CESifo WORKING PAPERS

9067
2021

May 2021

Deviant or Wrong? The Effects of Norm Information on the Efficacy of Punishment

Cristina Bicchieri, Eugen Dimant, Erte Xiao

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9067*

Deviant or Wrong? The Effects of Norm Information on the Efficacy of Punishment

Abstract

Research examining the effect of weak punishment on conformity indicates that punishment can backfire and lead to suboptimal social outcomes. We examine whether this effect is due to a lack of perceived legitimacy of rule enforcement, which would enable agents to justify selfish behavior. We address the question of legitimacy by shedding light upon the importance of social norms and their interplay with weak punishment in the context of a trust game. Across six conditions, we systematically vary the combination of existence of weak punishment and norm information. Norm information may refer either to what most others do (empirical) or to what most others deem appropriate (normative). We show that in isolation, neither weak punishment nor empirical/normative information increase prosocial, reciprocal behavior. We instead find that reciprocity significantly increases when normative information and weak punishment are combined, but only when compliance is relatively cheap. When compliance is more costly, we find that the combination of punishment and generic empirical information about others' conformity can have detrimental effects. In additional experiments, we show that this negative effect can be attributed to the punishment being perceived as unjustified, at least in some individuals. Our results have important implications for researchers and practitioners alike.

JEL-Codes: C910, D030, D730, H260.

Keywords: conformity, punishment, social norms, trust.

Cristina Bicchieri
University of Pennsylvania / Philadelphia / USA
cb36@sas.upenn.edu

Eugen Dimant
University of Pennsylvania / Philadelphia / USA
edimant@sas.upenn.edu

Erte Xiao
Monash University / Melbourne / Australia
erte.xiao@monash.edu

This version: February 1, 2021
The most recent version of this working paper can be downloaded by following this link:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3294371

This version of the paper is forthcoming at the Journal of Economic Behavior and Organization. We would like to thank Ernst Fehr, Daniel Houser, Rosemarie Nagel, Daniele Nosenzo, and David Rand, as well as seminar participants at Pompeu Fabra University and Yale University for their helpful comments and suggestions. We also appreciate the feedback received at the Social Norms in Epistemic Communities conference at Monash University, 2017's ANZWEE annual conference at the University of Melbourne, the 7th Biennial Workshop on Social Dilemmas, and 2017's North American Economic Science Association Meeting. We acknowledge financial support from the German Research Foundation (DFG) under Germany's Excellence Strategy — EXC 2126/1— 390838866.

1. Introduction

To encourage prosocial behavior, policy interventions have commonly used punishment, with mixed success. Since severe punishment usually requires costly monitoring and can have undesirable side effects, punishment is often weak (Tyler, 2006; Balafoutas et al., 2016), which means that the cost of punishment is lower than the cost of compliance. How can non-deterrent weak punishment effectively enforce cooperation?

We consider that punishment works via two channels: an incentive effect (e.g., increasing the cost of deviance) and a communication effect (e.g. conveying disapproval or signaling what the right action is).[1] With weak punishment, the incentive effect is not sufficient to change behavior. The outcome of punishment therefore is mainly determined by its communication effect. For example, weak punishment can effectively promote contribution to public goods when it is self-selected by the group members as the choice of imposing punishment signals their willingness to contribute (Tyran and Feld, 2006; Markussen et al., 2014; Romaniuc et al., 2020). Conversely, weak punishment, when interpreted as signaling that there is a small price for deviance, can crowd out intrinsic cooperative motivation and have the opposite effect of promoting transgressions (Gneezy and Rustichini, 2000).[2]

The importance of the communication function of weak punishment suggests that compliance is more likely when an agent thinks that enforcement is legitimate. In particular, our main hypothesis is that the combination of (weak) punishment and norm information has an advantage in inducing prosocial behavior when compared to independently implementing punishment. Our reasoning is that norm information signals that the enforcement refers to a shared agreement, indicating that non-conformity is *commonly* thought to be wrong and blameworthy.[3] We report results from three experiments – one laboratory behavioral experiment and two follow-up belief elicitation online experiments – to better understand the effect of the combination of punishment and norm information.

Norm information may take different forms, since social norms have both an empirical

[1] See, e.g., Xiao (2018). The expressive function of punishment has been widely discussed in the law literature (Cooter, 1998; Kahan, 1998; Sunstein, 1996).

[2] Studies have argued that the presence of social preferences (e.g., inequality aversion) explain why weak punishment can still promote public goods contributions (Engel, 2014; Kosfeld et al., 2009).

[3] We also differentiate between what might be thought of as 'arbitrary' punishment and punishment that is justified by the violation of an agreed upon rule of behavior (see the discussion in Xiao, 2018).

and a normative component (Bicchieri, 2006).[4] They tell us what is commonly done (empirical) as well as what is commonly approved of (normative). Empirical information alone may indirectly suggest the underlying normative appropriateness of a behavior. Normative information instead provides a direct and explicit signal about whether an action is appropriate, although it does not necessarily imply that most people behave accordingly (Bicchieri et al., 2020).[5] As we often have access to only one type of norm information, it is important to investigate their potentially different effects on behavior, especially when accompanied by punishment.[6] Adding normative information about an enforced behavior points out that people (usually in the form of a majority) view non-compliance as wrong. However, when an enforced behavior is only supported by empirical information and non-compliance is presented as a deviation from what is commonly done, agents can only indirectly infer that the violation is socially disapproved. (Bicchieri et al., 2019, for a recent discussion see Bicchieri and Dimant, 2019). We thus examine further whether punishment is more effective when the enforced behavior is presented as the right course of action rather than what is commonly done in the same situation. This consideration is also consistent with the observation that punishment in naturally occurring environments is usually associated with what is wrong and what should not be done, rather than what the majority does (Bicchieri, 2016; Dimant and Gesche, 2020).

Using a variant of the standard Trust Game, we first conducted a laboratory experiment to examine how providing the information that an enforced behavior is consistent with a shared norm influences the effectiveness of punishment and thus conformity. We extended our analysis with two additional online experiments. In the first experiment, we confirmed the existence of a shared perception of a norm of reciprocity in our setting. In the second experiment, we checked how the norm information provided in the experiment was interpreted by the participants. We deem our holistic experimental approach important to

[4] In social psychology, a distinction is made between descriptive and injunctive norms (Cialdini et al., 1990). Empirical information points to a descriptive norm, whereas normative information points to an injunctive one. Our definition of social norms (Bicchieri, 2006, 2016) includes both kinds of information.

[5] Both types of norm information have been used in policy interventions, e.g. to curb electricity or water use and often differ in their success (Schultz et al., 2007; Allcott and Mullainathan, 2010; Ferraro et al., 2011; Bicchieri and Dimant, 2019).

[6] In their littering studies, Cialdini et al. (1990) introduced explicit normative messages. They differ from our messages in that they are directly injunctive (e.g., "it is important to recycle" or "avoid littering") as opposed to informing people directly about what most others (dis)approve. The latter approach has the advantage of being explicitly focused on changing/creating new expectations, which are the foundations of behavior change (Bicchieri, 2006; Bicchieri and Dimant, 2019).

understanding both how and why norm information may or may not achieve the desired behavioral effect in our context. We briefly explain the designs of all three experiments.

In the laboratory experiment, we systematically vary the (co-)existence of norm-related information and punishment across six conditions. To investigate the effects of combining (weak) punishment[7] and norm information, we introduced a Baseline control (a standard Trust Game) and a total of five variations (treatments), in which norm information and punishment are systematically varied. Each treatment consisted of multiple rounds of play and random re-matching of pairs after each period. We introduced combinations of (weak) punishment, normative information, and empirical information at the beginning of each treatment. Participants were assigned either to the role of investor or trustee for the duration of their session and restricted to only one treatment variation (between-subjects design). All treatments were variations of the Baseline condition, in which the investor's decision space regarding how much to transfer to the trustee was limited to three choices (either none, half, or all of their initial endowment) and the decision whether to accompany a non-zero transfer with a return-request message. We employed a fixed-form message that asked the trustee to return at least half of the received amount. In our setup, a request was non-binding in conditions where punishment was not included. Where it was included, ignoring a request led to the automatic enforcement of punishment.

Any amount given by the investor was then tripled and transferred to the trustee who then decided how much, if anything, to return to the investor. The investor learned the actual return behavior of all trustees that s/he encountered throughout all periods only *after* the experiment concluded. We split up trustees in high (low) stakes groups when trustees received all (half) of the investor's endowment because the trustee's cost of conforming to the return request varied between the two cases. In addition to examining the average returns, we follow Xiao and Houser (2011) and categorize the trustees into different *types* and analyze the impact of punishment and norm-related information, both in isolation and combined, with respect to shifts of proportions of those types.

The three treatments with punishment varied on whether the request message was accompanied by empirical information (Pun_EmpInfo), normative information (Pun_NormInfo),

[7] We only focus on weak punishment (i.e., the cost of punishment does not exceed the cost of compliance and thus monetary incentives are not the dominant driver of decisions). As a result, the punishment is not equilibrium-shifting under selfish preferences, which sets our study apart from much of the existing research (Ensminger and Henrich, 2014, for an overview of the literature see Xiao, 2018).

or was accompanied with no information (Pun_NoInfo).[8] In these treatments, the investor's request message was binding. If the trustee returned less than 50%, she would receive a fixed monetary penalty. Since we were only interested in weak punishment, the penalty was always smaller than the 50% return, giving the trustee a monetary incentive to transgress. The empirical information let players know – based on data collected before the experimental session – that in a previous session most trustees *returned* at least 50%, and the normative information stated that most participants in a previous session thought that trustees **should** return at least 50%. To control for the effect of information alone, we included two more treatments where punishment was absent and only empirical information (NoPun_EmpInfo) or normative information (NoPun_NormInfo) was provided.

Our main goal was to study the effects of each type of information, especially when paired with weak punishment.[9] We find that only the joint effect of normative information and punishment significantly increases conformity, while the separate enforcement mechanisms of punishment alone and normative information alone do not achieve this result. Again, punishment alone could be perceived as the arbitrary choice of the investor and thus the norm communication effect of punishment is not sufficient to enforce targeted reciprocal behavior. Norm information (both normative and empirical) alone, without the risk of being punished, may not be sufficient either whenever compliance is too costly. An unexpected finding is that the combination of punishment and empirical information can have detrimental effects on conformity when conformity is especially costly. Our results raise concerns about linking punishment and empirical information, which is commonly done in the recent wave of social norm nudging (e.g., Hallsworth et al., 2017; Dimant and Gesche, 2020). We further explore this combination to explain its negative outcome with two different experiments. In the first experiment, we elicit third parties' assessment of whether a social norm of reciprocity is in place in either the low- or the high-stakes condition of the game. We follow Bicchieri and Chavez (2010) and Bicchieri and Xiao (2009) and elicit personal normative beliefs, normative expectations, and empirical expectations. In both

[8] To ensure the truthfulness of the information, the empirical and normative information used was borrowed from the behavior and beliefs of other participants from a previous study. This approach is commonly adopted in experimental social norms research (Bicchieri and Xiao, 2009; Krupka and Weber, 2009).

[9] Our results help to better understand recent research with conflicting findings about separately manipulated normative or empirical information. See for example Bicchieri, 2006; Goldstein et al., 2008; Ferraro et al., 2011; Keane and Nickerson, 2015; Hallsworth et al., 2017; Bursztyn et al., 2018; Bott et al., 2019; Bursztyn et al., 2019; Dimant et al., 2019; Bicchieri et al., 2020; Bolton et al., 2020.

conditions, we find that a basic norm of reciprocity – i.e., return at least what was sent by the investor – was endorsed. In the second experiment, we assess how people interpret the normative or empirical information that we have provided in the trust game experiment. Since the combination of empirical information and punishment backfires in the high-stakes condition, we hypothesized that players do not think the (generic) empirical message about a majority of players returning half of the amount applies to the high-stakes condition. In our case, the combination of punishment and generic information that may be perceived as *intentionally* exploited by the investor could be interpreted as forcing punishment on high-stake trustees who do not behave like a majority of low-stakes ones. Indeed, this is what happens. If the threat of punishment is accompanied by information that is perceived as not applicable, it is not surprising that the reaction may be negative.

Our work contributes to the understanding of how punishment impacts pro-social behavior. This is particularly important from a policy perspective, with regard to designing effective and sustainable behavioral interventions. The effect of punishment can be enhanced when it is accompanied by norm-relevant information. However, we show that the information must be both specific and credible to obtain the desired result.

2. Experiment Design and Procedures

Punishment has been widely studied in many contexts, such as public goods games (PG). In a topically related paper, Andrighetto et al. (2013) examine the relationship between punishment and injunctive information in public goods games. Our experiment utilizes a variant of a Trust Game (Berg et al., 1995) that has demonstrated the backfiring effect of weak punishment (e.g., Fehr and Rockenbach, 2003; Houser et al., 2008). We choose this game because the punishment inflicted by an investor has clear self-interest motives, whereas punishment in settings that encompass a strategic dilemma, e.g. a public goods game, can also benefit other group members and thus could be motivated by other-regarding preferences, introducing unnecessary noise in our design. Following established protocols, we can thus provide a controlled experimental environment to test whether adding norm-related information can change people's perception of the legitimacy of the enforcement.

For a given experiment session, each participant was randomly assigned to the role of an investor or a trustee and remained in that role for the duration of the experiment. Each participant played the game for 10 rounds. At the beginning of each round, each participant received an endowment of 8 Experimental Currency Units (ECU; 2 ECUs =

$1) and was randomly matched with another participant in a different role.

Treatments varied by punishment (absent, present), norm information (absent, a normative message about what other participants believed ought to be done, an empirical message about what other participants did), and combinations thereof. As in Bicchieri and Xiao (2009), all data from which the truthful messages were based on data generated in a pilot trust game. In this pilot session, the majority of participants returned at least 50% of the tripled amount and the majority also indicated that Player 2 should return at least half of the tripled amount.

Treatments	*No Punishment*	*Punishment*
No Information	Baseline (60)	Pun_NoInfo (68)
Normative Information	NoPun_NormInfo (58)	Pun_NormInfo (62)
Empirical Information	NoPun_EmpInfo (94)	Pun_EmpInfo (76)

Table 1: Treatment overview and number of participants (in parentheses). These are total numbers and are split equally ($=\frac{n}{2}$) between investors and trustees.

2.1. Treatments

Figure 1 outlines the game played in each round in each treatment.

2.1.1. Baseline

At the beginning of each round, the investor had to decide how much to transfer to the trustee. The transfer (T) could be either 0 ECU, 4 ECU, or 8 ECU. We limited the action space of the investor to allow differentiation between low- and high-stakes investments across all treatments (explained in more detail below). Participants were informed that the transferred amount was multiplied by a factor of 3 by an experimenter. When deciding how much to transfer, the investor also had to decide whether to send a costless request message to the trustee, indicating whether he/she wanted the trustee to return 50% of the transfer. The message was in a fixed form, with two quantitative components adjusted corresponding to the investor's selected transfer; for example, "I'd like you to transfer back to me at least half of the 12 ECU (i.e., at least 6 ECU)." All participants knew that the investor chose whether to send the return request message or not.

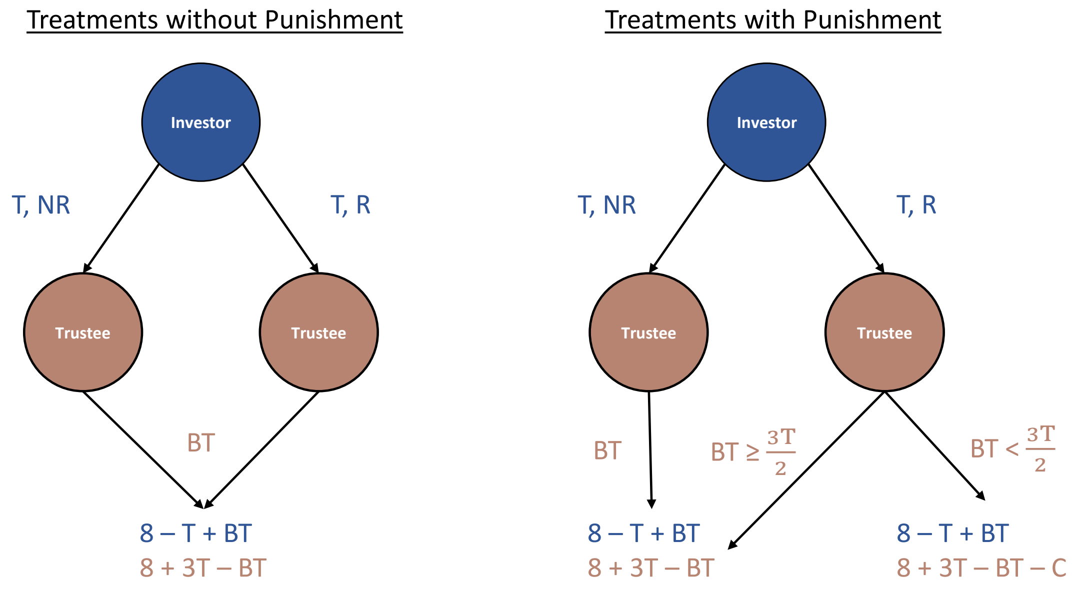


Figure 1: Sequence of actions and payoff structure in treatments with and without punishment. T: Investor's transfer to trustee. (N)R: Investor's decision to (not) send a return request message to the trustee. BT: Trustee's back-transfer to the Investor. C: Trustee's payoff cut (punishment)

Next, the trustee saw the transferred (tripled) amount and whether the investor sent a request message. Then, the trustee decided how much to transfer back to the investor. The back transfer amount (BT) is represented by any integer in the range of [0, 3T].

