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

Cristina Bicchieri[a], Eugen Dimant[a], Erte Xiao[b]

[a] *University of Pennsylvania*
[b] *Monash University*

Abstract

A stream of research examining the effect of 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, enabling agents to justify selfish behavior. We address the question of punishment legitimacy by shedding light upon the importance of social norms and their interplay with punishment. Often people are presented with incomplete norm information: either about what most others do (empirical) or what most others deem appropriate (normative). We show that neither punishment nor empirical/normative information in isolation result in prosocial behavior. In turn, we find that prosociality is significantly increased when normative information and punishment are combined, but only when compliance is relatively cheap. When compliance is more expensive, we find that the combination of punishment and empirical information about others' conformity can have detrimental effects on prosocial behavior. We attribute this outcome to the differential ability to distort one's own beliefs about applicable norms. Our results have important implications for researchers and practitioners alike.

Keywords: Conformity, Experiments, Punishment, Social Norms, Trust Game
JEL: C91, D03, D73, H26

*We thank Ernst Fehr, Daniel Houser, Matthew Jackson, Rosemarie Nagel, and David Rand, as well as seminar participants at the Pompeu Fabra University and Yale University for helpful comments and suggestions. We also appreciate the feedback received at the Social Norms in Epistemic Communities conference at Monash University, the 2017's ANZWEE annual conference at University of Melbourne, the 7th Biennial Workshop on Social Dilemmas, and the 2017's North American Economic Science Association Meeting.

Email addresses: `cb36@sas.upenn.edu` (Cristina Bicchieri), `edimant@sas.upenn.edu` (Eugen Dimant), `erte.xiao@monash.edu` (Erte Xiao)

This version: November 1, 2018

1. Introduction

A large body of research has examined the effect of punishment on conformity. The standard economic theory of punishment focuses on how sanctions can change payoffs and thereby influence outcomes (Becker, 1968). It follows that when punishment is severe enough to overwhelm the expected benefit of defection, it can prevent opportunistic behavior. However, severe punishment usually requires costly monitoring and can have undesirable side effects.[1] As a result, punishment is often weak (Tyler, 2006; Balafoutas et al., 2016), meaning that the cost of punishment is lower than the cost of compliance. Yet weak punishment can backfire and lead to negative behavior (Gneezy and Rustichini, 2000; Fehr and Rockenbach, 2003; Villatoro et al., 2014; Calabuig et al., 2016). For example, Fehr and Rockenbach (2003) showed that in a trust game, trustees return less when the investor imposes a weak punishment to enforce the desired return amount. Additionally, Houser et al. (2008) showed that the same detrimental effect occurs even when weak punishment is imposed by nature, rather than directly by an individual (i.e., the investor).[2]

Our paper investigates whether punishment is more effective if it signals a norm, informing the trustee of a shared agreement and thus indicating that non-conformity is wrong and will be rightfully punished.[3] Indeed, in a naturally occurring environment, it is often made clear that an action will be punished because it violates a socially held standard rather than someone's (e.g., a punisher's) self-interested preference.

As legal scholars have long emphasized, punishment has an important norm-expressing function that is independent of its effect on material payoffs (Sunstein, 1996; Cooter, 1998; Kahan, 1998). For example, in a public goods game, Galbiati et al. (2008) show that punishment informs people about what they should do and creates an obligation to cooperate. Since norms are socially shared standards, one can expect that a punishment's norm-signaling function is enhanced when it has a social dimension.[4] For example, it has been shown that publicly implemented weak punishment promotes conformity more effectively than privately implemented punishment (Xiao and Houser, 2011). What is more, in a public goods game, cooperation significantly improves when punishment is imposed by

[1] Stigler (1970) argues that severe sanctions suffer from a lack of marginal deterrence for serious crimes.

[2] For a review, see Bowles and Polania-Reyes (2012).

[3] Fehr and Williams (2017) examine various peer-sanctioning mechanisms and find, among other things, that normative consensus is key in facilitating high cooperation rates. See also Abbink et al. (2017).

[4] For a recent review and discussion of literature, see Xiao et al. (2018).

group members rather than exogenously (Tyran and Feld, 2006).

Social norms have both an empirical and a normative component (Bicchieri, 2006).[5] They tell us what people usually do, as well as what people approve of. Empirical information alone may only indirectly suggest the underlying normative appropriateness of the behavior. Normative information instead provides a direct and explicit signal that an action is appropriate, even though it does not necessarily imply that most people behave accordingly. Yet studies show that, when only normative information is provided, such information has a stronger effect on behavior than just providing empirical information about what most people do (Cialdini et al., 1990; Bicchieri, 2006). As we often have access to only one type of norm information rather than both, it is important to investigate their potentially different effects on behavior, especially when accompanied by punishment. By adding normative information about an enforced behavior, it is made clear that a group of people, besides the punisher, view noncompliance as wrong. By signaling a norm violation, punishment increases the psychological cost of violation, thus enforcing compliance. However, when an enforced behavior is only supported by empirical information and noncompliance is perceived simply as a deviation from what others do, norms-signaling punishment may be weakened and even lead to transgression (see Schultz et al., 2007; Bicchieri and Dimant, 2018). Thus, we hypothesize that punishment is more effective when an enforced behavior is presented as the right course of action rather than what others do or would do. This hypothesis is also consistent with the observation that punishment in naturally occurring environments is usually associated with what is wrong and what should be done rather than what a majority does or would do (Bicchieri, 2006).

We conducted a controlled laboratory experiment to examine how providing the information that an enforced behavior is consistent with a shared norm can affect the outcome of punishment. An important methodological ingredient of our design is the focus on weak punishment such that the cost of punishment is not higher than the cost of compliance and monetary incentives are not the dominant driver of decisions. As a result, the introduced punishment is not equilibrium shifting, which sets our study apart from much of the existing research.

Our experiment consisted of six variations of a standard trust game, each with multiple

[5] 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.

rounds of play. We systematically introduced punishment, normative or empirical information, and the combinations thereof within the games. Subjects were assigned either to the role of investor or trustee for the full duration of their experimental session and restricted to only one treatment variation (between-subjects design). All treatments were variations of the Baseline condition, in which the investor decided whether to transfer any amount of her endowment to the trustee and whether to accompany this with a return-request message.[6] Any amount given by the investor was then tripled and transferred to the trustee, who then decided how much, if any, to return to the investor.

In the three treatments with punishment, the investor's request message was binding in that if the trustee returned less than 50%, she would receive an automatically implemented penalty of a fixed amount. Since we were only interested in weak punishment and not equilibrium-shifting, the penalty was designed to be always smaller than the 50% return, giving the trustee a monetary incentive to deviate. The three treatments with punishment varied on whether participants were informed that the request message was consistent with normative or empirical information.[7] In particular, the empirical information provided a statistic (based on truthfully collected information preceding the experimental session) that in a previous session most trustees returned at least 50%, while normative information stated that most participants in a previous session thought trustees should return at least 50%. To control for the effect of information alone, we also included two more treatments where punishment was absent and only empirical or normative information was provided.

By capitalizing on this design, we study the different effects of each type of information, especially when paired with punishment. We find that only the composite effect of normative information and punishment significantly increases conformity, while the separate enforcement mechanisms of punishment and normative information do not achieve this result by themselves. Our results help to better understand recent research yielding conflicting finding about separately manipulated normative or empirical information (Bicchieri, 2006; Goldstein et al., 2008; Ferraro et al., 2011; Keane and Nickerson, 2015;

[6] A fixed message that asked the trustee to return at least half of the most recently transaction. In conditions where punishment was not included, such a request was non-binding as the trustee could return any amount or none at all. Where punishment was included, ignoring such a request led to the automatic enforcement of punishment.

[7] To ensure the truthfulness of the information, the empirical/normative information presented borrowed from the behavior and beliefs of other participants from a previous study. Such methodology is commonly adopted in experimental research on social norms (Bicchieri and Xiao, 2009; Krupka and Weber, 2013).

Kraft-Todd et al., 2015; Bursztyn et al., 2015; Hallsworth et al., 2017; Muthukrishna et al., 2017). Interestingly, we also find that the combination of punishment and empirical information has detrimental effects on conformity levels.

