 times 2 between subject design, with one factor being a 50% NES on the initial endowment and the other factor being a recall with two levels, violence recall and neutral recall. The belief over the counterpart's action was also elicited. In the analysis, we control for the dummy for priming. In this case, the initial endowment is unearned, participants directly made the decision in the PD after answering the recall question.

The subject pool was non standard and recruited among 18-30 y.o. residents of all the districts in Bogota. As shown in Table 4, left panel, the total number of observation is 223, of which 61 in the control and neutral recall, 60 in the NES and neutral recall, 51 in the control and violence recall and 51 in the NES and violence recall. Data collection took place in September-November 2018.

	Control	NES
Primed	51	51
Neutral	61	60

1/2	C	NC
C	-2, -2	-10, 0
NC	0, -10	-6, -6

Table 4: Experiment IV: design (left panel) and stage game (right panel)

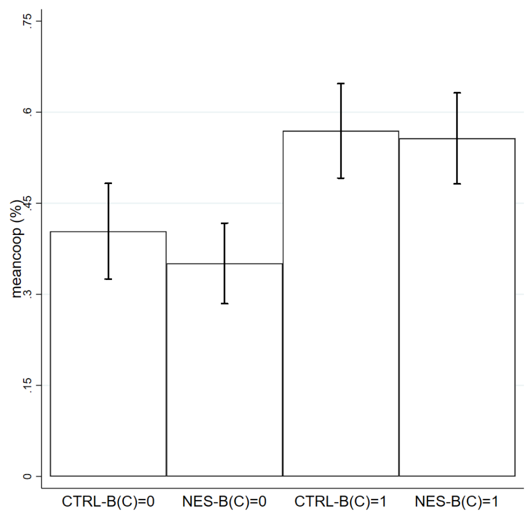


Figure 10: The impact of NES in the Cooperation Game

7.2 Results

These are the main results. Under the belief that the counterpart would not cooperate, participants chose cooperation 40.38% of the time. The shock decreased cooperation to 35.08%. The difference was not statistically significant ($F(1, 218) = 1.06$, $p = 0.15$ one sided). When expecting the counterpart to cooperate, participants chose to reciprocate 56.92% of the time. The cooperation rate was slightly lower under the shock (55.67%). The difference is not statistically significant ($F(1, 218) = 0.05$, $p = 0.41$ one sided).

The OLS regression is reported in Table 5. The outcome is the average cooperation and the independent variables are: a dummy equal to one if the participants is in the NES condition and has a belief of no cooperation ($NES - B(C) = 0$), a dummy equal to one if the participant is in the control condition and has a belief of cooperation ($CTRL - B(C) = 1$), and a dummy equal to one if the participant is in the NES condition and has a belief of cooperation ($NES - B(C) = 1$). In the last rows, we report the $p-$values for the $F-$tests for our two predictions. To wrap up, there is a marginal reduction in cooperation due to NES, but the effect is not statistically significant at the conventional level.

Table 5: OLS estimates of effect of NES on Cooperation

	(1) Cooperation
NES-B(C)=0	-0.053
	(0.051)
CTRL-B(C)=1	0.166***
	(0.055)
NES-B(C)=1	0.153***
	(0.054)
Recall	-0.009
	(0.037)
Constant	0.408***
	(0.045)
N	223
Test NES under B(C)=0	0.15
Test NES under B(C)=1	0.41

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

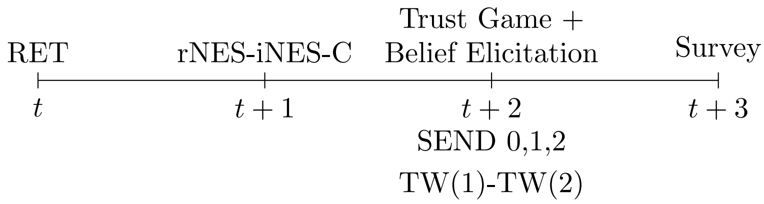


Figure 11: The timeline of Experiment V

8 Experimental Evidence from Intentional Shocks

The field (natural experiment) studies reviewed in the introduction elicited trust or other pro-social decisions around the exposure to a shock. The counterfactual is estimated according to different strategies: instrumental variable, randomized recall, or control for a reasonable set of covariates. These studies argue in favour a positive effect of a NES.