To provide clean evidence for the effect of punishment on the trustees' return decisions (see below), the investors did not know the trustees' return amount in each round until all ten rounds were completed. Specifically, all investors were shown a summary of the decisions and outcomes of each round only at the end of the experiment. Thus, our design avoids the possibility that trustees' return behavior in each round might influence investors' transfer decisions in the next round, which might in turn influence the trustees' behavior. One round was randomly chosen as the payoff round and participants were paid the amounts they earned in that round.

2.1.2. Punishment Treatments

In the three treatments with the punishment opportunity (Pun_NoInfo, Pun_NormInfo, and Pun_EmpInfo), participants were told that if the investor sends a return request message, the trustee would receive a payoff cut of 5 ECU if his/her back transfer amount was less than 50% of the tripled transfer amount. This amount would then go back to the experimenter.

Thus, the mechanism through which the investor obtains money is limited to reciprocation from the trustee. If the investor does not send the return request message, the trustee does not receive a payoff cut regardless of the amount of the back transfer.

Our design of punishment treatment is a simplified version of (Fehr and Rockenbach, 2003; Houser et al., 2008) where the investors had to decide both how much to request from the trustee and whether to impose punishment. We ask the investors to decide whether to make the request and, if the request is made, punishment will be automatically imposed if the request is not met. The simplification is important for testing our hypotheses. By allowing the investors to make only one decision (to send or not send a request for a fixed amount), we can better understand the treatment effect on the trustees' responses. An alternative design would have been to exogenously set the minimum amount to be returned and ask the investor to decide whether to impose punishment if the trustee returns less than that. As our focus is the importance of the trustees' perception of the legitimacy of the request, it is important to let the investor make the choice whether to send a request.[10] Lastly, by fixing the request to be 50% of the received amount, in the norm information conditions we only need to reveal to the trustees whether requesting 50% is consistent with a norm. This feature makes the design easier to implement.

2.1.3. Norm Information Treatments (Normative or Empirical)

We adopted the design of Bicchieri and Xiao (2009) in the four treatments with normative or empirical information (Pun_NormInfo, Pun_EmpInfo, NoPun_NormInfo, and NoPun_EmpInfo). In the treatments with normative information, the instructions read: *"In a previous survey, most participants said that Player 2 should return at least half of the tripled transferred amount."* In treatments with empirical information, the instructions read: *"In a previous experiment, most participants in the role of Player 2 returned at least half of the tripled transferred amount to Player 1."*[11]

[10] The choice to send a request may be viewed as a signal that the investor does not trust the trustee. The norm information may further interact with this signal. For example, the investor's decision whether to send the request may be affected by the message. However, as we report later, we do not observe the frequency of request to vary across treatments. In fact, the requests are made most of the time.

[11] Since our main focus is to study the relationship between norm information and punishment and to retain comparability within and across treatments, we used generic empirical/normative messages throughout the experiment. That is, following existing literature, we did not specify whether the truthfully-obtained information was the result of behavior/beliefs in low- or high-stakes situations. In our design, introducing this separation would have created comparability problems and potential information asymmetries. Hence, while the source of the information was transparent and unambiguous (i.e., taken from a previous sur-

To summarize, in the Baseline condition, subjects played a trust game and the investor could send a non-binding request message asking the trustee to return at least 50% of the transferred amount. In Pun_NoInfo, when the investor chose to send the request message, the trustee would receive a penalty if he/she returned less than 50%. In the Pun_NoInfo treatment, participants did not receive any information about previous participants' choices or beliefs, whereas those in the Pun_NormInfo (Pun_EmpInfo) treatment learned that most players in a previous game thought trustees should (did) return at least 50%. Finally, the NoPun_NormInfo and NoPun_EmpInfo treatments only differed from the Pun_NormInfo and Pun_EmpInfo treatments in that the return request message was not accompanied by punishment if the trustee did not return enough money. These last two treatments let us examine whether any difference between the Pun_NormInfo (Pun_EmpInfo) and the Baseline treatments can be attributed to the normative (empirical) information alone.

2.2. Procedure

The experimental sessions were conducted at the Behavioral Ethics Lab at the University of Pennsylvania using participants recruited through an institutional human-subjects research platform, Experiments@Penn. We recruited a total of 418 participants (in 34 sessions ranging from 10-16 participants) across six treatments at the University of Pennsylvania in the summer of 2016.[12] For each session, one treatment was chosen at random. Participants were then randomized to different computer booths and a second randomizer determined which participants would be assigned to the role of an investor or trustee. Irrespective of their role, each participant received the same full set of instructions. These were handed out as a hard copy and then read aloud by the experimenter to achieve transparency and common knowledge. The average duration of a session, which included the game and a post-experiment questionnaire, was 45 minutes. The average hourly compensation was $18, which included a $10 show-up fee. The experiment was programmed using z-Tree (Fischbacher, 2007). Across all treatments, the average age of the participants was 22.2 years old, and 62.7% of participants were female.

vey/experiment), the exact content of this information remained unspecified. We return to this point in our discussion section.

[12] In light of our surprising results that mainly occurred in the empirical information conditions, we over-proportionally collected data in those conditions to ensure the robustness of our findings.

3. Theoretical Analysis and Hypotheses

Our main question is whether punishment is more effective when combined with normative or empirical information. Presenting empirical or normative information may make the (reciprocity) norm more salient. Non-conformity (i.e., in the form of returning zero) in this situation might increase the psychological cost due to the disutility of norm violation.[13]

We argue that punishment accompanied by norm information increases the saliency that the punished behavior violates the respective norm. And, depending on one's sensitivity to the specific norm, this salience can increase the disutility of violation. As a result, punishment can change behavior even when the monetary cost alone is not sufficient to enforce conformity. To formalize this, we adopt the norm-based utility function framework introduced in Bicchieri (2006): the disutility from norm violations depends on (1) the difference between the payoff from a chosen action and the payoff from following the norm, and (2) the individual's sensitivity to the relevant norm.

Let $k \geq 0$ be Player 2's sensitivity toward the norm (denoted as r^0) then Player 2's disutility of deviating can be defined as:

$$k \times max\{[m - r - (m - r^0)], 0\} \tag{1}$$

where m is the transferred amount. Player 2 decides how much to return (r) to:

$$\max_r U = m - r - k \times max\{[m - r - (m - r^0)], 0\} \tag{2}$$

In Appendix A, we provide a detailed analysis of each treatment and stake condition: low-stakes (L) when half of the money was sent to Player 2 and high-stakes (H) when all of the money was sent to Player 2. Here we summarize the main findings and the hypotheses. We assume that without the norm information message (in both the Baseline and the Pun_NoInfo treatments), a trustee thinks the acceptable return amount (r^0) can be equal or less than 50% and no less than the investment amount (4 or 8 ECU).[14]

In the four treatments in which the normative/empirical information is received (two Pun_Info and two NoPun_Info treatments), the trustee will believe that $r^0 = 6$ in the low-stakes case and $r^0 = 12$ in the high stakes case. The trustee would not give more than r^0.

[13] We assume that the psychological cost is a direct function of a person's sensitivity to the norm, or caring about what the norm stands for.

[14] This is also why we design the request to be 50% so we can examine the effect of explicit norm information.

In the two Pun_Info treatments, the punishment imposes a cost of 5 ECU when the trustee returns less than requested. In the two NoPun_Info treatments, there is no monetary cost of returning less than r^0. We summarize the conditions below for each treatment to achieve a higher return than the Baseline condition:

Case 1: Low-Stakes

$$\begin{cases} \mathrm{r}^*_{PunNoInfo\_L} > r^*_{Baseline\_L}, & \text{if there is a significant number of individuals whose k}>\frac{1}{r^0_{PunNoInfo\_L}} \\ \mathrm{r}^*_{PunInfo\_L} > r^*_{Baseline\_L}, & \text{if there is a significant number of individuals whose k}>\frac{1}{6} \\ \mathrm{r}^*_{PunInfo\_L} > r^*_{PunNoInfo\_L}, & \text{if there is a significant number of individuals whose } \frac{1}{6}<\text{k}<\frac{1}{r^0_{PunNoInfo\_L}} \\ \mathrm{r}^*_{NoPunInfo\_L} > r^*_{Baseline\_L}, & \text{if there is a significant number of individuals whose k}>1 \end{cases}$$

Note when k$<\frac{1}{6}$ then $r^*_{PunInfo\_L} = r^*_{PunNoInfo\_L} = r^*_{Baseline\_L} = 0$.

Case 2: High-Stakes

$$\begin{cases} \mathrm{r}^*_{PunNoInfo\_H} > r^*_{Baseline\_H}, & \text{if there is a significant number of individuals whose k}>\frac{7}{r^0_{PunNoInfo\_H}} \\ \mathrm{r}^*_{PunInfo\_H} > r^*_{Baseline\_H}, & \text{if there is a significant number of individuals whose k}>\frac{7}{12} \\ \mathrm{r}^*_{PunInfo\_H} > r^*_{PunNoInfo\_H}, & \text{if there is a significant number of individuals whose } \frac{7}{12}<\text{k}<\frac{7}{r^0_{PunNoInfo\_H}} \\ \mathrm{r}^*_{NoPunInfo\_H} > r^*_{Baseline\_H}, & \text{if there is a significant number of individuals whose k}>1 \end{cases}$$

Note when k$<\frac{7}{12}$ then $r^*_{PunInfo\_H} = r^*_{PunNoInfo\_H} = r^*_{Baseline\_H} = 0$.

In both cases, the potential positive effect of punishment may be significantly diminished if there is a crowding out effect, in that punishment may change people's perception of the decision environment from norm-based to profit-maximizing (i.e., one could pay a small fine to transgress, see Gneezy and Rustichini, 2000). This means that in our framework, $r^0_{PunNoInfo} = 0$. If so, then $r^*_{PunNoInfo} = 0$, regardless of the value of k.

Hypotheses

Comparing the conditions of each treatment to achieve a higher return than the baseline, we derive our main hypothesis:

Hypothesis 1: Pun_EmpInfo and Pun_NormInfo are always more effective than the NoPun_EmpInfo and NoPun_NormInfo and the Pun_NoInfo treatments.

Furthermore, we note that the conditions for the two PunInfo treatments to achieve higher returns than the Baseline condition are stricter when the stakes are high than when they are low. Thus, our second hypothesis is:

Hypothesis 2: It is more likely to observe the PunInfo effect in the low-stakes than the high-stakes condition.

Our hypotheses assume that the provided norm information will lead subjects to update their belief to one that returning at least 50% is the right thing to do. As previously discussed, while the normative information directly and explicitly conveys what is socially approved, the empirical information only indirectly conveys this message. The impact of the empirical information thus may depend on how the trustees interpret it. If the empirical information is less effective in expressing the wrongness of returning less than 50% than the direct normative information, we expect that the Empirical information treatments may not be as effective as the Normative information treatments.

4. Results

We investigate the return behavior of trustees in different treatments varying punishment, norm information, and combinations thereof.[15] In the subsequent sections, we focus on the trustees' average return behavior.[16] Note again that we refer to the case of 8-ECU transfer as the High-Stakes (HS) situation, which requires trustees to return 12 ECU, and the case of 4-ECU transfer as the Low-Stakes (LS) condition, which requires trustees to return 6 ECU. Pursuing the same analytical strategy as Houser et al. (2008), we first examine the data both in pooled form as well as separately by its stakes (HS and LS).

In the spirit of a repeated strategic game situation (such as in the Trust Game and Prisoner's Dilemma Game, see, e.g., Anderhub et al. 2002; Dal Bó and Fréchette 2011) without rematch or feedback between rounds, we follow related literature (see, e.g., Huck

[15] In line with our motivation, we limit our attention to the role of punishment and norm information on trustee behavior, and control for investor behavior in our regression analyses.

[16] All investors sent a return request message at least once (overall, 93% of the time), with no significant differences between treatments. To allow for comparability between treatments, our main analyses focus on the cases where a return request message was sent. Our regression analyses in Section 4.4 are robust to the inclusion of the no-request cases (see Table A1 in the Appendix).

et al., 2012) in our analysis and treat each of the investors' decisions as independent (for a discussion see Camerer, 1997; Binmore and Shaked, 2010; Charness et al., 2012). To account for this, we will use the bootstrap approach as proposed by Moffatt (2015) for our mean comparisons of trustees' return behavior.[17] Controlling for covariates, trends, and clustering of standard errors yields coherent results (see Section 4.4).

As our interest is the effect of punishment, we compare trustees' behavior when the investors made the request across treatment. In all the treatments, most investors sent the request and we do not observe treatment differences in the frequency of sending the request messages (see Footnote 16 and Table A1 in the appendix for a regression analysis that incorporates all observations, including those where no return message was sent, indicating that our results are robust to the inclusion of those data). We find that punishment alone does not successfully improve return rates, especially in High-Stakes. Neither empirical nor normative information alone induces a return rate higher than that of the Baseline. The combination of punishment and normative information produces substantial positive behavioral change but only in Low-Stakes. These results are consistent with Hypotheses 1 and 2. Interestingly, the combination of punishment and empirical information is not only ineffective when the compliance cost is low, but is detrimental when the compliance cost is high. This detrimental effect suggests that a self-serving bias may arise when empirical information is ambiguous and can be interpreted in multiple ways, as we shall discuss later (Bicchieri et al., 2020; Bolton et al., 2020).

4.1. Effect of Punishment Alone

Figure 2 reports the average return (in percentage) for the Baseline and Punishment treatments. Punishment does not significantly increase the return levels by trustees, either in HS or LS. For the pooled results, introduction of punishment yields a non-significant increase from 32.4% to 35.6% (BSM, p=0.19). The results are the same for both LS and HS separately (p=0.10 and p=0.91, respectively).

Next, we classify the behavior of trustees into three types (for a related approach, see Houser et al., 2008): Complete Violation of trust if returned amount (r) equals 0%; Incomplete Conformity if 0% $<$ r $<$ 50%; Complete Conformity if r $\geq$ 50%.[18]

[17] We employ the bootstrap two-sample t-test method (BSM; see Moffatt 2015) with 9999 replications. BSM (significant at $p < 0.05$) retains the rich cardinal information in the data without making any assumptions about the distribution.

[18] For an analysis of types, we calculate three ratios for high- and low-stakes per participant, each of

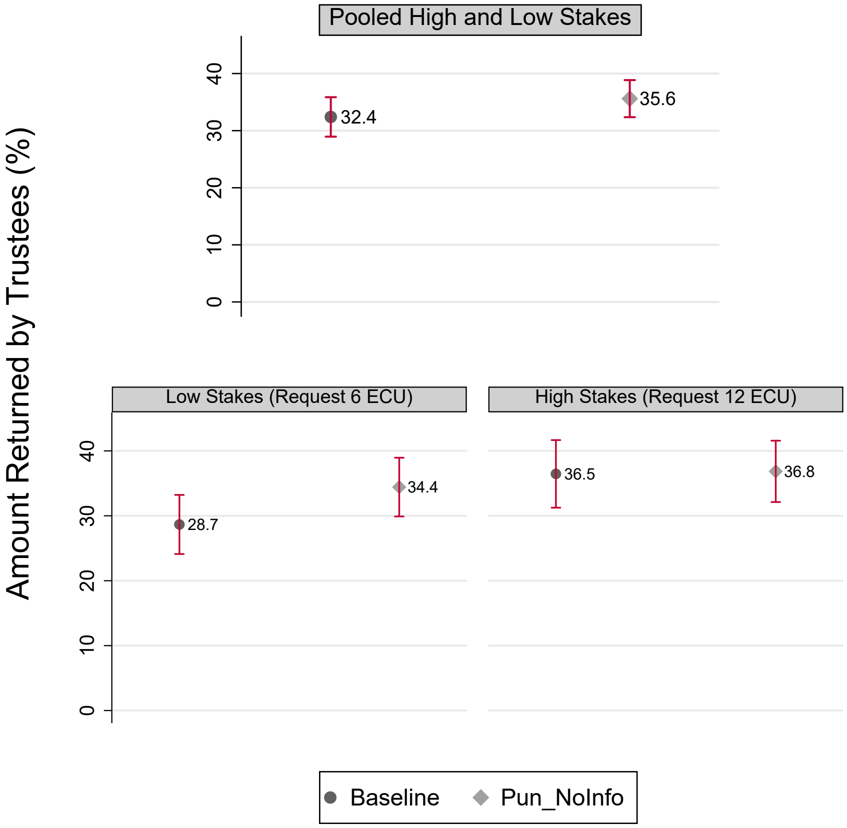


Figure 2: Amounts returned by trustees as percentage of amounts received from investors; upper part indicates pooled amounts; lower part indicates amounts per LS vs. HS; Baseline: no punishment or norm information; Pun_NoInfo: punishment (5 ECU) without norm information. None of the comparisons are significant at the conventional levels. Whiskers represent 95% CIs.