Individuals may not respond to the threat of punishment if they view the required behavior as insufficiently justified. For punishment to be effective, it is important to pay attention to its norm expressing function, which establishes the legitimacy of the enforced behavior. Our findings suggest that supporting the required behavior with normative information can serve such a purpose. In contrast, empirical information may not help and may even be abused for self-serving purposes (Bicchieri and Dimant, 2018). In effect, our results raise concerns about the combination of punishment and empirical information, which is part of the recent wave of social norm nudging (i.e., Hallsworth et al., 2017). In our experiment, it is more acceptable to punish wrongness (e.g., disregarding normative information) than to punish deviation (e.g., disregarding empirical information). Another important consideration is that it is less challenging to infer what behavior is desirable from normative information than to do so from an empirical message (Bicchieri and Dimant, 2018). Many common behaviors are just customary and do not involve any injunction or obligation. In this sense, the interpretation of an empirical message leaves room for self-serving biases when deciding whether it points to a normative standard. Accordingly, in our experiment We focused on a careful dissemination of norm-laden messages and would suggest similar caution for future research.

Our results contribute to the understanding of how punishment impacts pro-social behavior. This is particularly important from a policy perspective, especially with regards to designing effective and sustainable behavioral interventions. Recent evidence suggests that the introduction of punishment alone is often less effective or even measurably destructive in changing negative behaviors. Examples include the elimination of female genital cutting in Africa, dueling in Europe, and foot binding in China, as well as the reduction of smoking, corruption and tax evasion in numerous other countries. We offer a novel explanation for the often negligible and sometimes detrimental effect of punishment — when enforced alone — observed in previous studies.

2. Experiment Design and Procedures

We recruited a total of 418 participants across six treatments at the University of Pennsylvania. Our experiment utilizes a variant of a trust game (Berg et al., 1995) as

introduced in recent related literature Fehr and Rockenbach (2003); Houser et al. (2008). Per experiment session, each participant was randomly assigned the role of investor or trustee and remained in that role throughout the experiment. Each participant played the game for 10 rounds. At the beginning of each round, each participant received an endowment of 8 Experiment 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 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 generated were based on a pilot trust game. On average, the majority of participants returned at least 50% of the tripled amount. When asked, the majority also indicated that Player 2 should indeed 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).

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 cost of conformity across all treatment specifications (explained in more detail below). It was disclosed 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

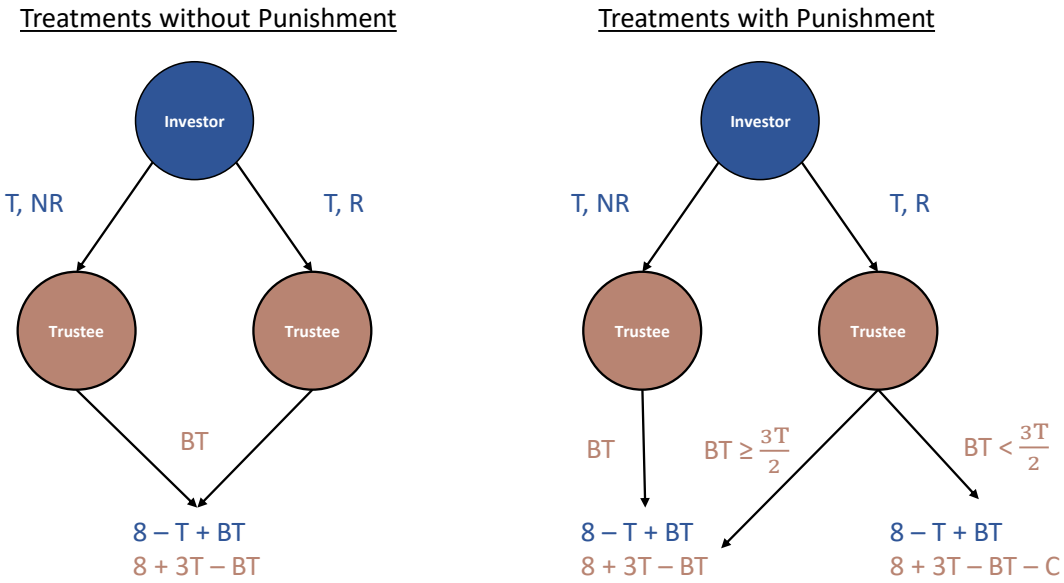


Figure 1: Sequence of actions and payoff structure in treatments with and without punishment. T: Investor's transfer to trustee. (N)R: Investor's decision to (not) send a return request message to the trustee. BT: Trustee's back-transfer to the Investor. C: Trustee's payoff cut (punishment)

transfer. The message was in a fixed form, with two quantitative components adjusted correspondingly 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.

Next, the trustee saw the transferred 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 participants were shown a summary of the decisions and outcomes of each round only at the end of the experiment. Thus, our design avoided 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 sent a return request message, the trustee would receive a payoff cut of 5 ECU if his/her back transfer amount were less than 50% of the tripled transfer amount. On the other hand, if the investor did not send the return request message, the trustee would not receive any payoff cut regardless of the amount of the back transfer.

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 the treatments with empirical information, the instructions read: *"In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transferred amount to Player 1."*

Since our main focus is to study the relation between norm information and punishment and to retain comparability within and across treatments, we used general empirical/normative messages throughout the experiment. That is, we did not specify whether the truthfully obtained message was the result of behavior/beliefs in a low conformity cost or a high conformity cost situation. 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 survey), the exact content of this information remained unspecified. We return to this point in our discussion section.

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 the Pun_NoInfo treatment, 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 statistics, 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 differed from the Pun_NormInfo and Pun_EmpInfo treatments only 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 experiment 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. The average duration of a session, which included the game and a post-experiment questionnaire, was 45 minutes, and 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, participants were 22.2 years old on average and 62.7% were female.

3. Predictions

Our main question is whether punishment is more effective when it is combined with normative or empirical information. When empirical or normative information is presented, it 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.

We argue that punishment accompanied by norm information makes salient that the punished behavior violates the norm. And, depending upon one's sensitivity to the specific norm, this salience increases the psychological cost of violation. As a result, punishment can effectively change behavior even when its 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 $\pi(r)$ denote the trustee's (i.e., Player 2's) payoff when returning r and $\pi(r^0)$ the Player 2's payoff when returning r^0, the amount that he/she thinks is approved by others. Let $k \geq 0$ be Player 2's sensitivity toward the norm then Player 2's disutility of deviating can be defined as

$$k|\pi(r) - \pi(r^0)|.$$

Player 2 decides how much to return (r) in order to:

$$\max_r U = \pi(r) - k|\pi(r) - \pi(r^0)| \tag{1}$$

In our experiment, punishment is always weak in that the required minimum return amount, either 6 ECU or 12 ECU, is always higher than the fixed penalty of 5 ECU. Previous studies have shown that weak punishment alone does not increase returns (Fehr and Rockenbach, 2003; Houser et al., 2008). For simplicity, we assume that in both the Baseline and the Pun_NoInfo treatments, players can easily hold the belief that the goal is to maximize profit, because there is no explicitly stated norm regarding what is the right amount to return. Thus, the utility maximizing decision is to return zero in these two treatments. When punishment is combined with empirical or normative information about returning 50%, returning zero is viewed as violating the norm and thus introduces an additional psychological cost due to the disutility of norm violation. In the experiment, there are two cost conditions and two types of information, normative or empirical, either alone or combined with punishment. We discuss each conformity cost case separately. In each case, we assume that Player 2 will not return more than the minimum amount specified in the normative/empirical information.

Case 1 (low conformity cost): Player 1 sends 4 ECU and requests 6 ECU to be returned.

$$\max_r U = \pi(r) - k|\pi(r) - \pi(r^0)| \tag{2}$$

In the Pun_NormInfo and Pun_EmpInfo treatments:

As the norm information points to a return approved by other players, $r^0 = 6$. With a dictated punishment of 5 ECU, we have

$$U = \begin{cases} (12 - r) - k((12 - r) - 6), & \text{if } r = 6 \\ (12 - r - 5) - k((12 - r) - 6), & \text{if } r < 6 \end{cases}$$

It follows that $r^* = 6$ if $k > 1/6$ and $r^* = 0$ if $k < 1/6$.
Given a significant number of Player 2's with $k > 1/6$, participants in the two punishment-plus-information treatments were expected to achieve a higher rate of compliance than those in the Baseline and Pun_NoInfo treatments where $r^* = 0$.