In these naturally occurring situations, the shock is typically confounded. Take violence, for example: suffering a violent act implies realizing that someone intentionally caused a loss. We build on this intuition to design a lab experiment that manipulates intentional shocks, similar to those studied in natural settings.

A minor advantage from the design of this experiment is that it allows for an intensive margin in the trust game, increasing statistical power.

8.1 Experimental Design and Procedures

Experiment V had a between subject design with two treatments and a control. In part I, subjects participated to a RET of four minutes after which they were assigned to the experimental conditions, at the individual level within session. In part II, they played a trust game with role switching and incentivized elicitation of beliefs. The timeline is illustrated in Figure 11.

The RET in Part I is a Niederle and Vesterlund (2007)'s task, as in Experiment I. In fact, these sessions were run in parallel to the latter. After receiving their earnings from the RET, participants are assigned to a random negative loss of 80% (rNES), are matched with a participant in Experiment I who decided intentionally to steal 80% of the earnings (iNES), or received no shock (C). This information is common knowledge: the initial instructions included examples and images to ensure full awareness of the nature of the shock.

The stage game in Part II was based on the Bogliacino et al. (2018)'s version of the trust game by Berg et al. (1995). Participants were randomly matched and received two ECUs each. The trustor decided to send zero, one or two tokens to the trustee, with the standard multiplier (3). Then the trustee decided whether to share (ending with equal payoffs) or not. The trustworthiness decisions were elicited with the strategy method. We adopted a neutral framing.

The elicitation of beliefs followed the standard procedures in Experiments I-III. Participants were asked to guess whether the counterpart had been shocked and what decisions she took. One belief was randomly selected for payment and paid with one ECU to prevent hedging.

We run a total of 19 sessions, in parallel with Experiment I, sending random invitations to the Unbiased subject pool. The average session had 14 observations, with minimal variation. We collected 261 data points. We had 80 observations in C, 85 in the rNES, and 96 in the iNES. Incentives included the earnings from the RET, one decision in the trust game, and a belief. Average payment was around 16000 COP. Participants solved on average 5.32 ± 2.12 additions. There were no differences across treatments ($\chi^2 = 17.38$, $p = 0.74$).

8.2 Results: Trustworthiness

Figure 12 plots the two trustworthiness decisions broken down by experimental condition.

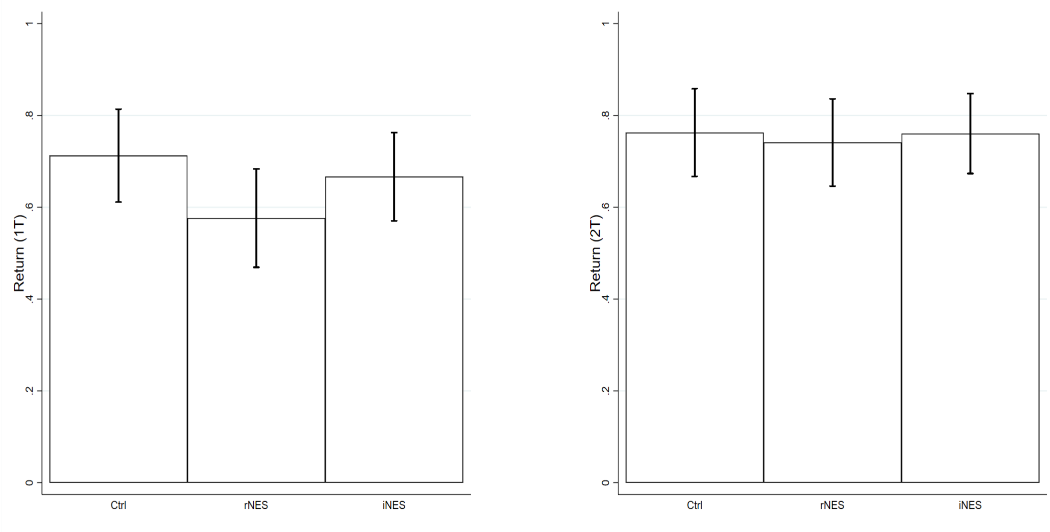


Figure 12: Trustworthiness: iNES and rNES

On average, in the control conditions, subjects share 71.25% of the time when one ECU is sent. This likelihood decreases to 57.64% in presence of a rNES. This difference is statistically significant (t-test controlling for unequal variance, $t = 1.83$, one sided $p = 0.03$).