Figure 3 plots the distribution of the three types in each of the four conditions. Kolmogorov-Smirnov (K-S) tests suggest that the distributions in the low-stakes condition are significantly different between the Baseline and punishment treatments (p<0.01).

Consistent with Houser et al. (2008), we observe a bimodal return pattern in the punishment conditions and a significant decrease in the proportion of Incomplete Conformity (0% vs. 25.0%, BSM, p<0.01). While the proportion of Complete Violation is insignificant (33.6% vs. 35.0, BSM, p=0.94), punishment significantly increases the proportion of Com-

which indicates the fraction of Complete Violation, Incomplete Conformity, or Complete Conformity at the individual level across all rounds. In so doing, we account for behavioral changes across all rounds and the fact that under different stakes, decisions could be impacted by the transferred amount.

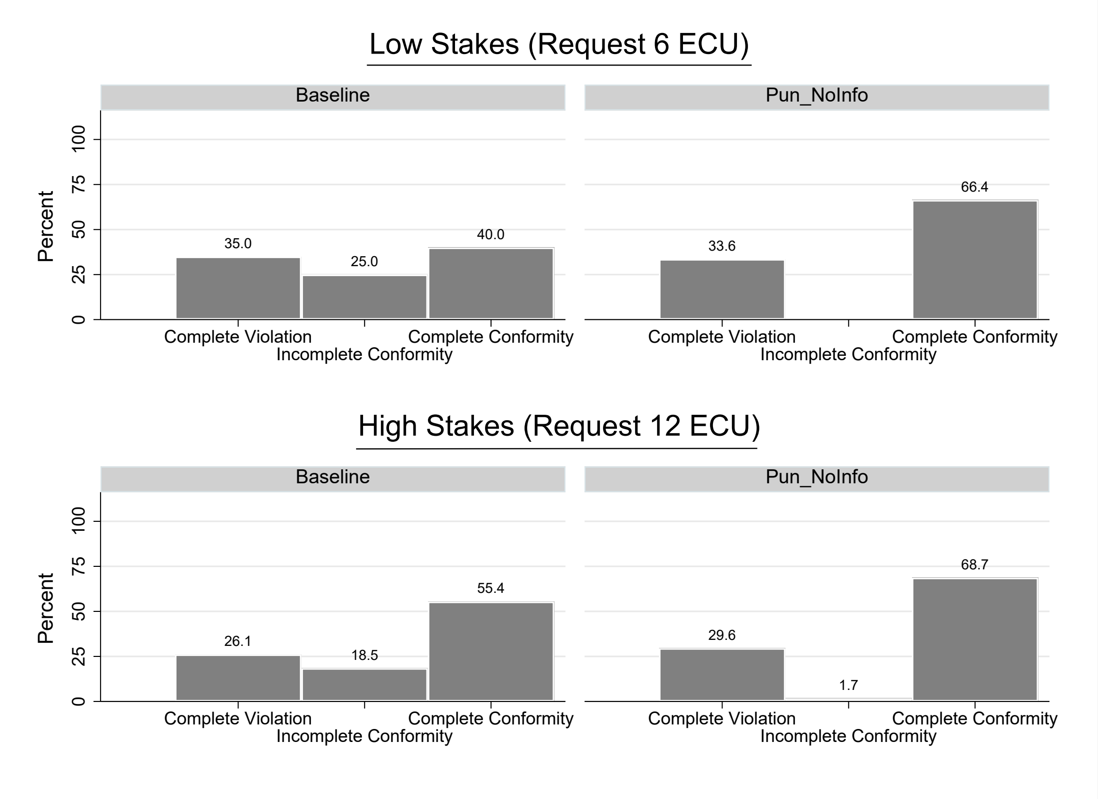


Figure 3: Distribution of return types in Baseline (NoPun_NoInfo) and Pun_NoInfo conditions.

plete Conformity (40% vs. 66.4%, BSM, p=0.04). The positive shift does not translate into a significant change in average return behavior, partly because many of the Incomplete Conformity types in the Baseline condition were right below the 50% cut-off.

In contrast, in HS the difference between the punishment condition and the Baseline condition is relatively small and non-significant (K-S, p=0.33). While we observe significantly fewer Incomplete Conformity types in the Punishment condition than in the Baseline condition (1.7% vs. 18.5%, BSM, p<0.01), the effect of Pun_NoInfo on the other two types is not statistically significant (Complete Violation: 29.6% vs. 26.1%, BSM, p=0.41; Complete Conformity: 68.7% vs. 55.4%, BSM, p=0.52).

Overall, we can conclude that – in our setting – the effect of weak punishment alone on behavior is nuanced but mostly negligible. As Figure 2 indicates, weak punishment does not change average return behavior, neither in the aggregate nor conditional on low or high stakes. However, in following Houser et al. (2008), we also examined the changes in 'types'. As in Houser et al. (2008), we observe that in the presence of punishment, investors either achieve a return they aimed for or nothing at all. Unlike Houser et al. (2008) who found that punishment increased the rate of Complete Violation when the requested return was more than double the penalty amount, we did not find such a detrimental effect of punishment in HS. In addition to individual differences, this discrepancy may exist because Houser et al. (2008) allowed return requests much higher than 50%, which led to less compliance.

4.2. Effect of Norm Information Alone

Our results in Figure 4 indicate that norm information in isolation does not affect behavior, which aligns with Dimant et al. (2019), for example. This null finding holds for both pooled data and for LS and HS separately. In particular, for LS and HS, the differences in average returns between the Baseline and both NoPun_NormInfo and NoPun_EmpInfo are not statistically significant (LS: 28.7% vs. 23.7%, BSM, p=0.17; 28.7% vs. 23.3%, BSM, p=0.11; HS: 36.5% vs. 30.5%, BSM, p=0.11; 36.5% vs. 30.9%, BSM, p=0.11).

Figure 5 reports distributions of the three return types for the information only and Baseline treatments. None of the pairwise distribution comparisons between the information only and the Baseline treatments reach statistical significance.

In sum, when looking at these results in conjunction with those reported previously, we do not observe an effect of punishment or norm information in isolation. In the next section, we will examine this effect in more detail when norm information is combined with punishment, rendering the rule and the cost of compliance even more salient. As will be shown, the combination of both is vital to behavioral change.

4.3. Effect of Punishment and Norm Information Combined

Figure 6 plots the average return in the Baseline, Pun_NoInfo, Pun_NormInfo and Pun_EmpInfo treatments. When pooling the two stakes situations, we observe a significant decrease in the trustees' average return in the Pun_EmpInfo condition compared to that in the Pun_NoInfo and Pun_NormInfo treatments (BSM, both p<0.01).

The combination of punishment and normative information leads to a significant increase in trustees' return behavior in LS over the Baseline (42.7% vs. 28.7%, BSM, p<0.01), well above the insignificant 5.7% increase in the Pun_NoInfo compared to the Baseline condition. The return rate is also significantly higher than that in the Pun_NoInfo treatment (42.7% vs. 34.4%, BSM, p=0.02). In the Pun_EmpInfo treatment, we did not observe a similar difference from the Baseline condition (32.1% vs. 28.7%, BSM, p=0.25). The return rate in the Pun_EmpInfo treatment is also significantly lower than that in the

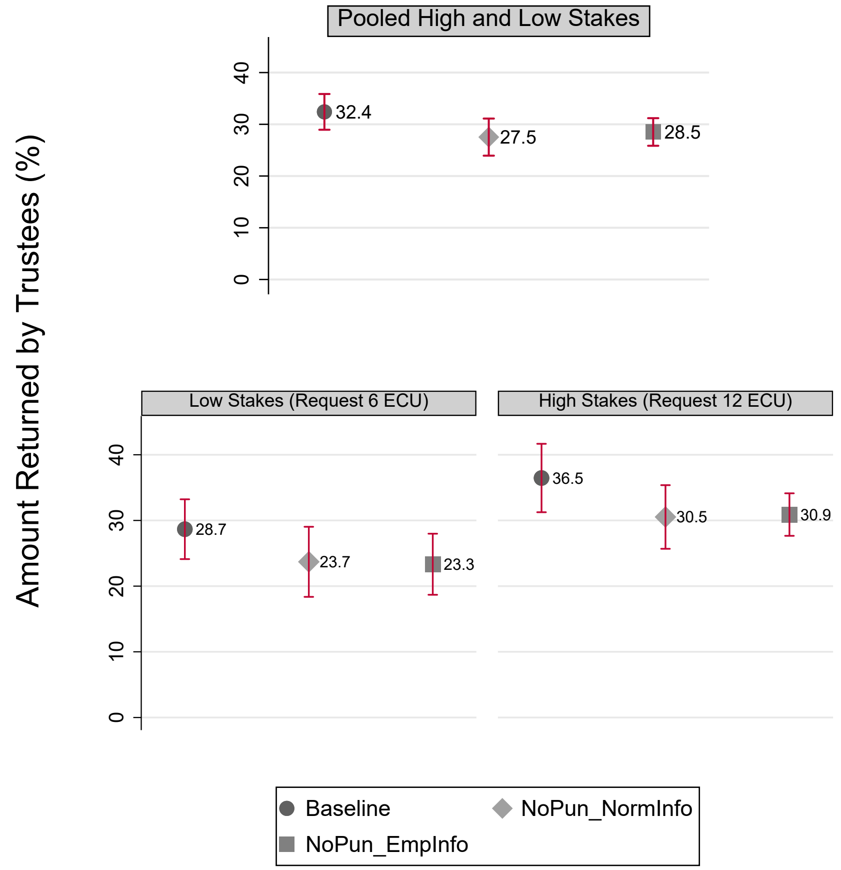


Figure 4: Amounts returned by trustees as percentages of amounts received from investors; upper part indicates pooled amounts; lower part indicates amounts per LS vs. HS; Baseline: no punishment or norm information; NoPun_NormInfo: no punishment, with normative information. NoPun_EmpInfo: no punishment, with empirical information. None of the comparisons are significant at the conventional levels. Whiskers represent 95% CIs.

Pun_NormInfo treatment (32.1% vs. 42.7%, BSM, $p<0.01$).[19] These results support Prediction 1: punishment is more effective when combined with normative information (about a socially disapproved behavior) than enforced alone or combined with empirical information (about a majority compliant behavior). Of particular interest, normative information plays only a negligible role in HS: the return rate in the Pun_NormInfo treatment is not significantly different from that in the Baseline and Pun_NoInfo treatments (31.6% vs. 36.5%,

[19] It should also be noted that the return rate in the Pun_EmpInfo is very close to that in the Punishment-NoInfo condition (32.1% vs. 34.4%, BSM, p=0.43).

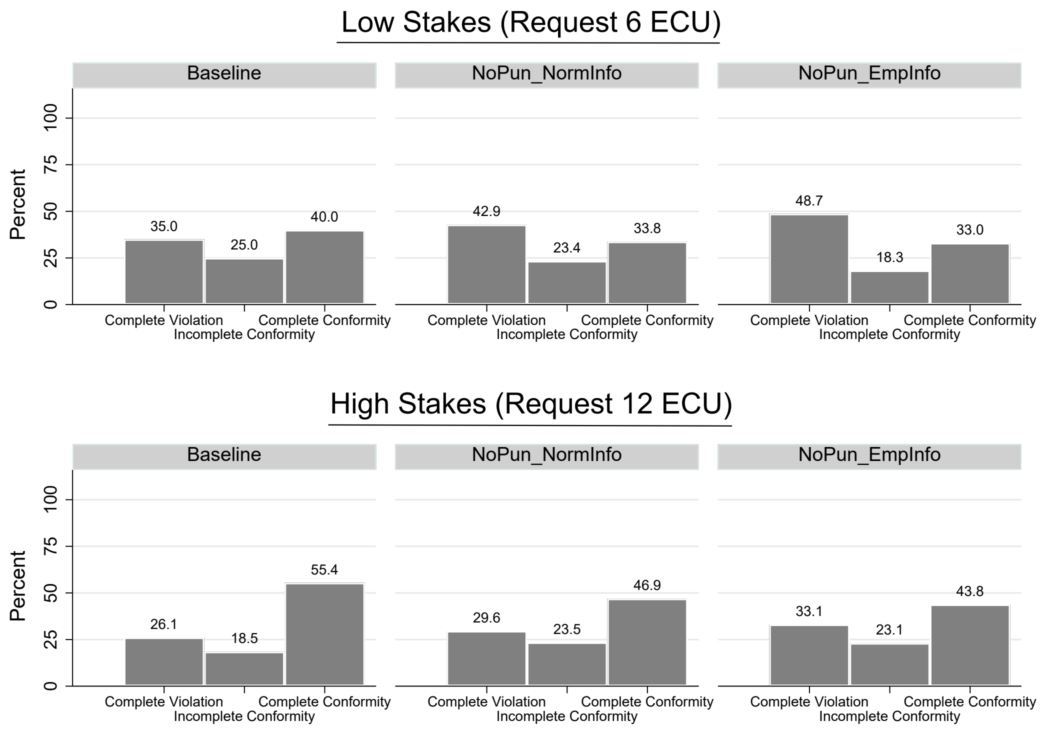


Figure 5: Distribution of return types in the Baseline (NoPun_NoInfo), NoPun_NormInfo, and NoPun_EmpInfo treatments.

BSM, p=0.18; 31.6% vs. 36.8%, BSM, p=0.18). Moreover, adding empirical information significantly decreases return rates as compared to those in the Baseline and Pun_NoInfo treatments (22.2% vs. 36.5%, BSM, p=0.01; 22.2% vs. 36.8%, BSM, p=0.01).

Here, we highlight the results as they relate to Hypothesis 1, which suggests that the observed effects of punishment combined with normative or empirical information are not due to the normative or empirical information alone, but due to their combination with punishment.[20] Compared to Pun_NormInfo, NoPun_NormInfo leads to lower conformity rates when pooled across stakes (27.5% vs. 36.3%, BSM, p<0.01). Consistent with the discussion of Hypothesis 2, the difference is mainly driven by the LS condition (23.7% vs. 42.7%, BSM, p<0.01). The difference in the HS condition is not significant (30.5% vs.

[20] For brevity, the full comparisons of the average return in all the six treatments are illustrated in Figure A.1 in the Appendix.

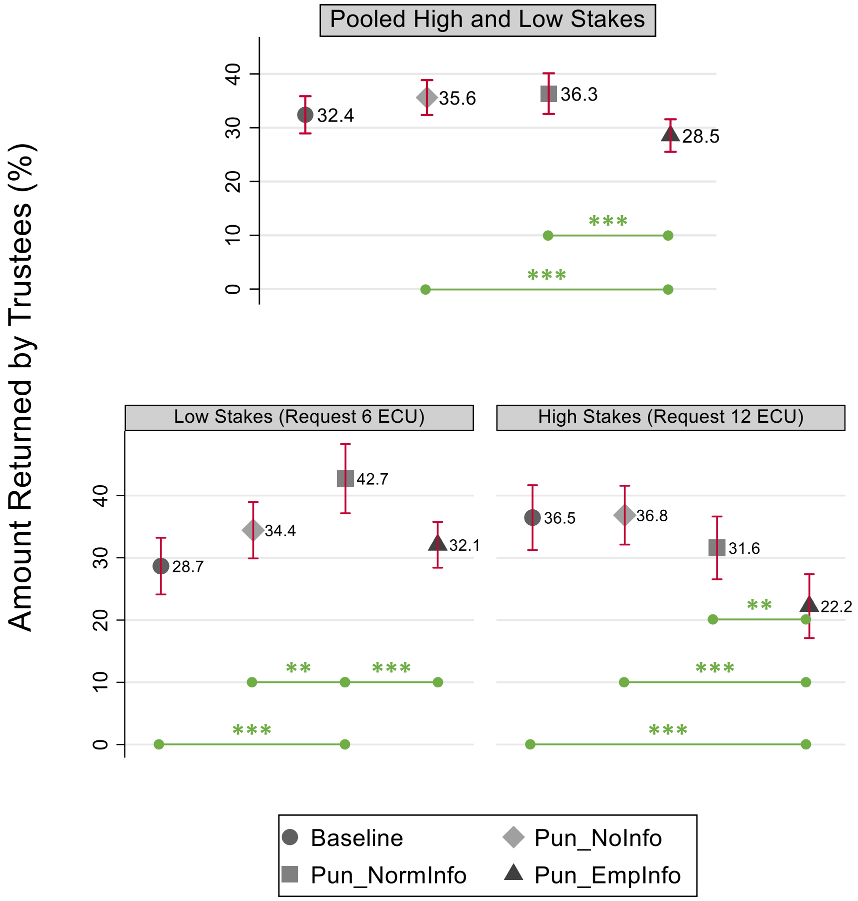


Figure 6: Amounts returned by trustees as percentages of amounts received from investors; upper part indicates pooled amounts; lower part indicates amounts per LS vs. HS; Baseline: no punishment or norm information; Pun_NoInfo: punishment (5 ECU) without norm information; Pun_NormInfo: punishment (5 ECU) and normative information; Pun_EmpInfo: punishment (5 ECU) and empirical information. Significant differences are indicated at the conventional levels of *p<0.1, **p<0.05, ***p<0.01. Whiskers represent 95% CIs.