In the NoPun_NormInfo and NoPun_EmpInfo treatments:

Similarly, $r^0 = 6$ as a result of present norm information. However, there was no punishment for returning less than 6 ECU. Thus,

$$U = (12 - r) - k((12 - r) - 6)$$

It follows that $r^* = 6$ if $k > 1$ and $r^* = 0$ if $k < 1$.

Given a significant number of Player 2's with $k > 1$, participants in the two information-only treatments were expected to achieve a higher rate of compliance than those in the Baseline and Pun_NoInfo treatments.

Case 2 (high conformity cost): Player 1 sends 8 ECU and requests 12 ECU to be returned.

$$\max_r U = \pi(r) - k|\pi(r) - \pi(r^0)| \tag{3}$$

In the Pun_NormInfo and Pun_EmpInfo treatments:

With $r^0 = 12$ and a punishment of 5 ECU for returns less than 12 ECU, we have,

$$U = \begin{cases} (24 - r) - k((24 - r) - 12), & \text{if } r = 12 \\ (24 - r - 5) - k((24 - r) - 12), & \text{if } r < 12 \end{cases}$$

It follows that $r^* = 12$ if $k > 7/12$ and $r^* = 0$ if $k < 7/12$.
Given a significant number of Player 2's with $k > 7/12$, participants in the two punishment-plus-information treatments were expected to achieve a higher rate of compliance than those in the Baseline and Pun_NoInfo treatments.

In the NoPun_NormInfo and NoPun_EmpInfo treatments:

Similarly, $r^0 = 12$ as a result of present norm information. However, there was no punishment for returning less than 6 ECU. Thus,

$$U = (24 - r) - k((24 - r) - 12)$$

It follows that $r^* = 12$ if $k > 1$ and $r^* = 0$ if $k < 1$.

Given a significant number of Player 2's with $k > 1$, participants in the two information-only treatments were expected to achieve a higher rate of compliance than those in the

Baseline and Pun_NoInfo treatments.

The above analyses suggest that, in each case, it is more likely to observe a higher compliance rate in the two punishment-plus-information treatments than in the two information-only treatments—as the latter would require a higher k. For example, if Player 1 sends 4 ECU and requests 12 ECU, a Player 2 with sensitivity k in range $(1/6, 1)$ would be more likely to comply in the Pun_NormInfo and Pun_EmpInfo treatments than in the other four treatments. Likewise, if Player 1 sends 8 ECU and requests 12 ECU, a Player 2 with sensitivity k in range $(7/12, 1)$ would be more likely to comply in the Pun_NormInfo and Pun_EmpInfo treatments than in the other four treatments

So far, we do not differentiate normative and empirical information. As we discussed earlier, while normative information directly points out that a return of 50% is the right thing to do, empirical information only indirectly suggests that. Whether Player 2 complies following the exhibited empirical information depends on how she interprets the information. Previous research has suggested that it is easier for individuals to distort their beliefs about the appropriateness of a norm in the presence of empirical rather than normative information (Bicchieri and Dimant, 2018). Thus, it is possible that Player 2 is less likely to form the belief that $r^0 = 6$ or $r^0 = 12$ when given the empirical rather than the normative information. As a result, the compliance rate would be lower in the empirical than in the normative information treatments. In the extreme case where Player 2 rejects the empirical information, profit maximizing behavior would erase any impact of empirical information on punishment. Together, our predictions are:

Prediction 1:

$$\text{Return}^{\text{Pun\_NormInfo}} > \text{Return}^{\text{Pun\_EmpInfo}} \geq \text{Return}^{\text{Baseline}} \approx \text{Return}^{\text{Pun\_NoInfo}}$$

Prediction 2:

$$\begin{aligned} \text{Return}^{\text{Pun\_NormInfo}} &> \text{Return}^{\text{NoPun\_NormInfo}} \\ \text{Return}^{\text{Pun\_EmpInfo}} &\geq \text{Return}^{\text{NoPun\_EmpInfo}} \end{aligned}$$

Furthermore, the differences in the cutoffs of k in Case 1 and Case 2 make it more likely to observe the positive effect of punishment combined with norm information in predictions 1 and 2 when the conformity cost is low (i.e. Case 1) compared to high (i.e. Case 2).

4. Results

We investigate the return behavior of trustees in different treatments varying punishment, norms, and combinations thereof.[8] In the subsequent sections, we focus on the trustees' average return behavior.[9] We reference the case of 8-ECU transfer as High Conformity Cost (HCC) condition, which requires trustees to return 12 ECU, while the case of 4-ECU transfer as Low Conformity Cost (LCC) 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 conformity cost (HCC and LCC).[10]

We find that punishment alone is not successful in improving return rates, especially in HCC, supporting Prediction 1. Neither empirical nor normative information alone induces a return rate higher than that of the Baseline treatment. The combination of punishment and normative information produces substantial positive behavioral change but only in LCC, which provides evidence for Prediction 2 and highlights the limits of normative information. Interestingly, the combination of punishment and empirical information is not only ineffective when the compliance cost is low, but is in fact 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 and Dimant, 2018; Bolton et al., 2018).

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, with or without an examination along HCC versus LCC. For the pooled results, introduction

[8] 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.

[9] All investors sent a return request message at least once (overall, in 93% of the time), with no significant differences across treatments. To allow for comparability across treatments, our analysis includes only the cases where a return request message was sent. An examination of open responses given by the trustees in the post-experiment survey reveals that most found the investors' requests appropriate. Our regression analyses as presented in Section 4.4 are robust to the inclusion of the absent-request cases.

[10] The bootstrap two-sample t-test method (BSM; see Moffatt 2015) with 9999 replications was chosen for our random rematch protocol and investigation of mean differences of average return behavior. BSM (significant at $p < 0.05$) retains the rich cardinal information in the data without making any assumptions about the distribution. Unless noted otherwise, non-parametric Mann-Whitney-U (MWU) tests support these findings. Our regression analyses—that controlled for covariates, periodical trends, and clustering of standard errors—yield results that are coherent with our econometric approach here (see Section 4.4).

of punishment yields a non-significant increase from 32.4% to 35.6% (BSM, p=0.19). The same is both for LCC and HCC separately (p=0.10 and p=0.91, respectively).

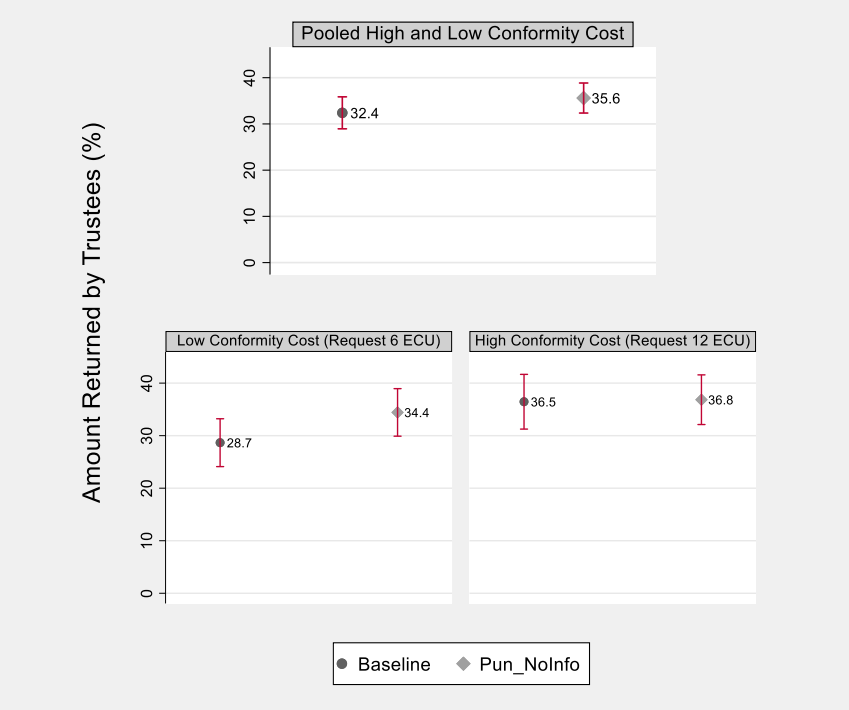


Figure 2: Amounts returned by trustees as percentage of amounts received from investors; upper part indicates pooled amounts; lower part indicates amounts per LCC vs. HCC; Baseline: no punishment or norm information; Pun_NoInfo: punishment (5 ECU) without norm information. None of the comparisons are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

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 $r \geq 50\%$.[11]

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 Cost 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 under the

[11]For an analysis of types, we calculate three ratios for high and low conformity costs per participant, each of 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 conformity costs, decisions could be impacted by the transferred amount.