The difference is softened when the shock is intentional, as the likelihood decreases to 66.66% ($t = 0.65$, $p = 0.25$). The difference between the two shocks is not statistically significant ($t = -1.24$, two sided $p = 0.21$).

The other trustworthiness decision is not different across conditions. The likelihood to share is 76.25% in the C, 74.11% in the rNES ($t = 0.31$, $p = 0.37$), and 76.04% in the iNES ($t = 0.03$, $p = 0.48$). The difference between shocks is not statistically significant ($t = -0.29$, $p = 0.76$).

At least for one of the trustworthiness decisions, data support our prediction for random shocks.

To shed further light on the role of intentional shocks, we condition on the belief that the counterpart has been shocked. When the trustee believed that the counterpart did not suffer the shock, she shares in case of one ECU 76.74% of the time. This likelihood decreased to 47.38% for rNES, and to 50% for iNES.

However, conditional on the belief that the counterpart suffered the shock, participants shared 64.86% of the times, i.e. a decrease of more than 10pp, in the control. In rNES, the share increased to 60.60% and in iNES, up to 69.51%. The difference between the two shock treatments is not statistically significant.

Table 6: OLS estimates of effect of rNES and iNES on Trust

	(1) Tw(1T)	(1) Tw(2T)
iNES-B(S)=1	-0.072	0.060
	(0.083)	(0.083)
rNES-B(S)=1	-0.161*	0.021
	(0.089)	(0.088)
C-B(S)=1	-0.119	0.090
	(0.103)	(0.095)
iNES-B(S)=0	-0.267*	-0.078
	(0.150)	(0.147)
rNES-B(S)=0	-0.294**	0.016
	(0.133)	(0.123)
Constant	0.767***	0.721***
	(0.065)	(0.069)
N	261	261

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

To understand these results, consider the regression in Table 6, column (1). Participants are divided into $C-B(S)=0$, $rNEs-B(S)=0$, $iNEs-B(S)=0$, $C-B(S)=1$, $rNEs-B(S)=1$, $iNEs-B(S)=1$. We added five dummies, leaving the $C-B(S)=0$ as the omitted category, captured by the constant.

The effect of intentional shock conditional on the counterpart being shocked, with respect to the effect of the shock conditional on the counterpart not being shocked is 31.39% and it is marginally significant ($F(1,255)=3.13$, $p=0.07$). The differential effect of intentional shock with respect to random shock is about 6.27% ($F(1,255)=0.10$, $p=0.74$).

The results for the other trustworthiness decision are less pronounced. They are shown in column (2). On average the second trustworthiness decision was 72.09% in the control and conditional on the belief that the counterpart was not shocked. This increased by 1.59% to 74.68% in presence of a rNES and decreased by 7.8% to 64.28% in presence of an iNES.

Conditional on believing the counterpart to be shocked, the likelihood increased to 81.08% in the control, to 74.24% in presence of an rNES and 78.04% in presence of an iNES.

The difference in difference is not statistically significant i.e. 4.77% ($F(1,255)=0.08$, $p=0.77$), and the intentional shock would be 13.20% larger than the random shock ($F(1,255)=0.54$, $p=0.46$).