31.6%, BSM, p=0.77). These results suggest that the significant effect of Pun_NormInfo on returns cannot be attributed to the normative information alone.

When comparing the Pun_EmpInfo and the NoPun_EmpInfo treatments, we observe a significant increase in conformity for the former in LS (32.1% vs. 23.3% , BSM, p<0.01). As we reported above, the positive effect of Pun_EmpInfo cannot be attributed to punishment alone. On the other hand, empirical information combined with punishment backfires in HS; specifically, we observe a significant decrease in the conformity rate (30.9% vs. 22.2%, BSM, p<0.01). As a result, there is no significant difference between the two treatments when data is pooled (28.5% vs. 28.5%, BSM, p=0.95).[21] These results suggest that the stakes, which directly affect the cost of conformity, and the kind of information (empirical or normative) influence the benefit of combining punishment with norm information. Consistent with our hypothesis, normative information is helpful, but its supplemental effect is moderated by the stakes. When the cost is high, neither normative nor empirical information improves the efficacy of punishment. Surprisingly, empirical information alone proves counterproductive when the cost is high (e.g, it decreases return rates).

To further understand these results, we plot the return distribution in Figure 7. The return patterns in the LS condition reveal significant dissimilarities between the Baseline and the Pun_NormInfo and Pun_EmpInfo treatments (K-S, p<0.01), the latter of which uncovers distinctive bimodal distributions with a significant decrease in Incomplete Conformity (25.0% vs. 2.3%, BSM, p<0.01; 25% vs. 2.9%, BSM, p<0.01). Compared with the Baseline treatment, the Pun_NormInfo treatment sees a significant increase of Complete Conformity (40.0% vs. 77.0%, BSM, p<0.01) and a substantial decrease of Complete Violation (35.0% vs. 20.7%, BSM, p<0.01). Such a significant shift in Pun_NormInfo cannot be attributed to punishment alone: if we compare the Pun_NormInfo and Pun_NoInfo treatments, we observe that the former exhibits a higher rate of Complete Conformity (77% vs. 66.4%, BSM, p=0.03) and a lower rate of Complete Violation (20.7% vs. 33.6%, BSM, p<0.01). These results show that normative information enhances the effectiveness of punishment by increasing the rate of complete conformity while reducing complete violation rates. Such an enhancement does not occur with empirical information.

Continuing with the analysis of LS, while Pun_EmpInfo offers significant increases in Complete Conformity (62.1% vs. 40.0%, BSM, p<0.01) over the Baseline treatment, this is very close to what we observe in the Pun_NoInfo treatment (62.1% vs. 66.4, BSM, p=0.48). This implies that the main effect results from punishment, which is corroborated by the substantially smaller amount of complete conformity in NoPun_EmpInfo (48.7%) as indicated in Figure 5. We find no significant change in Complete Violation in Pun_EmpInfo compared to the Baseline treatment (35.1% vs 35.0%, BSM, p=0.74); or in Pun_EmpInfo compared to the Pun_NoInfo treatment (35.1% vs. 33.6%, BSM, p=0.66).

[21] Further support for our results is illustrated in Figure A.2 in the Appendix in which we plot return behavior across treatments over all 10 periods. We can observe that – compared to the Baseline – the direction of the result for Pun_NormInfo in LS and the result for Pun_EmpInfo in HS persists.

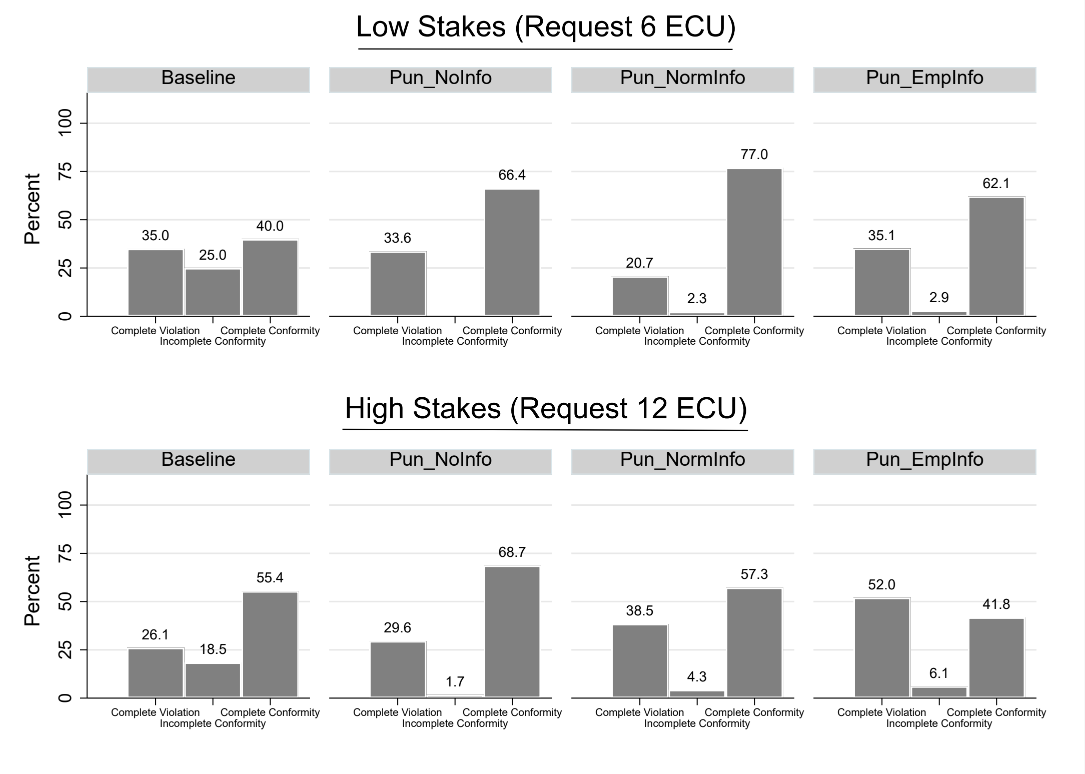


Figure 7: Distribution of return types in Baseline (NoPun_NoInfo), Pun_NoInfo, Pun_NormInfo, and Pun_EmpInfo treatments.

In sum, when the stakes are low, the return patterns across treatments are consistent with Hypothesis 1. Punishment can more effectively promote reciprocity by making the fact that returning less than the requested amount is socially disapproved salient. The interaction of information and punishment is particularly effective when the former is normative. As seen from the average returns, when the stakes are high, the benefit of both types of information is much less evident and empirical information is even detrimental. This implies that Hypothesis 1 holds for LS but not HS.

Figure 7 further reveals that the detrimental effect observed in the Pun_EmpInfo treatment in the HS condition is mainly driven by the significant increase in Complete Violation over the Baseline (52.0% vs. 26.1%, BSM, p<0.01). At the same time, we only observe a marginally significant increase in Complete Violation in the Pun_NormInfo compared to the Baseline treatment (38.5% vs. 26.1%, BSM, p=0.06). Additionally, Complete Conformity is marginally less frequent in the Pun_EmpInfo than in the Baseline treatment (41.8% vs. 55.4%, BSM, p=0.06), whilst such a negative shift does not occur in Pun_NormInfo-

Baseline (Complete Conformity: 57.3% vs. 55.4%, BSM, p=0.58). We reported in Section 3.1 and observe again in Figure 7 that there is no significant negative shift in Complete Conformity when comparing the Baseline and the Pun_NoInfo treatments.

In sum, we have shown that the combination of punishment and norm information can produce effects that are not uniform. By focusing on its effectiveness compared to the Baseline, we show that the detrimental outcomes of combining punishment with empirical information mainly show up in the high stakes condition. This *backfiring* occurs in two ways: first, it substantially increases the fraction of complete violators at the expense of complete conformists (see the bottom panel of Figure 7). Second, it produces lower average returns in the high stakes condition, which also leads to an overall average decrease when pooled across low and high stakes (see the top and bottom panel of Figure 6).

As we show, these results suggest that the detrimental effect in HS of the Pun_EmpInfo condition is mainly due to the addition of generic empirical information, rather than the punishment itself. Note that in HS, punishment alone hardly affects conformity whereas adding norm information decreases conformity, significantly so if the information is empirical. Since compliance is more costly in HS than in LS, there is an inherent tension between selfish behavior and conformity. To resolve the tension, one may exploit some wiggle room created by the generic information by forming a self-serving belief ("only individuals in the low-stakes condition followed the rule"). This is made possible by the lack of specificity about the relevant reference group in the empirical message. When conformity is cheap (LS) we do not see this effect. Existing evidence indicates that empirical information, but not normative information, gives rise to motivated beliefs and (self-serving) belief distortion to justify non-compliance (Kunda, 1990; Bénabou and Tirole, 2016; Gino et al., 2016; Bicchieri et al., 2020; Dimant, 2020). Moreover, studies show that investor's intentions play a role in the decision to reciprocate (Toussaert, 2017; Orhun, 2018). People may feel justified to pursue self interest when threatened by punishment combined with empirical information that is perceived as unapplicable. We return to this point in Section 5.

4.4. Regression Results

We analyze our data using different variants of multivariate regressions that examine the robustness of our results using a host of control variables.[22] In all cases, we employ random

[22] These include gender, a measure for self-control taken from Tangney et al. (2004), and a measure for risk taken from Dohmen et al. (2011). Although we report the coefficients in our regression tables, we will not

effects panel regressions with standard errors clustered at the participant level.[23] As Table 2 indicates, the examination of average return behavior across treatments yields three main results indicating that our previous findings are robust to the inclusion of various controls.The results are as follows:

DV: Amount Returned by Trustee (%)	**Low Stakes**		**High Stakes**	
	(1)	(2)	(3)	(4)
Treatment				
(Base Level: Baseline)				
Pun_NoInfo	6.108	5.191	-2.154	-3.151
	(5.388)	(5.853)	(5.685)	(5.940)
NoPun_NormInfo	-8.938	-9.543	-7.592	-7.727
	(5.750)	(6.182)	(5.948)	(5.961)
Pun_NormInfo	13.071**	13.537**	1.327	0.711
	(5.664)	(6.054)	(6.477)	(6.640)
NoPun_EmpInfo	-6.793	-7.640	-3.504	-4.388
	(5.193)	(5.586)	(5.374)	(5.472)
Pun_ EmpInfo	1.520	2.404	-10.299*	-12.308**
	(5.051)	(5.432)	(5.712)	(5.870)
Round	-0.636***	-0.486*	-0.340*	0.022
	(0.237)	(0.248)	(0.203)	(0.205)
Gender	-0.443	0.450	3.674	3.899
	(3.289)	(3.434)	(3.676)	(3.762)
Self-Control	3.886**	4.331***	4.051**	4.062**
	(1.612)	(1.677)	(1.829)	(1.868)
Risk	0.321	0.241	0.113	0.196
	(0.694)	(0.731)	(0.808)	(0.833)
L1.Amount Received from Investor		0.004		0.041
		(0.075)		(0.041)
Constant	32.599***	31.607***	34.050***	31.236***
	(5.543)	(6.200)	(6.352)	(6.484)
Observations	675	567	771	694

Table 2: Random effects model with robust standard errors (in parentheses) clustered on the participant level. Estimations only for periods in which return request message was sent. Control variables include stakes (1 = high), Round (1-10), Gender (1 = male), Self-Control (higher number indicates more self-control, standardized measure), Risk (higher number indicates more risk-seeking, standardized measure). L1.Amount Received from Investor (% amount received from an investor in previous round, which indicates whether trustee faced a high- or low-stakes situation). Significance levels: *p<0.10, **p<0.05, ***p<0.01.

Result 1: Neither punishment nor norm information alone significantly affect return behavior. This remains statistically supported across the stakes faced by trustees.

discuss them in further detail because these control variables are merely used to establish the robustness of our previously discussed unconditional results.

[23] In that we follow the previously motivated literature. A small number of sessions does not allow us to cluster standard errors at the session level. Other challenges with alternative clustering methods (and on session level in particular) are discussed by Cameron et al. 2008; Fréchette 2012; Abadie et al. 2017.

Result 2: The combination of punishment and normative information successfully increases return rates, but only when compliance is cheap. The increase is substantial and about 13% higher than the Baseline.

Result 3: The combination of punishment and empirical information triggers a substantial backlash in return behavior, but only when conformity is very costly. The reduction amounts to 10% to 12% relative to the Baseline condition.

The insignificant coefficient for previous round's investor behavior indicates that the possibility of learning is limited at best, which supports our methodological choice of random partner-rematch across rounds. Note that all results are robust even after the inclusion of the 7% of data in which investors did not send a return request message (see Table A1 in the Appendix). We provide a more detailed analysis of the drivers of trustee behavior across treatments in Table A2 in the Appendix. In conclusion, our regression results support the robustness of our previous analyses.

5. Understanding the Mechanisms: Two Follow-Up Experiments

We ran two follow-up experiments (total n = 475) with separate groups of participants in the summer of 2020: a *norm elicitation* experiment and an *information credibility* experiment. In the first experiment, we test the assumptions we made in the theoretical analysis. In the analysis, we derive the conditions for the Pun_Info treatments to increase the return from the assumption that in the Baseline the trustees think the acceptable return amount is less than 50% (but not less than the original amount sent). We test if this assumption holds when we elicit normative and empirical expectations about reciprocating behavior.

In the second experiment, we try to understand *why* the Pun_Info treatments do not achieve a higher return than the Baseline in the high-stakes condition and why the Pun_EmpInfo treatment even backfires in that condition. Note, again, that in the original Trust game, players received generic information about average behavior, without the low- and high-stakes distinction. We hypothesized that the uncertainty about the relevant reference network led the trustees to interpret the information differently in the two conditions. In particular, given the high cost of returning 50% when the investor transferred all the endowment, trustees may think the empirical information about 50% return is mostly driven by the behavior in the low-stakes case. In this second experiment, we thus ask about how credible the generic information is.

Norm Elicitation Experiment

In our first experiment, we follow Bicchieri and Xiao (2009) and Bicchieri and Chavez (2010) to measure the existence of a social norm and examine whether individuals hold sufficiently high and consistent expectations about what other people do (empirical expectations) and what other people think one ought to do (normative expectations) in the context of our Trust Game, separately for low- and high-stakes. We drew participants from the same participant pool as our original experiment and collected data from 178 University of Pennsylvania students, none of whom previously participated in the original experiment.[24] The experiment consisted of three parts (the order of parts (ii) and (iii) was randomized):

(i) Description of the original Trust game laboratory experiment

(ii) Elicitation of three beliefs for low-stakes behavior

(iii) Elicitation of three beliefs for high-stakes behavior

After receiving an explanation of the original Baseline version of the experiment (no inclusion of punishment or norm-based information), participants were placed in the role of Player 2 (the trustee) and were asked to express three beliefs for each of two parts (ii) and (iii); thus, 6 beliefs in total.[25] Within each part, the three sets of beliefs consisted of:[26]

1. **Personal normative beliefs (separately for high/low stake examples):** "Please chose the option corresponding to what you think one should do in the role of P2."
2. **Empirical Expectations (separately for high/low stake examples):** "Please guess what you believe the most frequent choice P2s made in the experiment."
3. **Normative Expectations (separately for high/low stake examples):** "We have asked all participants in this survey what they believe P2 should do. Please guess what you believe is the most frequent answer other participants gave about what they believe P2 should do."

[24] Participants received a show-up fee of $2 and received up to $4 in additional bonus payments. The average duration of participation was 10 minutes, yielding an average hourly pay of about $21.

[25] In the original repeated game, players experienced both high- and low-stakes conditions. As in the present experiment, players knew both conditions could occur.

[26] Incentivization of empirical expectations was based on actual behavior of participants in the main experiment. Incentivization of normative expectations was based on the stated personal normative beliefs of the other participants in this experiment. As is customary in this literature, we use this natural order to elicit and incentivize beliefs. d'Adda et al. (2016) show that the order in which norm-related beliefs are elicited has no effect on their validity.