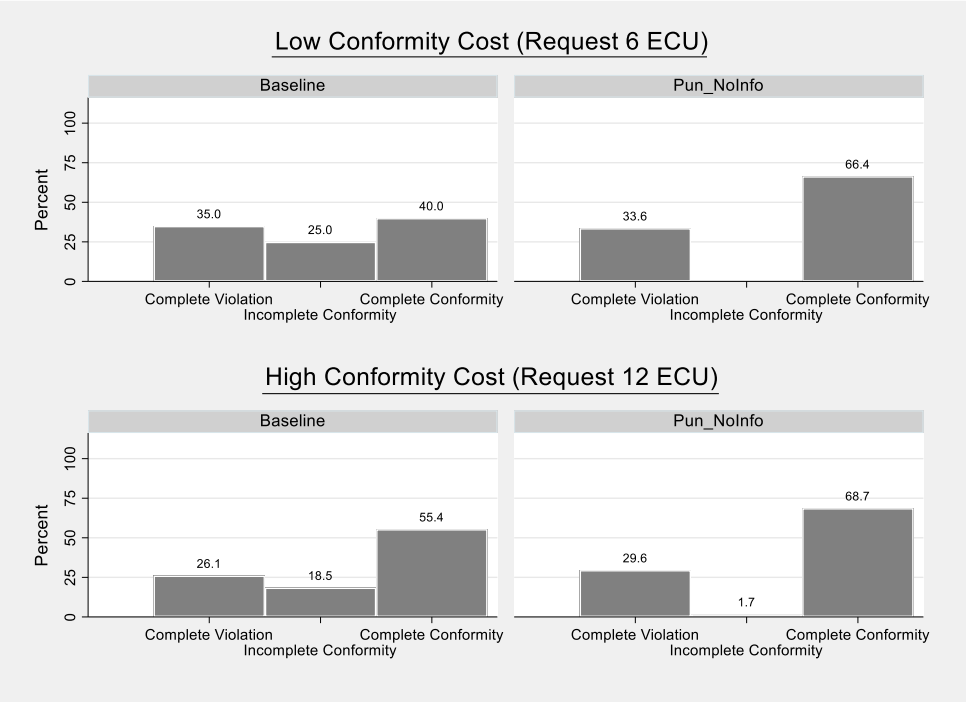


Figure 3: Distribution of return types in Baseline (NoPun_NoInfo) and Pun_NoInfo conditions.

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 changes non-signficantly (33.6% vs. 35.0, BSM, p=0.94), punishment significantly increases the proportion of Complete 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 were right below the 50% cut-off.

In contrast, in HCC the difference between the punishment condition and the Baseline is relatively small and non-significant (K-S, p=0.33). While we observe significantly less Incomplete Conformity types in the punishment 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, as in Houser et al. (2008), we observe that in the presence of punishment investors achieve either 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 HCC. In addition to individual differences, the reason for this result may be that Houser et al. (2008) allowed for requests of returns much higher than 50%, which in turn lead to lower levels of compliance with the request.

4.2. Effect of Norm Information Alone

For pooled data, we find empirical/normative information alone does not have a significant impact on the return rates either (Figure 4). For LCC and HCC, the differences in average return between the Baseline and both NoPun_NormInfo and NoPun_EmpInfo are not statistically significant (LCC: 28.7% vs. 23.7%, BSM, p=0.17; 28.7% vs. 23.3%, BSM, p=0.11; HCC: 36.5% vs. 30.5%, BSM, p=0.11; 36.5% vs. 30.9%, BSM, p=0.11).

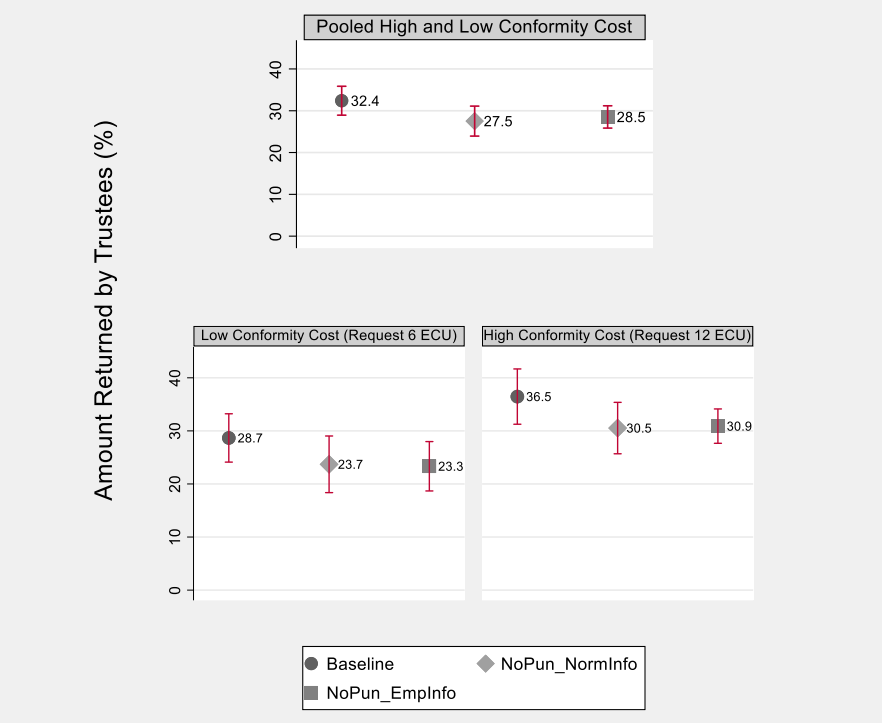


Figure 4: Amounts returned by trustees as percentages of amount received from investors; upper part indicates pooled amounts; lower part indicates amounts per LCC vs. HCC; Baseline: no punishment or norm information; NoPun_NormInfo: no punishment, with normative information. NoPun_EmpInfo: no punishment,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