Summing up, exposure to intentional shock seems to bias towards the in-group: victims of

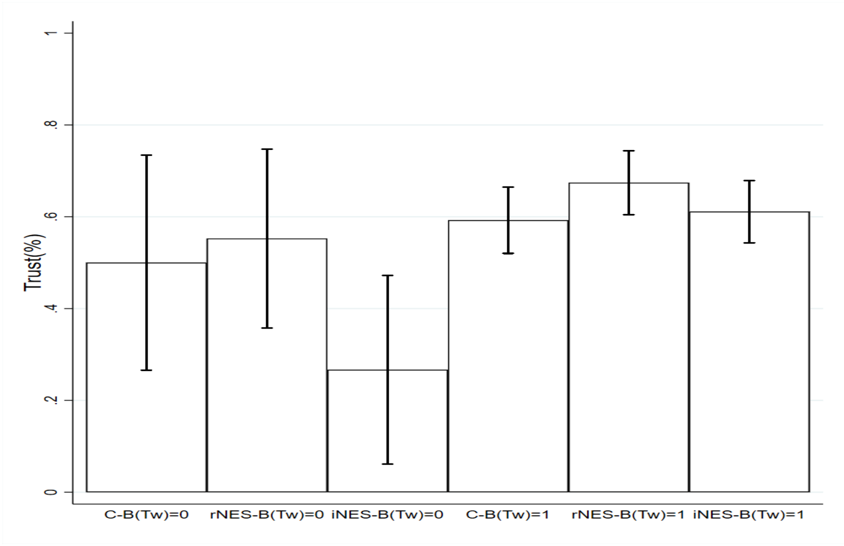


Figure 13: Trust: iNES and rNES

intentional shocks tend to treat more favorably those that went through the same experience than those who did not. Data from this controlled environment suggest that the estimated coefficient in a standard field study is likely to reflect this in-group bias and ends up inflating pro-sociality with respect to the pure random shock.

8.3 Results: Trust

Figure 13 plots trust broken down by experimental condition.

The outcome variable is computed as the amount sent divided by two, i.e. the intensive margin. Also, since we elicited the two beliefs over trustworthiness, we normalized them and created a single dummy equal to one if the first or the second belief was equal to one.

On average, the trust was 50.00% in the control and under the belief of not sharing in return, it is slightly higher in presence of a rNES (55.26%, $t = -0.36$, controlling for unequal variance, one sided $p = 0.64$), and instead lower in presence of an iNES, where it drops to 26.66% ($t = 1.60$, $p = 0.05$). The difference between the two shocks is statistically significant ($t = 2.14$, two sided $p = 0.04$).

Under the belief of sharing in return, trust increased to 59.23% in the control, to 67.42% in the rNES condition ($t = -1.62$, $p = 0.10$ two sided), and to 61.11% in the iNES ($t = -0.37$, $p = 0.70$). The difference between the two shocks was not statistically significant ($t = 1.29$, two sided $p = 0.19$).

We now look at trust considering both the belief that the counterpart had been shocked and the belief of trustworthiness. There are twelve cells, and we can run an OLS model with eleven dummies, reported in Table 7.

Table 7: OLS estimates of effect of rNES and iNES on Trust

	(1) Trust
iNES-B(Tw)=1-B(S)=1	0.247*
	(0.139)
rNES-B(Tw)=1-B(S)=1	0.268*
	(0.140)
iNES-B(Tw)=0-B(S)=1	-0.181
	(0.164)
rNES-B(Tw)=0-B(S)=1	0.178
	(0.171)
C-B(Tw)=1-B(S)=1	0.208
	(0.143)
C-B(Tw)=0-B(S)=1	0.278
	(0.206)
iNES-B(Tw)=1-B(S)=0	0.066
	(0.169)
rNES-B(Tw)=1-B(S)=0	0.344**
	(0.157)
C-B(Tw)=1-B(S)=0	0.199
	(0.145)
iNES-B(Tw)=0-B(S)=0	0.111
	(0.276)
rNES-B(Tw)=0-B(S)=0	0.111
	(0.225)
Constant	0.389***
	(0.134)
N	261

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

As can be seen from the regression, participants sent 38.88% of the endowment in the baseline, i.e. without any shock, not expecting any trustworthiness, and believing that the counterpart had not been shocked. In the other two treatments, under the same beliefs, the trust was slightly higher, by 11.11%. Expecting positive trustworthiness led to more trust, *ceteris paribus.* Trust increased to 58.82% in the C, to 73.33% in the rNES, but slightly decreased to 45.45% in the iNES. Under the belief that the counterpart had been shocked, and in absence of expected trustworthiness, the average trust is 66.66% in the control, 56.66% in the rNES, and only 20.83% in the iNES. In presence of expected trustworthiness, trust became 59.67% in the control, 65.68% in the rNES and 63.57% in the iNES.