For each of the questions, participants had four discrete choices that mirrored our previous analyses: returning nothing, returning a non-zero amount but less than what Player 1 sent, returning at least what Player 1 sent but less than half of the tripled amount, returning at least half of the tripled amount. We present the results in Figure 8 below.[27]

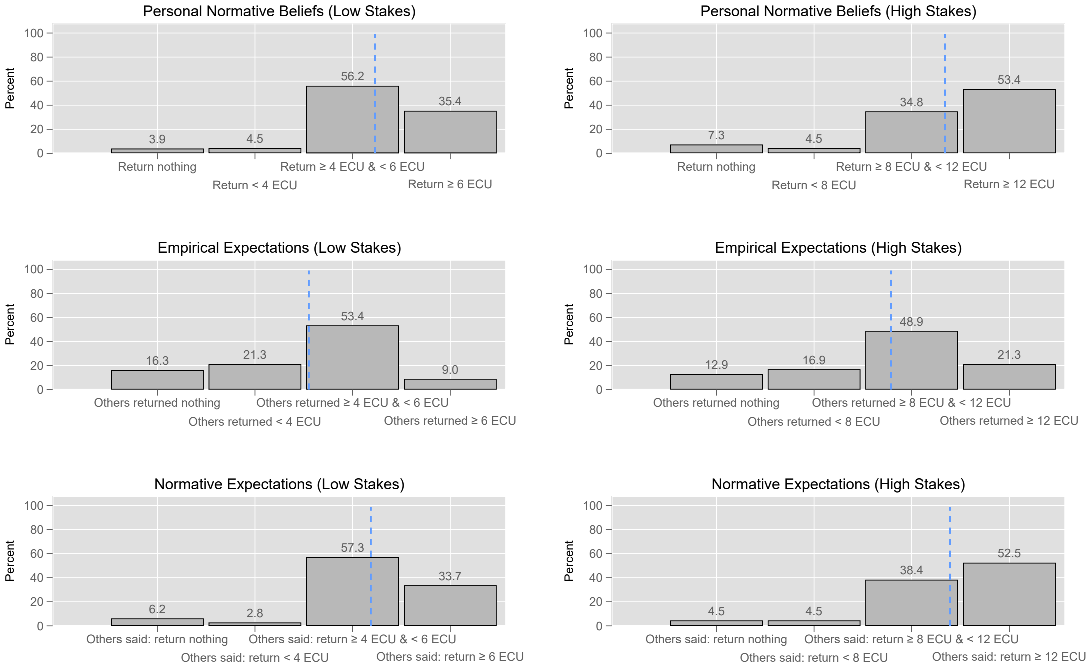


Figure 8: Distribution of personal normative beliefs, empirical expectations, and normative expectations across low-stakes and high-stakes conditions. Vertical blue lines represent averages. Consistent with the naming convention in our main experiment, the label "≥ 6 ECU" ("≥ 12 ECU") corresponds to returning at least half of the requested tripled amount in the low stakes (high stakes) condition.

We use the norm elicitation results presented in Figure 8 to shed more light on the previous results from the Trust Game, in particular those discussed in Figures 2, 4, and 6. As we

[27] It is interesting to observe how normative expectations are lower in the low-stakes than in the high-stakes condition. In the low-stakes, few people seem to think that Player 2 should give back at least half of the tripled amount, likely considering that the investor did not show excessive trust towards the recipient to begin with. On the contrary, in the high-stakes condition, both normative expectations and personal normative beliefs are much higher, indicating that since the investor gave all, he/she deserves to receive at least half of the tripled amount. If we also consider the relationship between empirical and normative expectations, we notice that they are highly consistent in the low-stakes condition, and not in the high-stakes condition. Here, the normative conviction that it is appropriate to give half or more of the tripled amount is at odds with what is in fact expected to occur.

already noted, the beliefs of low-stakes and high-stakes participants are quite different. In the low-stakes condition, all three beliefs point to a social norm of returning at least the received amount but less than half of the tripled amount (the majority of participants chose that option across all three belief elicitations). These beliefs are consistent with the assumptions we made in the the theoretical analysis. However, this is not the case for the high-stakes conditions: here, both personal normative beliefs and normative expectations signal that the majority of participants believe that one should return at least half of the tripled amount.[28] Yet, this is not what participants believe other participants would actually do, which – in light of our behavioral results – is the correct expectation. Such inconsistency in expectations suggests that a social norm exists (high and consistent normative expectations) but is not expected to be followed under high-stakes.[29] This is not uncommon, as a social norm may exist but not be followed at a given time. This happens because, as opposed to moral norms, social norms involve *conditional preferences*,i.e., the preference for compliance depends upon having sufficiently high (and consistent) empirical *and* normative expectations. When one of these expectations is low, one is justified in not following the norm Bicchieri (2006).[30]

Since the players in the norm-elicitation experiment are drawn from the same student population of the original Trust Game, we can assume that players' expectations in the original game were similar to participants' elicited expectations in the new experiments. We may conclude from that experiment that most low-stakes players in the original game held a strong norm of minimal reciprocity (returning 4, but less than 6), consistent with the assumption in our theoretical analysis. Providing normative information of returning at least 50%, according to our theoretical analysis, can potentially have a positive effect on behavior. Such an effect is most likely to be observed when the normative message is combined with punishment. This combination may lead players to update their original normative expectations in addition to the (small) monetary cost imposed on deviation. The null effect of the empirical message in the low-stakes condition indicates that empirical

[28] A high normative expectation in the Baseline is the likely reason why we did not observe a positive effect of Pun_NormInfo in the HS condition. If trustees already hold the belief that one should return more than 50% in the Baseline, adding a message that is consistent with the belief should not have any impact.

[29] When comparing the differences between low- and high-stakes, the results from Wilcoxon matched-pairs signed-ranks tests suggest that there are significant differences at the 1% level for all three beliefs.

[30] There are two cases in which a norm deviation disappears: either we are not focused on the norm (Cialdini et al., 1990), or the norm is suspended, as when we do not expect people to conform.

information is not as effective in changing people's normative expectations. In the high-stakes condition, normative expectations indicate that players believe the investor deserves to get at least half of the tripled amount since he/she gave all of the money (8 ECU). This suggests that, inconsistent with the assumption in the theoretical analysis, players have already held normative expectations in the Baseline that are consistent with the normative message. Therefore, the combination of punishment and normative information yielded no behavior change (Figure 6), as there was no reason to update normative expectations.

The puzzling result is that when empirical information is combined with punishment in the high-stakes condition, there is a significant decline in returns. Note that as in the low-stakes condition, the empirical expectation in the high-stakes condition is also to return less than 50%. We expect that the empirical information may not effectively update the player's normative expectations, but we did not predict a backfire effect. We speculate that the high cost of returning 50% in the high-stakes condition may motivate players to process the information in a self-serving manner. For example, a trustee can discount the empirical information as irrelevant to their situation, rendering punishment unjustied. As the message does not make it clear whether the 50% return occurs in both low- and high-stakes conditions or just one of them, we suspect that trustees may interpret the information as mostly relevant to the low-stakes condition. When punishment is combined with a message that is viewed as irrelevant to the decision context, punishment may be perceived as unjustified and thus backfires. This speculation would be in line with existing research showing that trust in messages and their effectiveness in changing behavior are closely connected (e.g., Gifford et al., 2018).

Information Credibility Experiment

In our second follow-up experiment (n=297 collected from Amazon Mechanical Turk), we find evidence consistent with this possibility.[31] After explaining the original game to the participants, we gave them either the original empirical or normative information (between-subjects design), after which we elicited their beliefs about how credible they found this information, on a scale from 0-10. We asked this question for low-stakes and high-stakes (within-subject design, random order). The results from this experiment show that there is a significant difference in the credibility of the empirical information in the high-stakes

[31] We turned to data collection on MTurk due to the inability to use physical labs during the COVID-19 pandemic. Reassuringly, recent literature indicates that the findings on MTurk are robust, generalizeable, and reproducible (Coppock et al., 2018; Snowberg and Yariv, 2020).

condition compared to the low-stakes condition (Wilcoxon matched-pairs signed-ranks test p<0.01; McNemar's chi-squared p=0.025), whereas the credibility of the normative information remains unchanged (p=0.95; McNemar's chi-squared p=0.99).[32] The high cost of compliance can be a reason for the low credibility assessment in the empirical information condition and, in our original experiment, may even lead trustees to suspect that the investor had *intentionally* exploited the ambiguity of the information.

6. Conclusion

The evidence presented in this paper shows that weak punishment can be effective when combined with norm information. This combination serves to legitimize punishment, linking it to what is considered socially appropriate behavior. However, not all norm information is equally helpful, and the cost of compliance may act as a moderator of the potentially positive effect of the combination. Norm information may be conveyed as an empirical message (what others do) or as a normative one (what others approve/disapprove of). Intuitively, one may expect the punishment/normative message combination to be the strongest motivator of pro-social behavior, since the normative message unequivocally signals what the right action is, as opposed to information about how frequent an action is. Cost of compliance, however, moderates this expected effect. With low cost (low-stakes), the combination of normative information and punishment significantly raises the rate of return compared to the Baseline (no punishment and no norm information), punishment alone, and normative information alone. With high compliance cost (high-stakes), however, we find no significant effect of the combination of normative information and punishment.[33]

The effect of combining punishment and empirical information is not as strong, and may even be negative. Note that empirical information, in our experiment, did not differentiate between the behavior of different groups (high and low-stakes). As we discuss

[32] Note that the average credibility score across all conditions was very high to begin with: around 7 on a scale from 0-10. This emphasizes that the participants had no doubt about the truthfulness of the norm information in general, but were simply less convinced in the high-stakes empirical information condition.

[33] In the context of a Trust game, this differential response is intuitive: the high stakes condition is one in which the investor sent all the money, and the trustees' response is already positive even in the absence of punishment. Adding normative information does not change much. The low stakes condition is one in which the investor only sent half of the money, an action that might be interpreted as lack of trust, and in this case the Baseline response is weaker if compared with the high-stakes response. We support this explanation with a second experiment in which we show that the initial normative expectations of the high-stakes group are already consistent with the message, whereas those of the low-stakes group are not, and can thus be (positively) changed by the message they receive.

above, ambiguity about the reference group may induce motivated reasoning, i.e. an interpretation of the information that favors selfish behavior. In the low stakes condition, the reported frequency of compliant behavior is credible if referred to low-stake trustees, and thus this combination does not induce behavior significantly different from the Baseline. Interestingly, we find that in high-stakes condition the combination of empirical information and punishment can have a detrimental effect by significantly decreasing the rate of return compared to Baseline and punishment alone. A reason for the negative effect of the combination of punishment and empirical information in the high-stakes group is the belief that punishment is unjustified, since trustees may easily believe that, given that the information does not differentiate between reference groups, compliance may be limited to the low-stakes condition, where it is much cheaper. In a third experiment we test our hypothesis about the different credibility of the empirical message and found that, indeed, it is less credible for the high-stakes condition. The negative reaction may thus be due to a combination of low credibility and seeing the investor as having *intentionally* exploited the ambiguity of the reference group information.

There is mounting interest in applying social norm methods to enhance nudge interventions (OECD, 2015; Miller and Prentice, 2016; Reijula et al., 2018). Our findings suggest that norm-based interventions can lead to significant improvements but can also backfire, even if the norm is embodied in a cooperative context and clearly stated (as opposed to when the state of the world is left uncertain, as is the case in Bicchieri et al. 2020) An important insight for policymakers is to avoid sending empirical information that is inconsistent with what people already believe to be true and accompany this information with negative sanctions, especially when compliance is costly.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Allcott, H. and Mullainathan, S. (2010). Behavior and Energy Policy. *Science*, 327(5970):1204–1205.

Anderhub, V., Engelmann, D., and Güth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization*, 48(2):197–216.

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., and Villatoro, D. (2013). Punish and voice: punishment enhances cooperation when combined with norm-signalling. *PloS one*, 8(6):e64941.

Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature communications*, 7:13327.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C. and Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, 23(2):161–178.

Bicchieri, C. and Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*.

Bicchieri, C., Dimant, E., Gaechter, S., and Nosenzo, D. (2019). Social proximity and the erosion of norm compliance. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3355028.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2020). It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3326146.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Binmore, K. and Shaked, A. (2010). Experimental economics: Where next? *Journal of Economic Behavior & Organization*, 73(1):87–100.

Bolton, G., Dimant, E., and Schmidt, U. (2020). Observability and social image: On the robustness and fragility of reciprocity. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294375.

Bott, K. M., Cappelen, A. W., Sorensen, E., and Tungodden, B. (2019). You've got mail: A randomised field experiment on tax evasion. *Management Science*.

Bursztyn, L., Egorov, G., and Fiorin, S. (2019). From extreme to mainstream: How social norms unravel. NBER Working Paper.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2018). Misperceived social norms: Female labor force participation in saudi arabia. NBER Working Paper.

Camerer, C. (1997). Rules for experimenting in psychology and economics, and why they differ. In *Understanding Strategic Interaction*, pages 313–327. Springer.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Charness, G., Gneezy, U., and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1–8.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015.

Cooter, R. (1998). Expressive law and economics. *The Journal of Legal Studies*, 27(S2):585–607.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Dal Bó, P. and Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–29.

Dimant, E. (2020). Hate trumps love: The impact of political polarization on social preferences. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3680871.

Dimant, E., Gerben, A. v. K., and Shalvi, S. (2019). Requiem for a nudge: Framing effects in nudging honesty. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3416399.

Dimant, E. and Gesche, T. (2020). Nudging enforcers: How norm perceptions and motives for lying shape sanctions. Working Paper Available at SSRN: https://ssrn.com/abstract=3664995.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

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

Ensminger, J. and Henrich, J. (2014). *Experimenting with social norms: Fairness and punishment in cross-cultural perspective.* Russell Sage Foundation.

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137.

Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *American Economic Review*, 101(3):318–22.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fréchette, G. R. (2012). Session-effects in the laboratory. *Experimental Economics*, 15(3):485–498.

Gifford, R., Lacroix, K., and Chen, A. (2018). Understanding responses to climate change: Psychological barriers to mitigation and a new theory of behavioral choice. In *Psychology and Climate Change*, pages 161–183. Elsevier.

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, 30(3):189–212.

Gneezy, U. and Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, 29(1):1–17.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3):472–482.

Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

Houser, D., Xiao, E., McCabe, K., and Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, 62(2):509–532.

Huck, S., Lünser, G. K., and Tyran, J.-R. (2012). Competition fosters trust. *Games and Economic Behavior*, 76(1):195–209.

Kahan, D. M. (1998). Social meaning and the economic analysis of crime. *The Journal of Legal Studies*, 27(S2):609–622.

Keane, L. D. and Nickerson, D. W. (2015). When reports depress rather than inspire: a field experiment using age cohorts as reference groups. *Journal of Political Marketing*, 14(4):381–390.

Kosfeld, M., Okada, A., and Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, 99(4):1335–55.

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic psychology*, 30(3):307–320.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3):480.

Markussen, T., Putterman, L., and Tyran, J.-R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *Review of Economic Studies*, 81(1):301–324.

Miller, D. T. and Prentice, D. A. (2016). Changing norms to change behavior. *Annual review of psychology*, 67:339–361.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for experimental economics*. Macmillan International Higher Education.

OECD (2015). Behavioral insights and new approaches to policy design. the views from the field. international seminar report. Technical report, OECD.

Orhun, A. Y. (2018). Perceived motives and reciprocity. *Games and Economic Behavior*, 109:436–451.

Reijula, S., Kuorikoski, J., Ehrig, T., Katsikopoulos, K., Sunder, S., et al. (2018). Nudge, boost, or design? limitations of behaviorally informed policy under social interaction. *Journal of Behavioral Economics for Policy*, 2(1):99–105.

Romaniuc, R., Dubois, D., Dimant, E., Lupusor, A., and Prohnitchi, V. (2020). Understanding cross-cultural differences in peer reporting practices: Evidence from tax evasion games in moldova and france. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3725208.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological science*, 18(5):429–434.

Snowberg, E. and Yariv, L. (2020). Testing the waters: Behavior across participant pools. Forthcoming American Economic Review.

Sunstein, C. R. (1996). On the expressive function of law. *University of Pennsylvania law review*, 144(5):2021–2053.

Tangney, J. P., Baumeister, R. F., and Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of personality*, 72(2):271–324.

Toussaert, S. (2017). Intention-based reciprocity and signaling of intentions. *Journal of Economic Behavior & Organization*, 137:132–144.

Tyler, T. R. (2006). *Why people obey the law.* Princeton University Press.

Tyran, J.-R. and Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *scandinavian Journal of Economics*, 108(1):135–156.

Xiao, E. (2018). Punishment, social norms, and cooperation. *Research Handbook on Behavioral Law and Economics*, page 155.

Xiao, E. and Houser, D. (2011). Punish in public. *Journal of Public Economics*, 95(7-8):1006–1017.

Appendix

A. Theoretical Model

Below we discuss the predictions separately by treatment and stakes: low-stakes (L) when half of the money was sent to Player 2 and high-stakes (H) when all of the money was sent to Player 2.