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

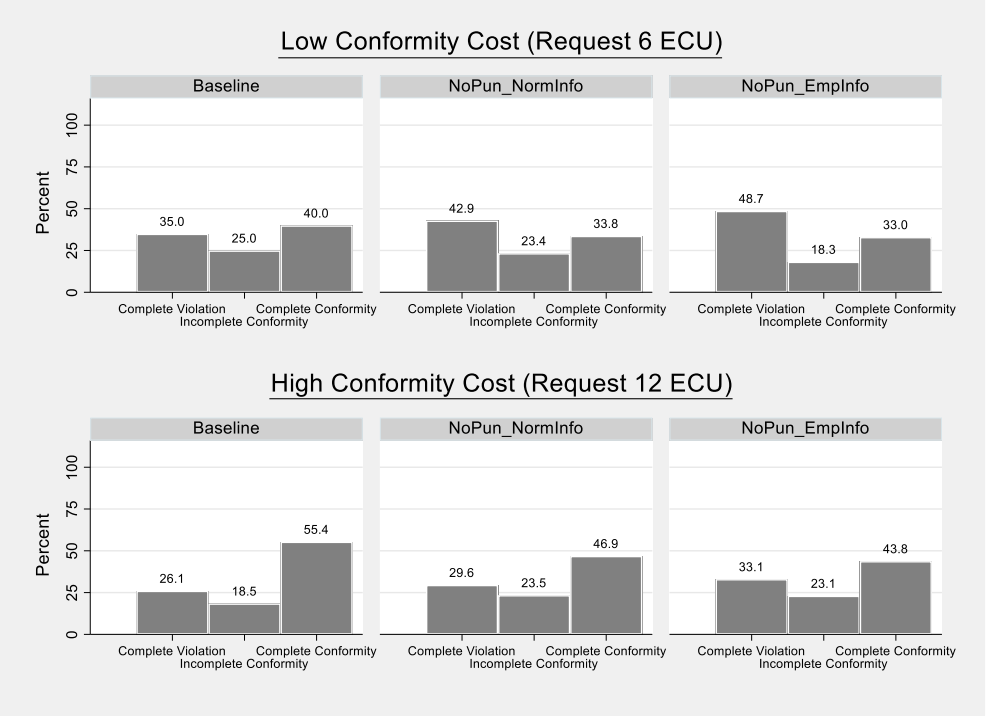


Figure 5: Distribution of return types in the Baseline (NoPun_NoInfo), NoPun_NormInfo, and NoPun_EmpInfo treatments.

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 cost conditions, we observe a significant decrease in the trustees' return in the Pun_EmpInfo condition as 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 LCC over the Baseline (42.7% vs. 28.7%, BSM,

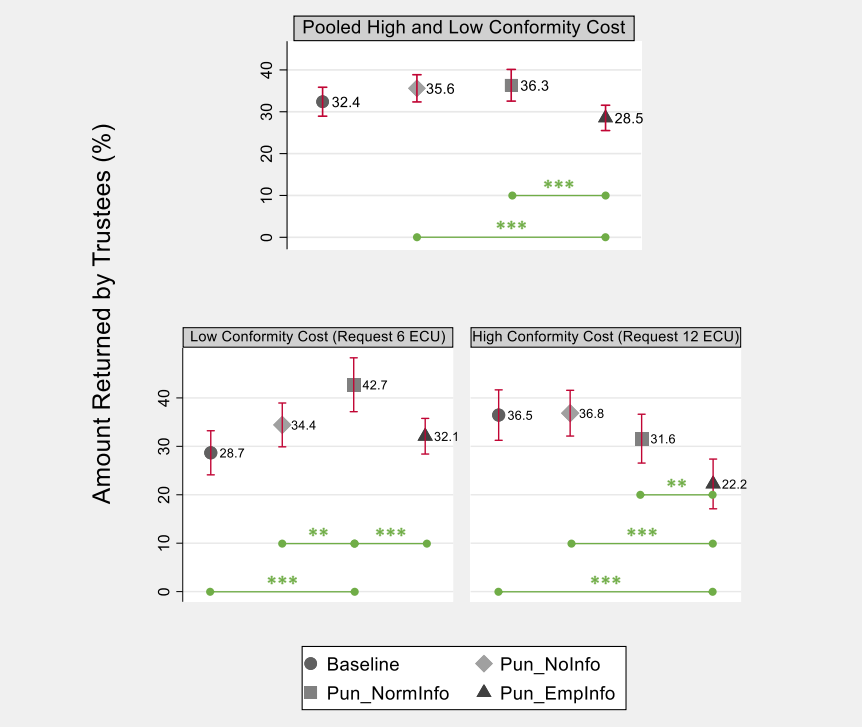


Figure 6: Amounts returned by trustees as percentages of amount received from investors; upper part indicates pooled amounts; lower part indicates amounts per LCC vs. HCC; Baseline: no punishment or norm information; Pun_NoInfo: punishment (5 ECU) without norm information; Pun_NormInfo: punishment (5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

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 in the Pun_NormInfo treatment (32.1% vs. 42.7%, BSM, p<0.01).[12] These results support Prediction 1 that punishment is more effective when it is combined with normative information (about a socially disapproved behavior) than enforced alone or with empirical

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

information (about a behavior deviating from the majority).

Of particular interest, normative information plays only a negligible role in HCC: 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%, BSM, p=0.18; 31.6% vs. 36.8%, BSM, p=0.18). Moreover, adding empirical information statistically 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).

For brevity, the full comparisons of the average return in all the six treatments are illustrated in Figure A.1 in the Appendix. Here we highlight the results related to Prediction 2 that suggest the observed effects of punishment combined with normative or empirical information are not due to the normative or empirical information alone, but to their combination with punishment. Compared to Pun_NormInfo, NoPun_NormInfo commands lower conformity rates when pooled across conformity costs (27.5% vs. 36.3%, BSM, p<0.01). Consistent with the discussion of Prediction 2, the difference is mainly driven by the LCC condition (23.7% vs. 42.7%, BSM, p<0.01). The difference in the HCC condition is not significant. (30.5% vs. 31.6%, BSM, p=0.77). These results suggest that the significant effect of Pun_NormInfo on return cannot be attributed to the normative information alone. When comparing the Pun_EmpInfo with the NoPun_EmpInfo treatments, we observe a significant increase in conformity for the former when LLC is in effect (23.3% vs. 32.1%, BSM, p<0.01). As we reported above, the positive effect of Pun_EmpInfo can be mostly attributed to punishment alone. On the other hand, empirical information with punishment backfires in HCC — 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 are pooled (28.5% vs. 28.5%, BSM, p=0.95)

These results suggest that the cost of conformity and the kind of norm information (empirical or normative) influence the benefit of combining punishment with a norm. Consistent with our hypothesis, normative information is helpful and its supplemental effect is subject to the cost of conformity. When the cost is high, neither normative or empirical information help to improve the efficacy of punishment. A surprising finding, however, is that 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 LCC condition reveal significant dissimilarities between the Baseline and the Pun_NormInfo and Pun_EmpInfo treatments (K-S, p<0.01), the latter of which

uncover 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 those in 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.

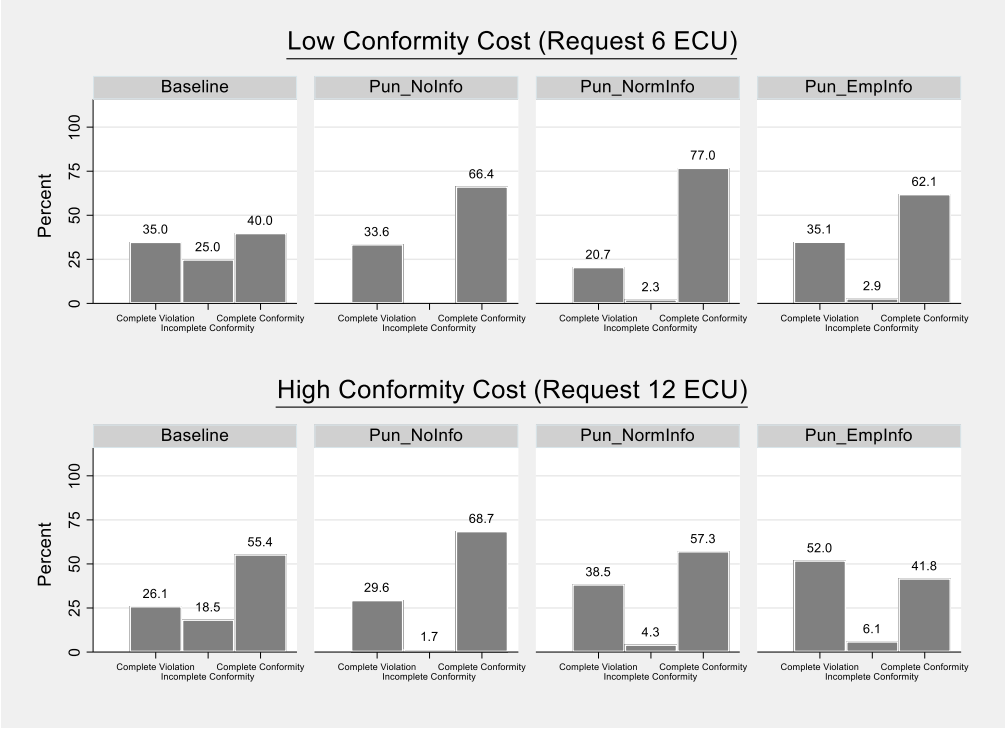


Figure 7: Distribution of return types in Baseline (NoPun_NoInfo), Pun_NoInfo, Pun_NormInfo, and Pun_EmpInfo treatments.