In other words, exposure to iNES made the in-group dimension very salient. The level of trust towards the out-group is lower in iNES than in C and rNES, and not reactive to expected trustworthiness. At the opposite, iNES-affected participants became extremely reciprocal towards the in-group, barely willing to share if they do not expect trustworthiness but reaching the highest level of trust towards the trustworthy. Differently from the iNES, participants in the rNES and C conditions do not show an in-group bias.

9 Experimental Evidence from Rule Following

The last experiment follows the standard design of Experiment I-III but with three variations: this time the norm is not separately elicited but exogenously declared to the participants, wealth effects are separately controlled for in two additional treatments, and the shock is not fully specified ex ante (participants are told that the payment could change in a second phase, to avoid deception). 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 shock is announced but no loss is administered (No shock treatment); (2) each correct transcription is paid one point, the potential shock is announced and an 80% loss is 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 point (Low treatment) and there is no exposure to shock. Differently from Experiment I-III, we simply mentioned in treatments (1) and (2) that the first part is divided in two phases, and that the payment could change between phase one and two. After finishing this first part, participants were endowed with 30 seconds and asked to play with a stick man that should cross five traffic lights. Each second costed one point. The rule announced to participants was to wait for the green. Each traffic light took five seconds to change colour.

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

[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

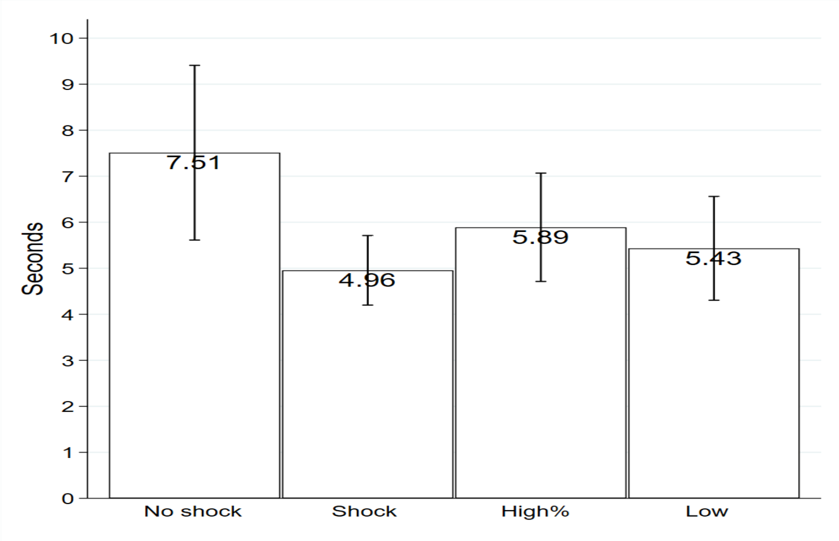


Figure 14: Results from the rule following task

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 to cross the path. On average, subjects spent 5.92 seconds (4.80 sd). In treatment (1) which is identical to the control in Experiment I-III, the outcome was 7.51 ($\pm$6.80). In treatment (2) (the NES condition in Experiment I-III) the outcome was 4.95 ($\pm$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 ($\pm$4.34) whereas in treatment (4) it was 5.43 ($\pm$4.28). A t-test controlling for unequal variance returns $t = 0.55$ (one sided $p = 0.29$). In other words, wealth effect did not generate the same result as the shock. Results are plotted in Figure 14.

Since it was preregistered, we run the OLS using the seconds as outcome variable, with robust standard errors. The difference between the shock effect (dummy for treatment 2) and the wealth effect (the difference in performance between treatment 3 and 4) was 2.94 seconds and is statistically significant ($t = -2.27$, $p = 0.02$).