Baseline

It is straightforward that in this setting, a utility-maximizing trustee would not give more than $r^0_{Baseline}$. Previous studies have shown the average return rate in a trust game is often less than 50% (Camerer, 2011). Thus, we may assume that in the Baseline, a trustee thinks the acceptable return amount can be equal or less than 50% and no less than the investment amount (4 or 8 ECU). This is also why we design the request to be 50% so we can examine the effect of explicit norm information. Thus, in the low-stakes case, $4 \leq r^0_{Baseline\_L} \leq 6$. Conversely, in the high-stakes case, one obtains $8 \leq r^0_{Baseline\_H} \leq 12$. It is straightforward to see that:

Case 1: Low-Stakes

$$\begin{cases} \mathrm{r}^*_{Baseline\_L} = r^0_{Baseline\_L} \leq 6, & \text{if } k > 1 \\ \mathrm{r}^*_{Baseline\_L} = 0, & \text{if } k < 1 \end{cases}$$

Case 2: High-Stakes

$$\begin{cases} \mathrm{r}^*_{Baseline\_H} = r^0_{Baseline\_H} \leq 12, & \text{if } k > 1 \\ \mathrm{r}^*_{Baseline\_H} = 0, & \text{if } k < 1 \end{cases}$$

Thus, in the Baseline, the norm-sensitive agents (k>1) will comply with the norm and return r^0. Meanwhile, norm-insensitive agents (k<1) will return 0.[34]

Pun_NoInfo Treatment

Recall that punishment imposes a cost of 5 ECU if the trustee returns less than 50% of the tripled amount. It is straightforward that a utility-maximizing trustee would not give more than the amount that is enforced by the punishment. Similar to the Baseline, we assume that a trustee thinks the acceptable return amount is less than 50% and no less than the investment amount (4 or 8 ECU). Thus, in the low-stakes case, $4 \leq r^0_{\text{PunNoInfo\_L}} \leq 6$. Conversely, in the high-stakes case one obtains $8 \leq r^0_{PunNoInfo\_H} \leq 12$. With this, we can derive different predictions for low- and high-stakes scenarios:

[34] For the special case of k=1, it is easy to see that an individual is indifferent.

Case 1: Low-Stakes

If the trustee complies with the punishment (r=6) then: U=12-r-0=6. However, if the trustee does not comply with the punishment and returns less than 6 (again, the trustee will not return more than r^0), the trustee solves the following maximization problem:

$$\max_r U = 12 - r - 5 - k \times [12 - r - (12 - r^0_{PunNoInfo\_L})] = 12 - 5 - r \times (1 - k) - k \times r^0_{PunNoInfo\_L} \quad (3)$$

We get:

$$\begin{cases} \mathrm{r}^* = 0, & \text{if } k < 1 \\ \mathrm{r}^* = r^0_{PunNoInfo\_L}, & \text{if } k > 1 \end{cases}$$

By comparing the utility of compliance and non-compliance, we find:

$$\begin{cases} \mathrm{r}^*_{PunNoInfo\_L} = 6, & \text{if } k > \frac{1}{r^*_{PunNoInfo\_L}} \\ \mathrm{r}^*_{PunNoInfo\_L} = 0, & \text{if } k < \frac{1}{r^*_{PunNoInfo\_L}} \end{cases}$$

Case 2: High-Stakes

If the trustee complies with the punishment (r=12) then: U=24-r-0=12. However, if the trustee does not comply with the punishment and returns less than 12 (again, the trustee will not return more than r^0), the trustee solves the following maximization problem:

$$\max_r U = 24 - r - 5 - k \times [24 - r - (24 - r^0_{PunNoInfo\_H})] = 24 - 5 - r \times (1 - k) - k \times r^0_{PunNoInfo\_H} \quad (4)$$

We get:

$$\begin{cases} \mathrm{r}^* = 0, & \text{if } k < 1 \\ \mathrm{r}^* = r^0_{PunNoInfo\_H}, & \text{if } k > 1 \end{cases}$$

By comparing the utility of compliance and non-compliance, we find:

$$\begin{cases} \mathrm{r}^*_{PunNoInfo\_H} = 12, & \text{if } k > \frac{7}{r^*_{PunNoInfo\_H}} \\ \mathrm{r}^*_{PunNoInfo\_H} = 0, & \text{if } k < \frac{7}{r^*_{PunNoInfo\_H}} \end{cases}$$

Comparing the Pun_NoInfo and Baseline conditions, we would expect punishment to enforce cooperation as long as there is a significant number of k> $\frac{1}{r^0_{PunNoInfo\_L}}$ in the low-stakes condition or k> $\frac{7}{r^0_{PunNoInfo\_H}}$ in the high-stakes one, respectively. Previous studies suggest that punishment alone can be detrimental. One reason proposed in the literature is that punishment changes people's perception of the decision environment from norm-based to profit-maximizing (i.e., one would pay to transgress, see Gneezy and Rustichini, 2000). This means in our framework, $r^0_{PunNoInfo} = 0$. If so, then $r^*_{PunNoInfo} = 0$

Two Pun_Info (Pun_NormInfo and the Pun_EmpInfo) Treatments
When the normative/empirical information is received, the trustee will believe that $r^0_{PunNoInfo\_L}$ $= 6$ in the low-stakes case and $r^0_{PunNoInfo\_H} = 12$ in the high conformity case. The trustee would not give more than $r^0_{PunInfo}$. The punishment imposes a cost of 5 ECU when the trustee returns less than requested. We obtain the following findings:

Case 1: Low-Stakes

$$\begin{cases} \mathrm{r}^*_{PunInfo\_L} = 6, & \text{if } k > \frac{1}{6} \\ \mathrm{r}^*_{PunInfo\_L} = 0, & \text{if } k < \frac{1}{6} \end{cases}$$

Case 2: High-Stakes

$$\begin{cases} \mathrm{r}^*_{PunInfo\_H} = 12, & \text{if } k > \frac{7}{12} \\ \mathrm{r}^*_{PunInfo\_H} = 0, & \text{if } k < \frac{7}{12} \end{cases}$$

Two NoPun_Info (NoPun_NormInfo and the NoPun_EmpInfo) Treatments
In these two treatments, participants receive only the normative/empirical information. As in the Pun_NormInfo and Pun_EmpInfo treatments, the trustees will believe that $r^0_{NoPunInfo\_L} = 6$ when the enforced amount is 6 and $r^0_{NoPunInfo\_H} = 12$ when the enforced amount is 12. However, unlike the above two punishment treatments, there is no monetary cost of returning less than 6. It is straightforward to find that in these two treatments:

Case 1: Low-Stakes

$$\begin{cases} \mathrm{r}^*_{NoPunInfo\_L} = 6, & \text{if } k > 1 \\ \mathrm{r}^*_{NoPunInfo\_L} = 0, & \text{if } k < 1 \end{cases}$$

Case 2: High-Stakes

$$\begin{cases} \mathrm{r}^*_{NoPunInfo\_H} = 12, & \text{if } k > 1 \\ \mathrm{r}^*_{NoPunInfo\_H} = 0, & \text{if } k < 1 \end{cases}$$

Taking all these together, we summarize below the conditions for each treatment to achieve a higher return than the Baseline condition:

Case 1: Low-Stakes

$$\begin{cases} \mathrm{r}^*_{PunNoInfo\_L} > r^*_{Baseline\_L}, & \text{if there is a significant number of individuals whose k>}\frac{1}{r^0_{PunNoInfo\_L}} \\ \mathrm{r}^*_{PunInfo\_L} > r^*_{Baseline\_L}, & \text{if there is a significant number of individuals whose k>}\frac{1}{6} \\ \mathrm{r}^*_{PunInfo\_L} > r^*_{PunNoInfo\_L}, & \text{if there is a significant number of individuals whose } \frac{1}{6}\text{<k<}\frac{1}{r^0_{PunNoInfo\_L}} \\ \mathrm{r}^*_{NoPunInfo\_L} > r^*_{Baseline\_L}, & \text{if there is a significant number of individuals whose k>1} \end{cases}$$

Note when k<$\frac{1}{6}$ then $r^*_{PunInfo\_L} = r^*_{PunNoInfo\_L} = r^*_{Baseline\_L} = 0$.

Case 2: High-Stakes

$$\begin{cases} \mathrm{r}^*_{PunNoInfo\_H} > r^*_{Baseline\_H}, & \text{if there is a significant number of individuals whose k>}\frac{7}{r^0_{PunNoInfo\_H}} \\ \mathrm{r}^*_{PunInfo\_H} > r^*_{Baseline\_H}, & \text{if there is a significant number of individuals whose k>}\frac{7}{12} \\ \mathrm{r}^*_{PunInfo\_H} > r^*_{PunNoInfo\_H}, & \text{if there is a significant number of individuals whose } \frac{7}{12}\text{<k<}\frac{7}{r^0_{PunNoInfo\_H}} \\ \mathrm{r}^*_{NoPunInfo\_H} > r^*_{Baseline\_H}, & \text{if there is a significant number of individuals whose k>1} \end{cases}$$

Note when k<$\frac{7}{12}$ then $r^*_{PunInfo\_H} = r^*_{PunNoInfo\_H} = r^*_{Baseline\_H} = 0$.

B. Robustness Checks and Additional Figures for Trustee Behavior

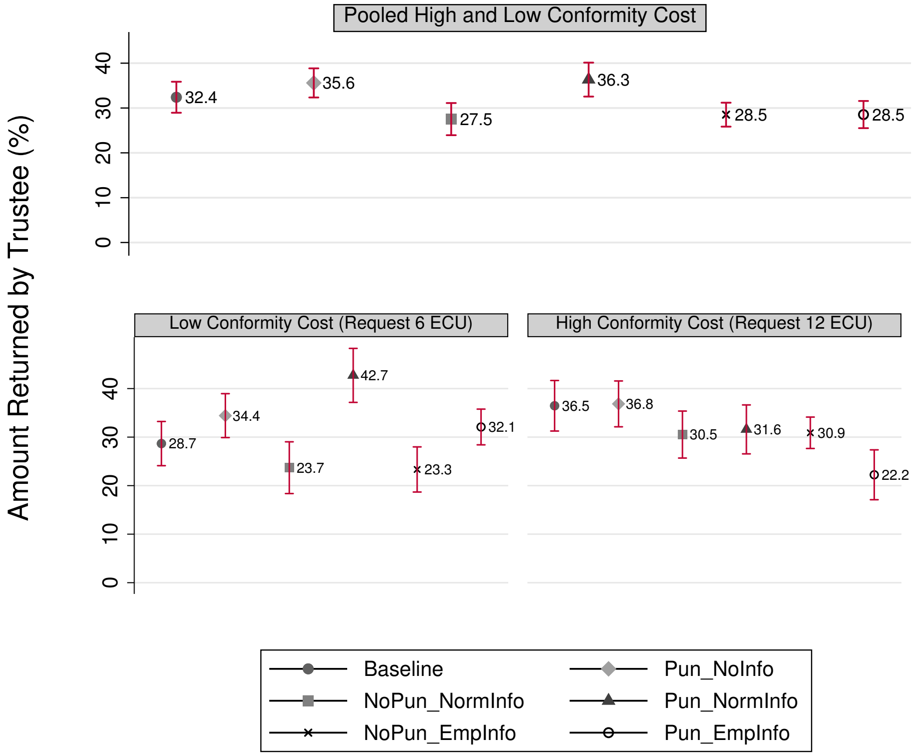


Figure A.1: Amounts returned by trustees as percentages of amounts received from investors; upper part indicates pooled amounts; lower part indicates amounts per LS vs. HS; Baseline: no punishment or norm information; Pun_NoInfo: punishment (5 ECU) without norm information; NoPun_NormInfo: no punishment with normative information; Pun_NormInfo: punishment (5 ECU) and normative information; NoPun_EmpInfo: no punishment with empirical information; Pun_EmpInfo: punishment (5 ECU) and empirical information. Whiskers represent 95% CIs.

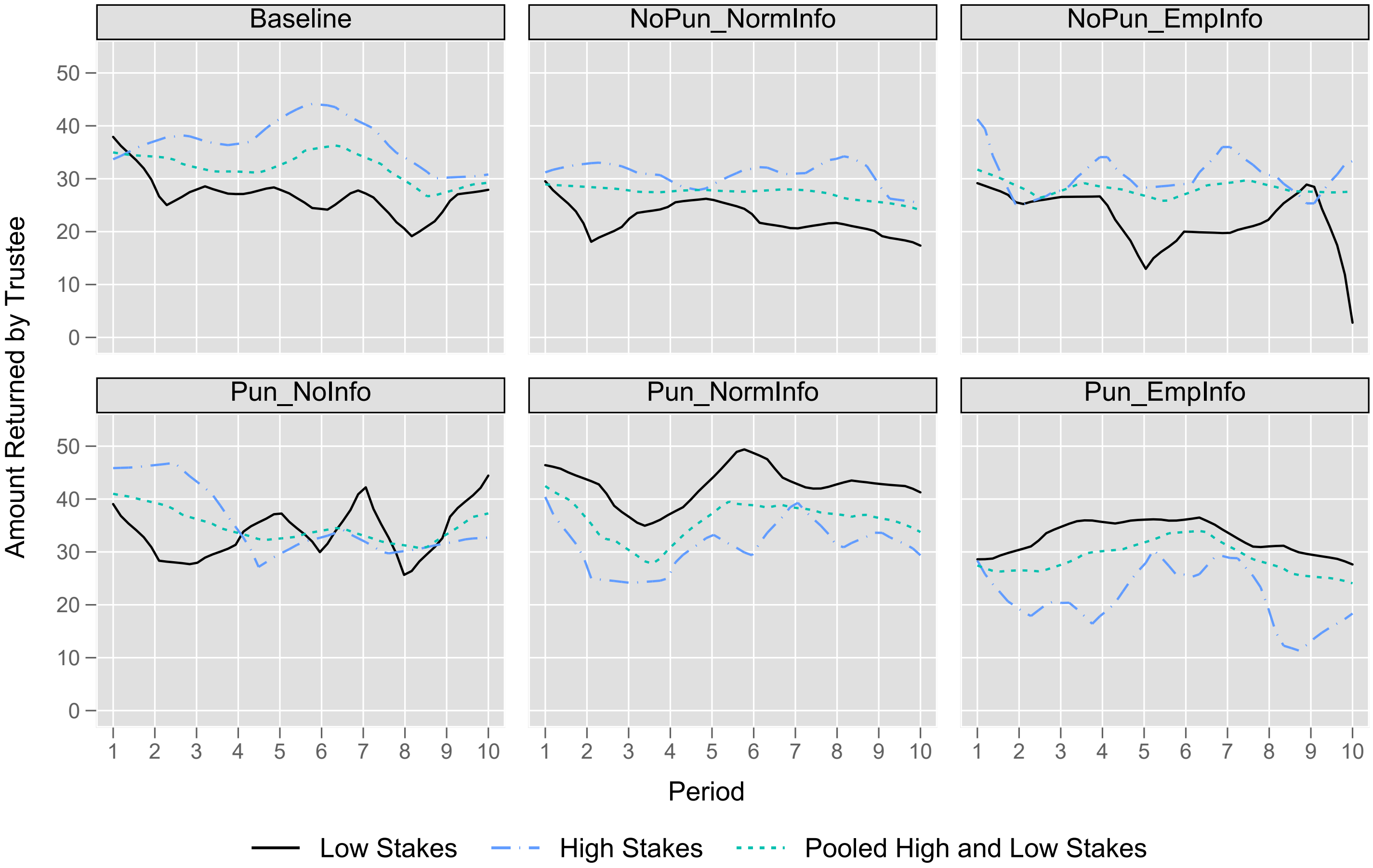


Figure A.2: Amounts returned by trustees as percentages of amounts received from investors over all periods.

DV: Amount Returned by Trustee (%)	**Low Stakes**		**High Stakes**	
	(1)	(2)	(3)	(4)
Treatment				
(Base Level: Baseline)				
Pun_NoInfo	6.036	5.711	-2.908	-3.931
	(5.363)	(5.767)	(5.664)	(5.898)
NoPun_NormInfo	-8.678	-8.696	-8.511	-8.579
	(5.748)	(6.110)	(5.846)	(5.843)
Pun_NormInfo	12.760**	13.841**	0.332	0.108
	(5.688)	(6.048)	(6.397)	(6.561)
NoPun_EmpInfo	-6.643	-7.771	-3.824	-4.477
	(5.187)	(5.492)	(5.363)	(5.445)
Pun_ EmpInfo	1.784	3.231	-10.145*	-12.066**
	(5.035)	(5.367)	(5.688)	(5.820)
Round	-0.597***	-0.382	-0.429**	-0.092
	(0.227)	(0.240)	(0.191)	(0.187)
Gender	-0.631	-0.131	3.324	3.410
	(3.286)	(3.405)	(3.650)	(3.728)
Self-Control	3.942**	4.261**	4.067**	4.015**
	(1.620)	(1.674)	(1.818)	(1.847)
Risk	0.338	0.257	0.102	0.167
	(0.699)	(0.733)	(0.809)	(0.831)
L1.Amount Received from Investor		0.049		0.027
		(0.072)		(0.039)
Constant	32.062***	29.648***	34.394***	31.889***
	(5.574)	(6.133)	(6.351)	(6.420)
Observations	711	599	844	763

Table A1: Random effects model with robust standard errors (in parentheses) clustered on the participant level. Estimations for all periods, including those in which no return request message was sent. Control variables include stakes (1 = high), Round (1-10), Gender (1 = male), Self-Control (higher number indicates more self-control, standardized measure), Risk (higher number indicates more risk-seeking, standardized measure). L1.Amount Received from Investor (% amount received from an investor in previous round, which indicates whether trustee faced a high- or low-stakes situation). Significance levels: *p<0.10, **p<0.05, ***p<0.01.