Continuing with the analysis of LCC, while Pun_EmpInfo offers significant increases in Complete Conformity (62.1% vs. 40.0%, BSM, p<0.01) over the Baseline treatment, such development 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 via punishment, which is corroborated by the substantially smaller number 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 treatments (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).

To reiterate, when the cost of conformity is low the return patterns across treatments are consistent with Prediction 1. Punishment can more effectively promote reciprocity by making salient the fact that returning less than the requested amount is socially disapproved of. The interaction of information and punishment is particularly effective when the former is normative. We will return to this finding in the discussion section. As seen from the average return data, when the cost of conformity is high, the benefit of both types of information is much less evident and empirical information is even detrimental. This implies that prediction 1 only holds for LCC, but not HCC.

Figure 7 further reveals that the detrimental effect observed in the Pun_EmpInfo treatment in the HCC 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 treatment.

These results suggest that the detrimental effect in HCC of the Pun_EmpInfo condition is mainly due to adding the empirical information to punishment rather than the punishment itself. Note that, in HCC, punishment alone hardly affects conformity, whereas adding norm information decreases conformity and significantly so in the empirical information case. Since compliance is relatively more costly in HCC than in LCC, compliance creates a tension between selfish behavior and obeying the rule. To solve the tension one may use some wiggle room, for example, forming a self-serving belief in the empirical information case ("only individuals in the low conformity cost condition followed the rule"). When conformity is cheap (LCC) we do not see this effect. Existing experimental evidence indicates that empirical information, but not normative information, gives rise to (self-serving) belief distortion to justify non-compliance (e.g., Bicchieri and Dimant 2018; also

Discussion in the current paper).

4.4. Regression Results

We analyze our data through different variants of multivariate regressions that examine the robustness of our results. In all cases, we employ random effects panel regressions with standard errors clustered at the participant level. As Table 2 indicates, the examination of average return behavior across treatments yields three main results that mirror exactly our results from the previous sections, showing that the findings are robust to the inclusion of various controls.[13] The results are as follows:

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

Result 2: The combination of punishment and normative information is successful at increasing 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 specification.

The coefficients from our controls suggest that return behavior declines over time and that participants with higher self-control (adapted from Tangney et al. 2018) accumulate higher return rates. We find no significant gender heterogeneity. The non-significant coefficient for previous round's investor behavior indicates that the possibility of learning is minuscule at best, which supports our methodological choice of random partner-rematch across rounds. In conclusion, our regression results showcase the robustness of our mean analyses.

[13]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 **??** in the Appendix). We provide a more detailed analysis of the drivers of trustee behavior across treatments in Table **??** in the Appendix.

DV: Amount Returned by Trustee	**Low Conformity Cost**		**High Conformity Cost**	
	(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 Conformity Cost (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 (ECU amount received from an investor in previous round, which indicates whether trustee faced a high or low conformity cost condition). Significance levels: $p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

5. Discussion and Conclusion

There is mounting interest in applying social norm methods to enhance nudge interventions (OECD, 2015; Miller and Prentice, 2016; **?**). Our findings suggest that norm-based interventions can lead to significant improvements but also backfire, even if the norm is embodied in a cooperative context and clearly stated (as opposed to left uncertain, as is the case in Bicchieri and Dimant 2018). We find two main effects of combining different types of norm information with punishment depending on the cost of conformity. With a low cost of conformity, we find that the combination of normative information and punishment

significantly raises the rate of return compared to baseline (in which both punishment and norm information is absent), punishment alone, and normative information alone. Behavior when the empirical information is combined with punishment does not display any significant differences relative to the other conditions. With a high cost of conformity, however, we find no significant effect of the combination of normative information and punishment.

Interestingly, we find that, when the conformity cost is high, the combination of empirical information and punishment can produce a detrimental effect by significantly decreasing the rate of return compared to baseline and just punishment. One possible explanation for this result that is in line with recent experimental literature is that individuals attempt to exploit the wiggle-room of norm-based interventions to avoid compliance (Konow, 2000; Dana et al., 2007; Spiekermann and Weiss, 2016; Bicchieri and Chavez, 2013; Bicchieri and Dimant, 2018). These experimental findings indicate that individuals tend to choose self-serving beliefs (and behavior) more often with respect to the empirical information about a norm due to a more flexible interpretation of what other people do, whereas individuals have a harder time distorting their own understanding of what behavior is normatively appropriate. In our experiment,the negative effect of empirical information with punishment in the high conformity cost condition may be due to individuals interpreting the empirical information as referring the low cost condition only, since it is cheaper to comply in that condition. In other words, participants who do not wish to conform in the high cost condition are uncertain about the reference group: the empirical information may refer to the low cost group, high cost group, or both. They use this uncertainty to form the belief that the empirical information refers to the low cost group only, since it's cheaper to comply in it. In the low cost condition, this rationalization does not work, since if participants complied in the high cost condition then they surely complied in the low cost condition as well.

The previously discussed literature suggests that it is harder to distort normative beliefs than empirical ones (as found in Bicchieri and Dimant, 2018) and thus we do not see a significant reduction in the return rate in the normative information case. Indeed, it would be hard to assume that reciprocity is only appropriate in the low cost condition. In line with this reasoning, there is no distortion in the low conformity cost condition, yielding a significant upward reaction of conformity.[14]

[14] Another possible explanation is that people view punishment as illegitimate when the reason for punishment is perceived simply as a divergence of their behavior from others'. Empirical information might

In sum, our experiment shows that providing normative information about socially disapproved behavior enhances the efficacy of punishment as long as compliance is not too costly. An important insight for policy-makers is that weak punishment itself may not be sufficient to enforce positive behavior: it may also be critical to highlight the social desirability of the enforced behavior. On the other hand, the detrimental effect of empirical information and punishment is, to the best of our knowledge, a novel finding and illustrates a potential pitfall of common norm-based interventions. We suggest closely examining normative and empirical information beyond the simple "majority rule" implemented in typical social norms interventions in order to understand the thresholds at which the interplay of norms and punishment becomes (in)effective. A future avenue of research should examine interventions that prevent self-serving justifications in the presence of descriptive/normative information that is used in both small- and large-scale norm-based interventions (Hallsworth et al., 2017; Schultz et al., 2018).

signal only a descriptive norm, and thus provides a much weaker justification for punishment (Eriksson et al., 2015; Bicchieri, 2016). As a result, it is much less successful than normative information in making punishment effective. The consequent resentment is consistent with the shift from complete conformity to complete violation. If true, such resentment towards punishment is particularly likely when the enforced rule requires the agent to give up a large amount of earnings (i.e. in the high conformity cost condition). An examination of open-ended questions in our post-experimental questionnaire did not yield much support for this explanation in our context, which allows us to conclude that participants perceived punishment as completely appropriate. A comprehensive text file including all written answers is available upon request.

References

Abbink, K., Gangadharan, L., Handfield, T., and Thrasher, J. (2017). Peer punishment promotes enforcement of bad social norms. *Nature communications*, 8(1):609.

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.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer.

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. K. (2013). Norm manipulation, norm evasion: experimental evidence. *Economics & Philosophy*, 29(2):175–198.

Bicchieri, C. and Dimant, E. (2018). It's not a lie if you believe it: Lying and belief distortion under norm-uncertainty. Technical report, Philosophy, Politics and Economics, University of Pennsylvania.

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.

Bolton, G., Dimant, E., and Schmidt, U. (2018). When a nudge backfires: Using observation to promote pro-social behavior. Technical report, Mimeo.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

Bursztyn, L., Fiorin, S., Gottlieb, D., and Kanz, M. (2015). Moral incentives in credit card debt repayment: Evidence from a field experiment. Technical report, National Bureau of Economic Research.