10 Concluding remarks

In this article, we show how experiencing major losses can be a source of norm violation. Although this objective fact has been documented by studies on crime against property, the literature failed to grasp the underlying mechanism and its implications for our theories of

variability is much lower than expected.

strategic behavior.

The unanswered question is why the asymmetry in the results between pro and anti-social behavior. Whereas the evidence seems robust when we studied cheating, stealing or JoD, it does not show results at conventional statistical level in trust and cooperation. For the case of the trust game, the high level of trust and trustworthiness in the baseline may have determined a loss in statistical power, with respect to our expectations. In the case of the prisoner's dilemma, the use of data from an experiment that included other treatments may have weakened the salience of NES.

As usual, further tests are warranted, especially to understand the presence of competing norms, which may have further affected our results in pro-social settings.

This article has implications for the ongoing discussion on global challenges such as the pandemic, global warming, and the war threats. Theories of cultural evolution suggest that the Western world behave as it does because, among other things, the *W.E.I.R.D.* Acronym of Western, Educated and from an Industrialized, Rich, and Developed country. It has been coined by Heinrich et al. (2010). psychology and the related norms evolved 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.

Finally, this article has implications for the discussion around welfare state reform. It is commonplace to listen to arguments in favour of letting people learn the hard way, attenuating "that diaphragm of protections which, in the course of the twentieth century, have progressively distanced the individual from direct contact with the hardness of life" (Padoa-Schioppa, 2003). While it is true that market *disciplines*, it also increases vulnerability to shocks due to downward adjustment of price and earnings. The evidence from this article suggests that attenuating protection may have unintended consequences that should be factored in.

References

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

Adger, W. Neil, Terry P. Hughes, Carl Folke, Stephen R. Carpenter, and Johan Rockström, "Social-ecological resilience to coastal disasters," aug 2005.

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.

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 and Mariana Blanco, "Trust, COVID and Negative income shocks," 2021. Working Paper.

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

Basu, Kaushik, "Social Norms and the Law," may 1998.

_ , "The role of norms and law in economics: An essay on political economy," 2000.

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.

Bejarano, H., J. Gillet, and I. Rodriguez-Lara, "Do negative random shocks affect trust and trustworthiness?," *Southern Economic Journal*, 2018, *85* (2), 563–579.

Bejarano, Hernán, Joris Gillet, and Ismael Rodriguez-Lara, "Trust and trustworthiness after negative random shocks," *Journal of Economic Psychology*, oct 2021, *86*, 102422.

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

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, jul 1995, *10* (1), 122–142.

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.

_ and Eugen Dimant, "It's Not A Lie If You Believe It. Lying and Belief Distortion Under Norm-Uncertainty," 2018.

_ and _ , "Nudging with Care: The Risks and Benefits of Social Information," *Public Choice*, 2019, (July).

_ , _ , 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.

_ , Cristiano Codagnone, Felipe Montealegre, Frans Folkvord, Camilo Gómez, Rafael Charris, Giovanni Liva, Francisco Lupiáñez-Villanueva, and Giuseppe A. Veltri, "Negative shocks predict change in cognitive function and preferences: assessing the negative affect and stress hypothesis," *Scientific Reports*, 2021, *11* (1), 1–10.

_ , Laura Jiménez Lozano, and Gianluca Grimalda, "Consultative democracy and trust," *Structural Change and Economic Dynamics*, 2018, *44*, 55–67.

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.

Castillo, Geoffrey, Lawrence Choo, and Veronika Grimm, "The stability of norms elicited with coordination games," 2020.

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.

Cialdini, Robert B, Carl A Kallgren, and Raymond R Reno, "A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior," *Advances in Experimental Social Psychology*, 1991, *24*, 201–234.

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.

Fehr, Ernst and Klaus M. Schmidt, "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

Fershtman, Chaim, Uri Gneezy, and John A List, "Equity Aversion: Social Norms and the Desire to be Ahead," *American Economic Journal: Microeconomics*, nov 2012, *4* (4), 131–144.

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.

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

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.

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.

Padoa-Schioppa, Tommaso, "Da Berlino e Parigi ritorno alla realtà," 2003.

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 $d = 1$.