DV: Amount Returned by Trustees (%)	(1)	(2)
Punishment	1.672 (5.202)	1.389 (4.894)
Normative Information	-8.328 (5.667)	-8.889 (5.444)
Empirical Information	-5.457 (5.112)	-5.246 (5.027)
Stakes	1.681 (1.496)	2.000 (1.476)
Punishment × Normative Information	16.232** (7.740)	19.343*** (7.478)
Punishment × Empirical Information	3.467 (7.070)	3.654 (6.889)
Punishment × Normative Information × Stakes	-10.785*** (3.912)	-10.636*** (3.906)
Punishment × Empirical Information × Stakes	-8.940** (3.796)	-9.543** (3.837)
Period		-0.525*** (0.164)
Gender		2.651 (3.143)
Self-Control		4.043** (1.570)
Risk		0.153 (0.684)
Constant	32.643*** (3.888)	33.270*** (5.511)
Observations	1446	1446

Table A2: Random effects model with robust standard errors (in parentheses) clustered on the participant level. Punishment (1 = punishment implemented), Normative Information (1 = normative information implemented), Empirical Information (1 = empirical information implemented), stakes (1 = high), Remaining coding of control variables the same as in previous tables. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

C. Experimental Instructions

Subsequently, we present the instructions exemplary for Pun_EmpInfo (Punishment + Empirical Information). Differences with our other treatments are highlighted in the text. More specifically, the part highlighted red was presented only in this treatment and in NoPun_EmpInfo (No Punishment + Empirical Information) to the participants. In NoPun_NormInfo (No Punishment + Normative Information) and Pun_NormInfo (Punishment + Normative Information), the sentence was replaced with: "*In a previous survey, most participants said that Player 2 should return at least half of the tripled transfer amount.*" The part highlighted in green was only included in treatments that involved punishment.

Instructions

Thank you for coming! You have earned $10 for showing up on time. The following instructions explain how you can potentially earn more money by making a number of decisions. To maximize your chances to earn more money, please read these instructions carefully! If you have a question at any time, please raise your hand, and an experimenter will assist you.

For the purpose of the experiment, it is important that you do not talk or communicate in other ways with the other participants. Please turn off your cell phone and all other electronic devices. You are asked to abide by these rules. If you do not abide, we would have to exclude you from this and future experiments, and you will not receive any compensation for the experiment.

The experiment consists of **a total of 10 rounds**. At the end of the experiment, one round will be chosen at random, and you will be paid privately in cash based on your earnings from that round and your initial earnings for showing up on time. Your decisions remain anonymous to other participants throughout the experiment. No participant will know who has made what decisions. Please do not talk to each other during the experiment.

During the experiment, all amounts will be presented in ECU (Experimental Currency Unit). At the end of the experiment all the ECU you have earned will be converted to Dollars as follows:

2 ECU = 1 Dollar

General Procedure

- There are two types of Players: **Player 1** and **Player 2**.
- Player 1 acts first and Player 2 acts second.
- In each of the 10 rounds, a participant in the role of Player 1 will be **randomly** matched with one participant who is in the role of **Player 2** (and vice versa).
- No one will know the identity of his/her matched participant in any of the 10 rounds.

Endowment

- Each participant (both Player 1 and Player 2) receives an initial endowment of **8 ECU**.

Decisions of Player 1:

1. Transfer Decision

- **Player 1** will have the opportunity to send none, half or all of his/her initial endowment to **Player 2**. In this case, Player 1 can transfer **0 ECU**, **4 ECU**, or **8 ECU** to Player 2.
- Each ECU transferred will be **tripled**. For example, if **Player 1** decides to transfer **4 ECU**, **Player 2** will receive **12 ECU**. If **Player 1** decides to transfer **8 ECU**, **Player 2** will receive **24 ECU**.

2. Request decision

If Player 1 decides to transfer 4 ECU or 8 ECU to Player 2, **Player 2** will then decide how much to transfer back to Player 1 (further detail of Player 2's possible decisions are provided in the following section, 'Decision of Player 2'). ***In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transfer amount to Player 1.***

In addition, Player 1 is given the option to ask Player 2 to transfer back at least half of the tripled transfer amount. For example, if Player 1 transfers 4 ECU to Player 2 (so that Player 2 receives 12 ECU), Player 1 will decide whether to send Player 2 the return request message "I'd like you to transfer back to me at least half of the 12 ECU (i.e. at least 6 ECU)". Alternatively, if Player 1 transfers 8 ECU to Player 2 (so that Player 2 receives 24 ECU), Player 1 will decide whether to send Player 2 the return request message "I'd like you to transfer back to me at least half of the 24 ECU (i.e. at least 12 ECU)".

Decision of Player 2:

After Player 1 has made his/her decision(s), Player 2 will see Player 1's transfer decision. In the case that Player 1 transfers 4 ECU or 8 ECU, Player 2 will also see whether Player 1 asks him/her to transfer back at least half of the tripled amount. Player 2 will then decide how much (if anything) to transfer back to Player 1 as described below.

- If Player 1 transfers 0 ECU, Player 2 will have no decision to make. The final earnings of Player 2 and Player 1 will be their initial endowment of 8 ECU each.
- If Player 1 transfers 4 ECU or 8 ECU, Player 2 will decide how much money to transfer back to Player 1 and how much money to keep to himself/herself. This could be any amount between 0 and the tripled amount of what Player 1 has sent, regardless of whether Player 1 asks Player 2 to transfer back at least half of the tripled amount.
- In addition, conditional on Player 1's decision to ask Player 2 to transfer back at least half of the tripled amount, Player 2 will face a **Payoff-cut** if his/her back-transfer does not meet this request. In particular:

- If Player 1 decided to request Player 2 to transfer back **at least half** of the tripled transfer amount, Player 2's payoff will be reduced by **5 ECU** if his/her actual back-transfer is **less** than the requested amount. However, Player 2 will not face a Payoff-cut if his/her back-transfer amount satisfies the request.
- For example, suppose that Player 1 send 4 ECU (or 8 ECU) to Player 2, so that Player 2 receives 12 ECU (or 24 ECU), and suppose that Player 1 requests a back-transfer of at least half of the tripled amount, at least 6 ECU (or 12 ECU). In this case, if Player 2 decides to transfer some amount less than 6 ECU (or 12 ECU), his/her payoff will be reduced by 5 ECU.
- If Player 1 decides **not** to request that Player 2 transfer back at least half of the tripled transfer amount, then Player 2 will not receive any payoff cut irrespectively of the actual amount he/she sends back.

Payoffs:

Player 1

(**8 ECU**) – (**potential transfer to Player 2**) + (**potential back-transfer from Player 2**)

Player 2

(**8 ECU**) + (**3 x potential transfer from Player 1**) – (**back-transfer to Player 1**) – (**potential payoff cut**)

Final Remarks:

A new round starts after Player 1 and 2 has made his/her decision. In the beginning of each new round, Player 1 will be randomly matched with another Player 2. No one will know the identity of his/her matched participant. Each round will proceed in the same way.

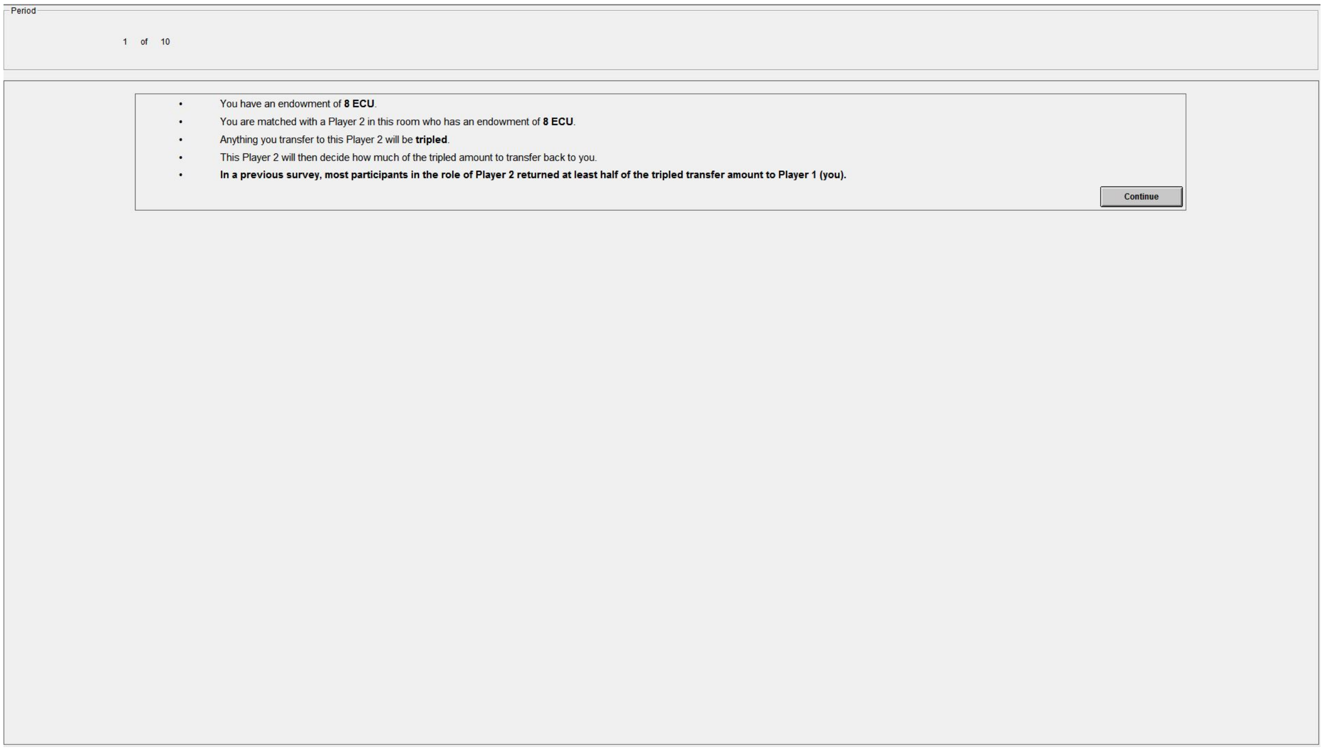
Player 1 will not know the result of each round (i.e. Player 1 will not know Player 2's decision in each round) until all the 10 rounds have finished. After all the 10 rounds have finished, each Player 1 will learn the matched Player 2's decision and the payoff outcomes in each round. Each Player 2 will also see a summary of the decision and payoff outcomes in each round.

One round will be chosen at random and Player 1 and 2 will be paid according to the outcome of that round.

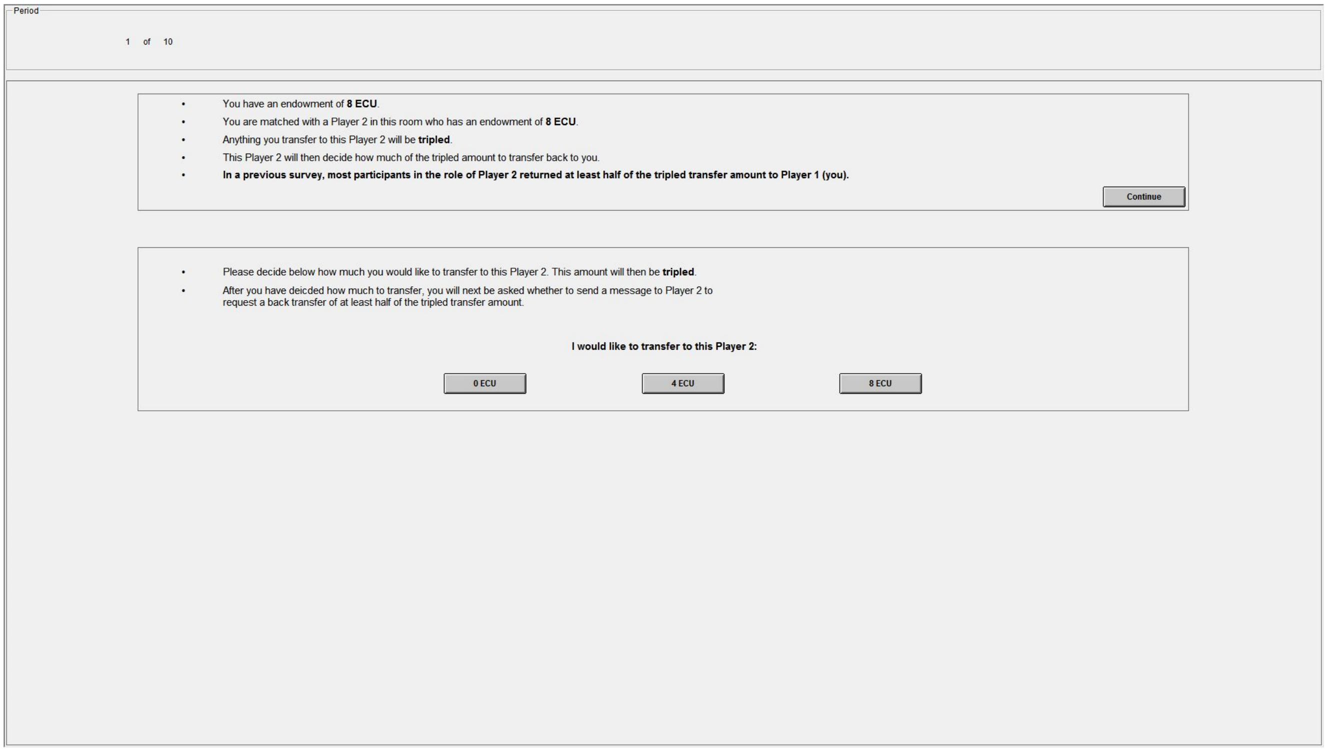
C. Screenshots of Experimental Procedure

Here, we present the screenshots for Treatment 5 (Punishment + Empirical Information). Differences from the other treatments are as previously explained in the experimental instructions. That is, indication of punishment and normative / empirical information was presented where the experimental design dictated. Screenshots are presented in the order in which the decisions occurred during one single round.

Investor



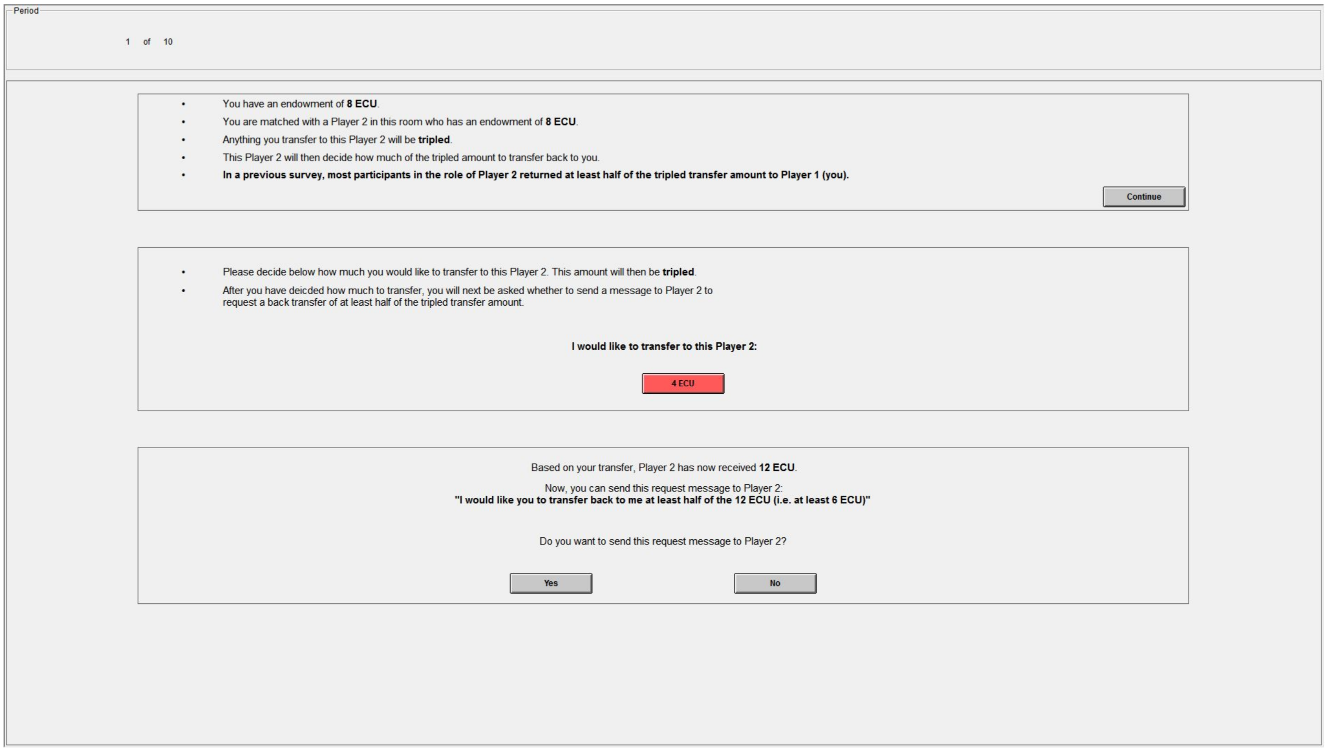
- You have an endowment of **8 ECU**.
- You are matched with a Player 2 in this room who has an endowment of **8 ECU**.
- Anything you transfer to this Player 2 will be **tripled**.
- This Player 2 will then decide how much of the tripled amount to transfer back to you.
- **In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transfer amount to Player 1 (you).**



- You have an endowment of **8 ECU**.
- You are matched with a Player 2 in this room who has an endowment of **8 ECU**.
- Anything you transfer to this Player 2 will be **tripled**.
- This Player 2 will then decide how much of the tripled amount to transfer back to you.
- **In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transfer amount to Player 1 (you).**

- Please decide below how much you would like to transfer to this Player 2. This amount will then be **tripled**.
- After you have deicded how much to transfer, you will next be asked whether to send a message to Player 2 to request a back transfer of at least half of the tripled transfer amount.