Calabuig, V., Fatas, E., Olcina, G., and Rodriguez-Lara, I. (2016). Carry a big stick, or no stick at all: Punishment and endowment heterogeneity in the trust game. *Journal of Economic Psychology*, 57:153–171.

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.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

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

Fehr, E. and Williams, T. (2017). Creating an efficient culture of cooperation.

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.

Galbiati, R., Vertova, P., et al. (2008). Obligations and cooperative behaviour in public good games. *Games and Economic Behavior*, 64(1):146–170.

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.

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.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American economic review*, 90(4):1072–1091.

Kraft-Todd, G., Yoeli, E., Bhanot, S., and Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, 3:96–101.

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

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.

Muthukrishna, M., Francois, P., Pourahmadi, S., and Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, 1(7):0138.

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

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.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2018). The constructive, destructive, and reconstructive power of social norms: Reprise. *Perspectives on psychological science*, 13(2):249–254.

Spiekermann, K. and Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, 96:170–183.

Stigler, G. J. (1970). The optimum enforcement of laws. *Journal of Political Economy*, 78(3):526–536.

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

Tangney, J. P., Boone, A. L., and Baumeister, R. F. (2018). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In *Self-Regulation and Self-Control*, pages 181–220. Routledge.

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.

Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., and Conte, R. (2014). The norm-signaling effects of group punishment: combining agent-based simulation and laboratory experiments. *Social science computer review*, 32(3):334–353.

Xiao, E. et al. (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. Robustness Checks and Additional Figures for Trustee Behavior

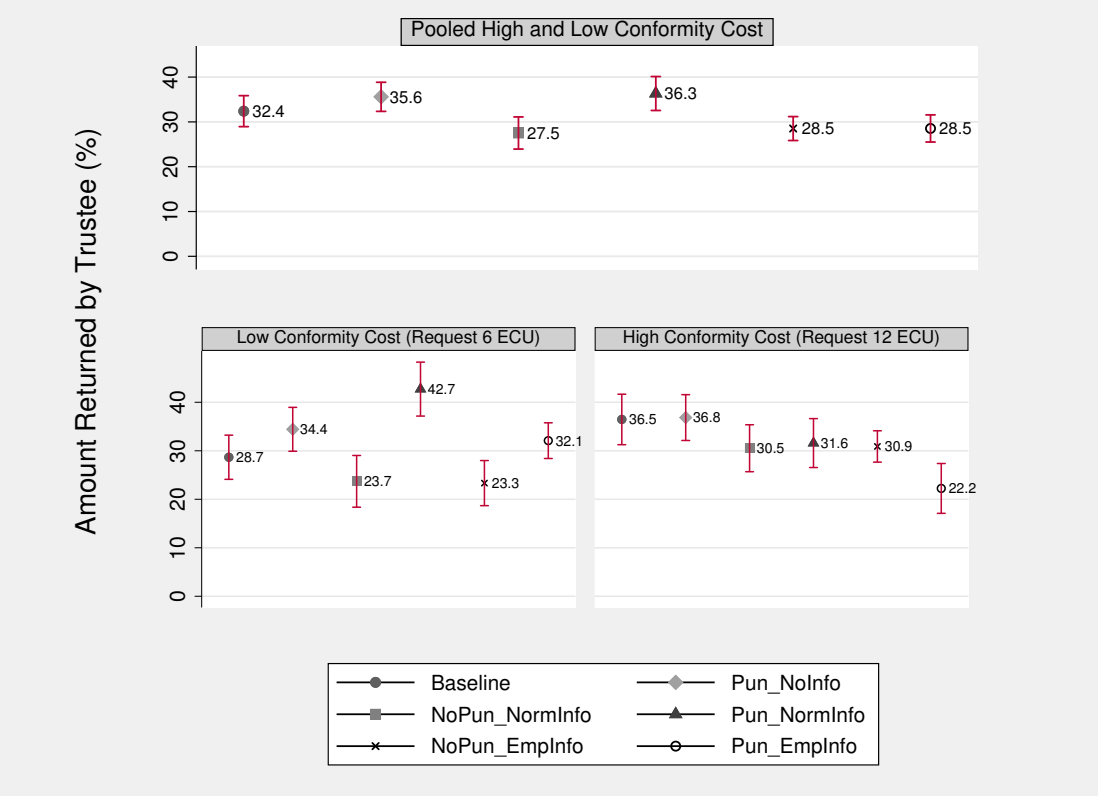


Figure A.1: Amounts returned by trustees as percentages of amount received from investors; upper part indicates pooled amounts; lower part indicates amounts per LCC vs. HCC; Baseline: no punishment or norm information; Pun_NoInfo: punishment (5 ECU) without norm information; NoPun_NormInfo: no punishment 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. Only significant differences are indicated at the conventional levels of *p<0.1, **p<0.05, ***p<0.01. Vertical lines represent 95% CIs.

DV: Amount Returned by Trustee	**Low Conformity Cost**		**High Conformity Cost**	
	(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 A2: 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 Conformity Cost (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 (ECU amount received from an investor in previous round, which indicates whether trustee faced a high or low conformity cost condition). Significance levels: $p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

DV: Amount Returned by Trustee	(1)	(2)
Punishment	1.672	1.389
	(5.202)	(4.894)
Normative Information	-8.328	-8.889
	(5.667)	(5.444)
Empirical Information	-5.457	-5.246
	(5.112)	(5.027)
Conformity Costs	1.681	1.999
	(1.496)	(1.4755)
Punishment × Normative Information	16.232**	19.343***
	(7.740)	(7.478)
Punishment × Empirical Information	3.467	3.654
	(7.070)	(6.889)
Punishment × Normative Information × Conformity Costs	-10.785***	-10.636***
	(3.912)	(3.906)
Punishment × Empirical Information × Conformity Costs	-8.940**	-9.542**
	(3.796)	(3.837)
Round		-0.525***
		(0.164)
Gender		2.651
		(3.143)
Self-Control		4.043**
		(1.570)
Risk		0.153
		(0.684)
Constant	32.643***	33.270***
	(3.888)	(5.511)
Observations	1446	1446

Table A1: 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), Conformity Costs (1 = high), Remaining coding of control variables the same as in Table 2. Significance levels: * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

B. 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.

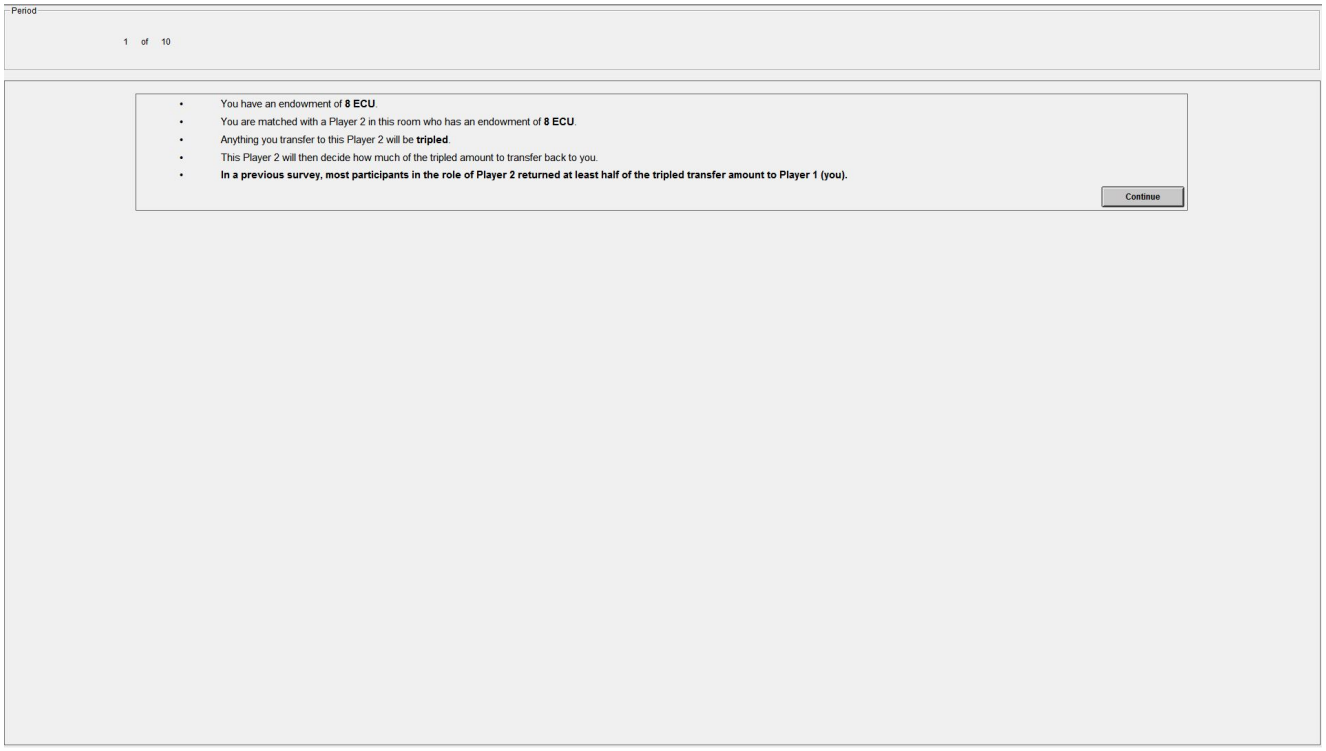
Player 1 will not know the result of each round (i.e. Player 1 will not know Player 2's decision in each round) until all the 10 rounds have finished. After all the 10 rounds have finished, each Player 1 will learn the matched Player 2's decision and the payoff outcomes in each round. Each Player 2 will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