I would like to transfer to this Player 2:



- You have an endowment of **8 ECU**.
- You are matched with a Player 2 in this room who has an endowment of **8 ECU**.
- Anything you transfer to this Player 2 will be **tripled**.
- This Player 2 will then decide how much of the tripled amount to transfer back to you.
- **In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transfer amount to Player 1 (you).**

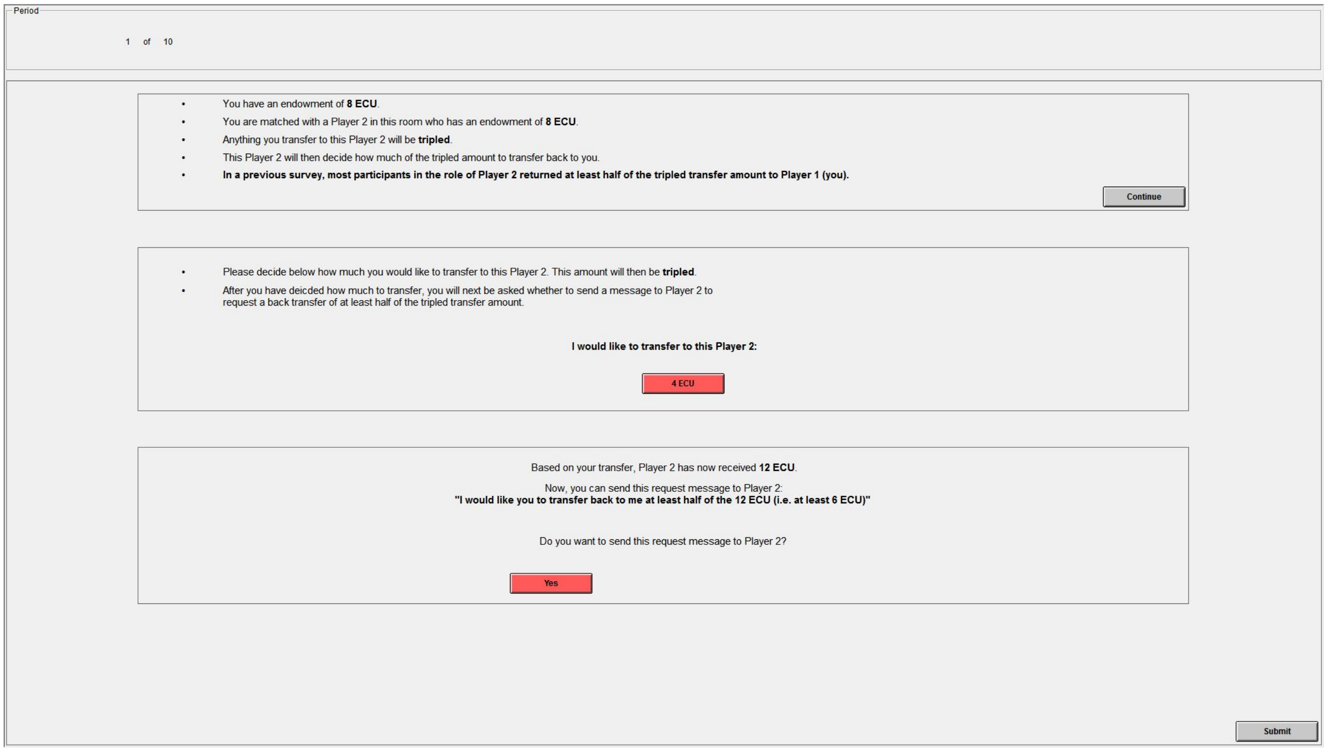
- Please decide below how much you would like to transfer to this Player 2. This amount will then be **tripled**.
- After you have deicded how much to transfer, you will next be asked whether to send a message to Player 2 to request a back transfer of at least half of the tripled transfer amount.

I would like to transfer to this Player 2:

Based on your transfer, Player 2 has now received **12 ECU**.

Now, you can send this request message to Player 2:
"I would like you to transfer back to me at least half of the 12 ECU (i.e. at least 6 ECU)"

Do you want to send this request message to Player 2?



Period

1 of 10

- You have an endowment of **8 ECU**.
- You are matched with a Player 2 in this room who has an endowment of **8 ECU**.
- Anything you transfer to this Player 2 will be **tripled**.
- This Player 2 will then decide how much of the tripled amount to transfer back to you.
- **In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transfer amount to Player 1 (you).**

- Please decide below how much you would like to transfer to this Player 2. This amount will then be **tripled**.
- After you have deicded how much to transfer, you will next be asked whether to send a message to Player 2 to request a back transfer of at least half of the tripled transfer amount.

I would like to transfer to this Player 2:

4 ECU

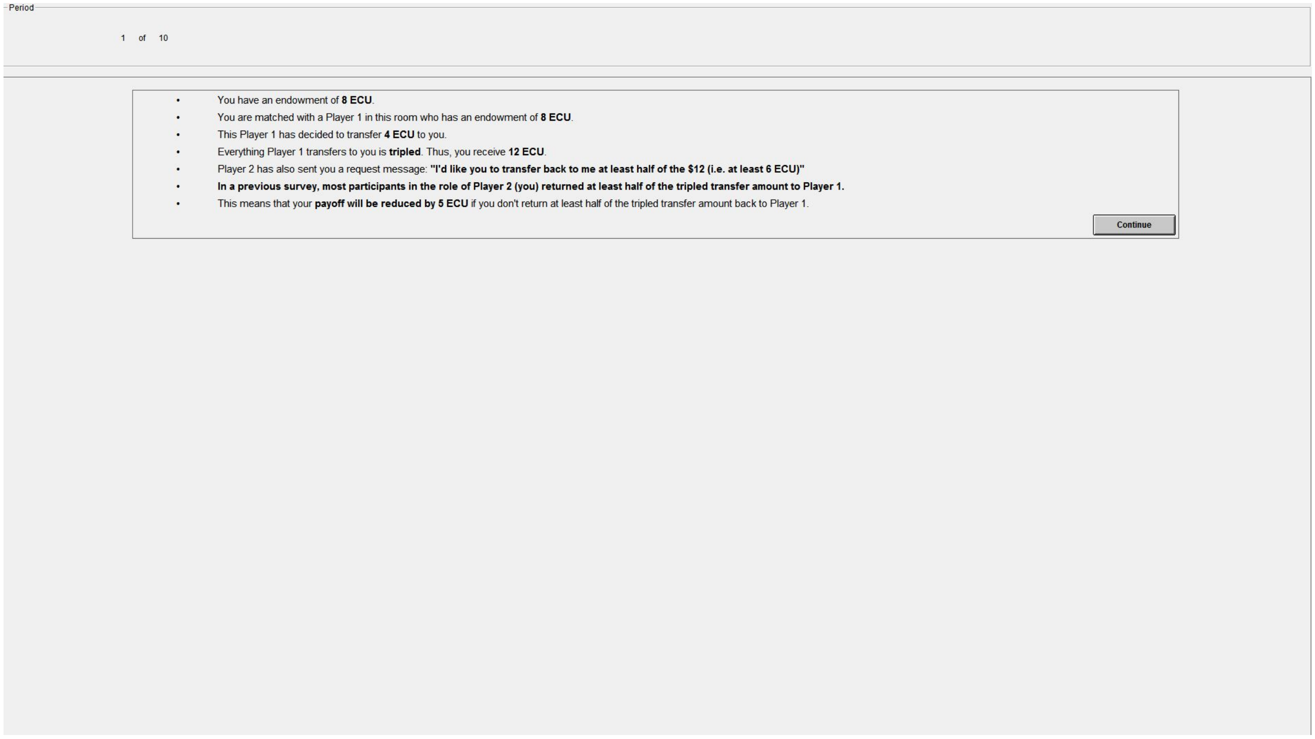
Based on your transfer, Player 2 has now received **12 ECU**.

Now, you can send this request message to Player 2:
"I would like you to transfer back to me at least half of the 12 ECU (i.e. at least 6 ECU)"

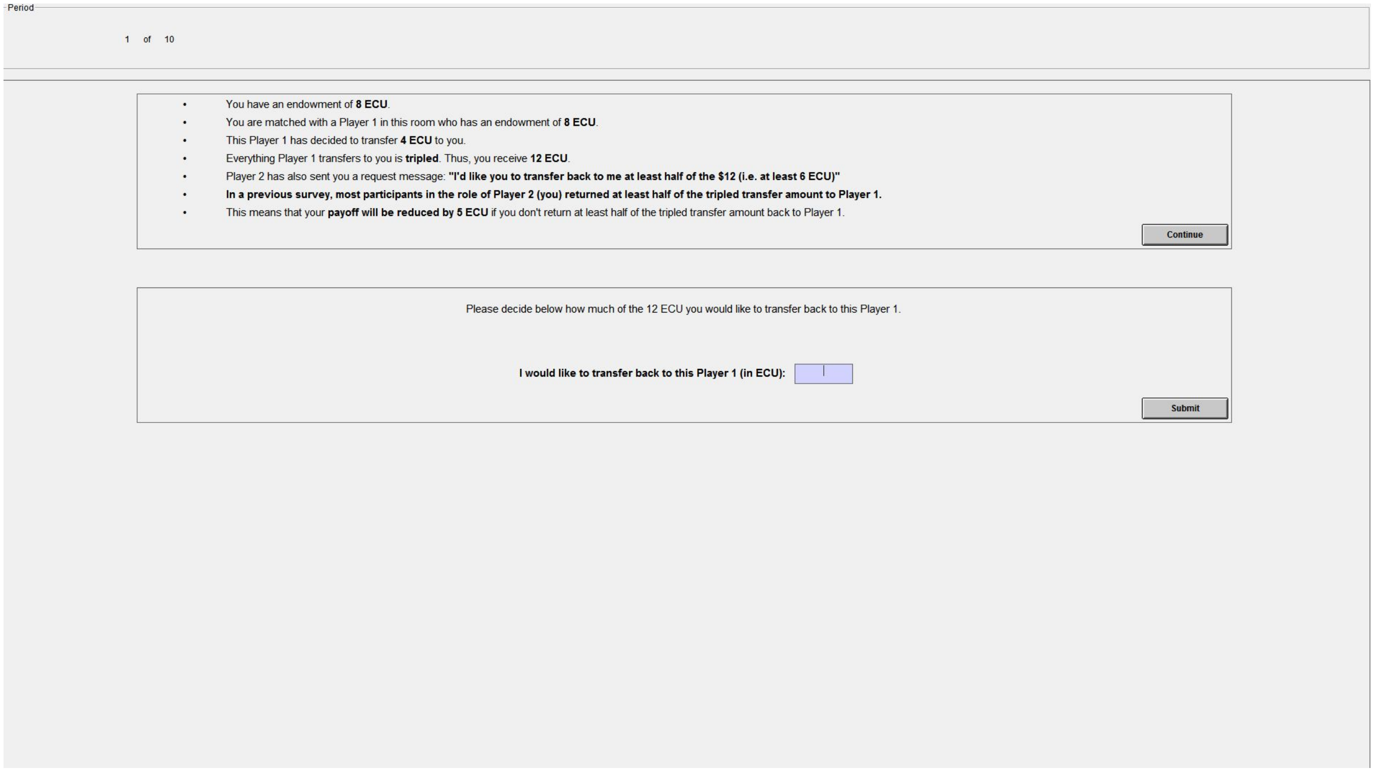
Do you want to send this request message to Player 2?

Yes

Trustee



- You have an endowment of **8 ECU**.
- You are matched with a Player 1 in this room who has an endowment of **8 ECU**.
- This Player 1 has decided to transfer **4 ECU** to you.
- Everything Player 1 transfers to you is **tripled**. Thus, you receive **12 ECU**.
- Player 2 has also sent you a request message: **"I'd like you to transfer back to me at least half of the $12 (i.e. at least 6 ECU)"**
- **In a previous survey, most participants in the role of Player 2 (you) returned at least half of the tripled transfer amount to Player 1.**
- This means that your **payoff will be reduced by 5 ECU** if you don't return at least half of the tripled transfer amount back to Player 1.

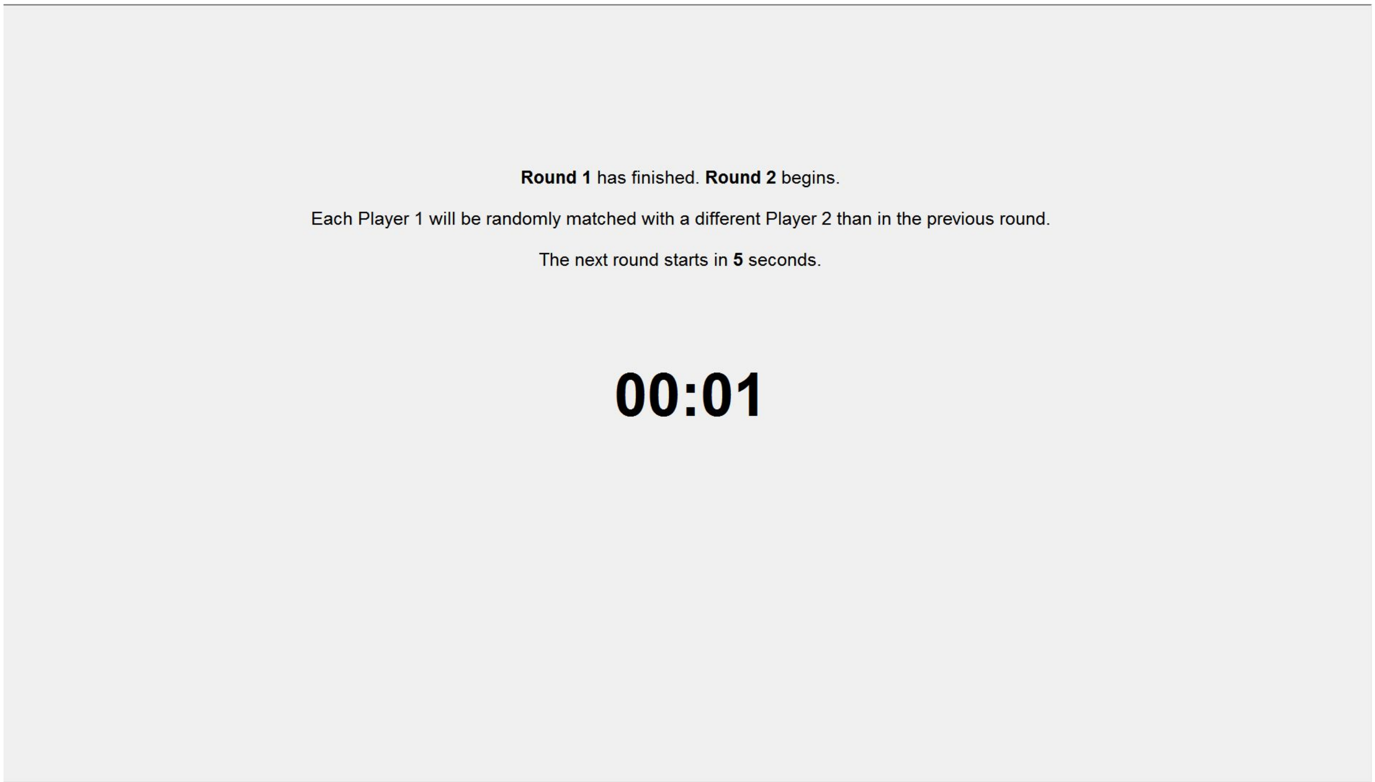


- You have an endowment of **8 ECU**.
- You are matched with a Player 1 in this room who has an endowment of **8 ECU**.
- This Player 1 has decided to transfer **4 ECU** to you.
- Everything Player 1 transfers to you is **tripled**. Thus, you receive **12 ECU**.
- Player 2 has also sent you a request message: **"I'd like you to transfer back to me at least half of the $12 (i.e. at least 6 ECU)"**
- **In a previous survey, most participants in the role of Player 2 (you) returned at least half of the tripled transfer amount to Player 1.**
- This means that your **payoff will be reduced by 5 ECU** if you don't return at least half of the tripled transfer amount back to Player 1.

Please decide below how much of the 12 ECU you would like to transfer back to this Player 1.

I would like to transfer back to this Player 1 (in ECU):

End of the round screenshot (Investor and Trustee)



Round 1 has finished. **Round 2** begins.

Each Player 1 will be randomly matched with a different Player 2 than in the previous round.

The next round starts in **5** seconds.

00:01