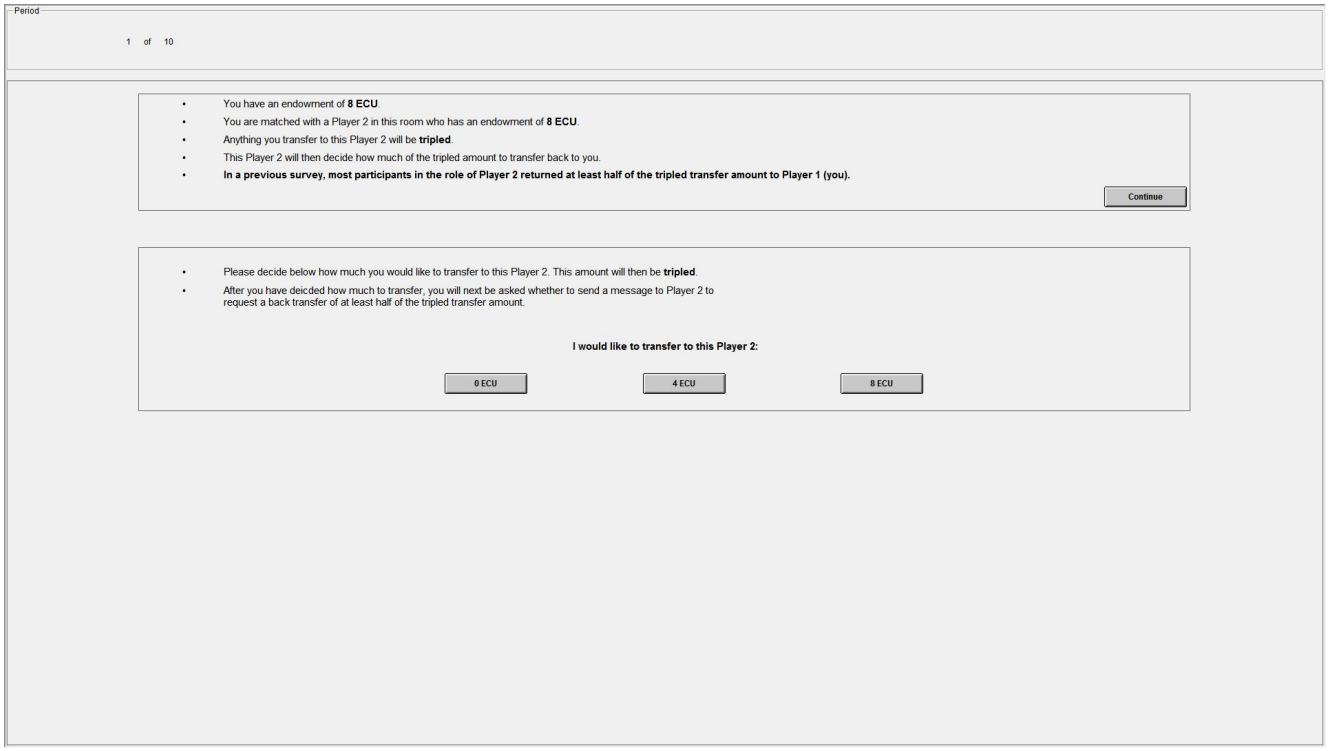
One round will be chosen at random and Player 1 and 2 will be paid according to the outcome of that round.

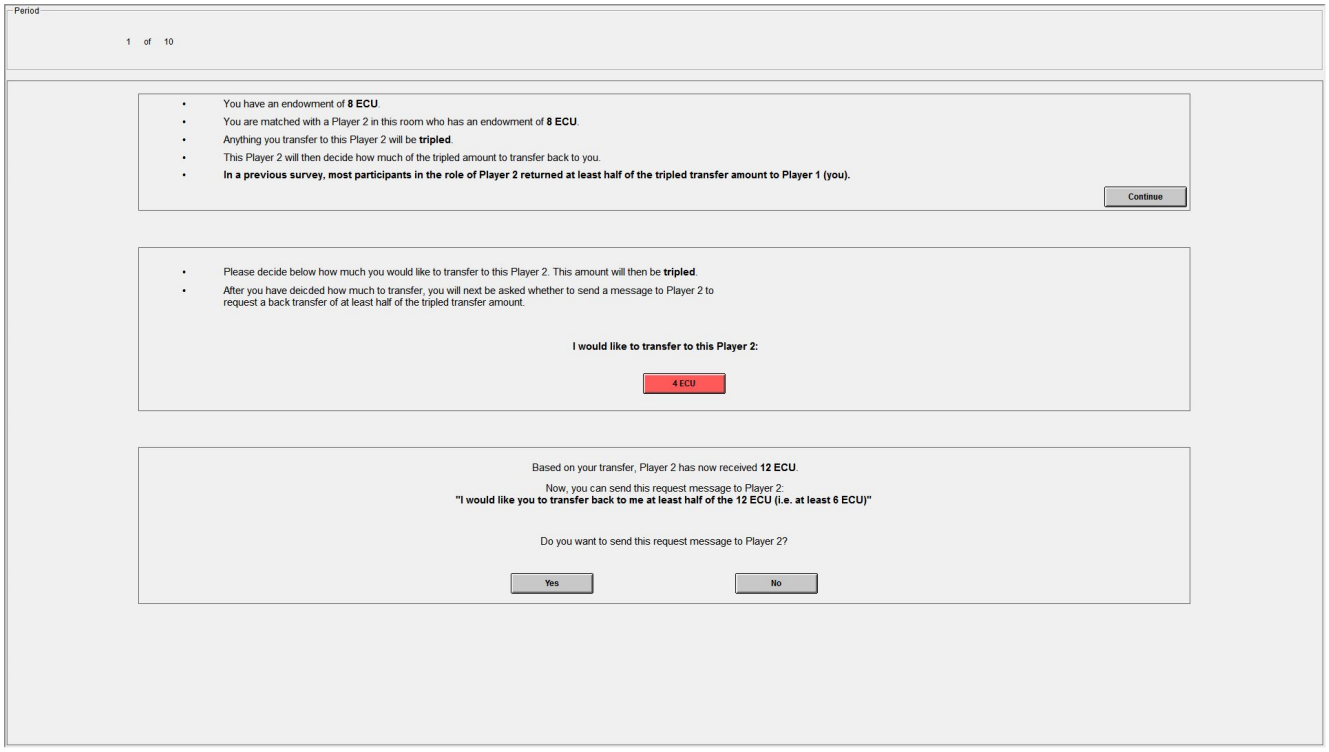
C. Screenshots of Experimental Procedure

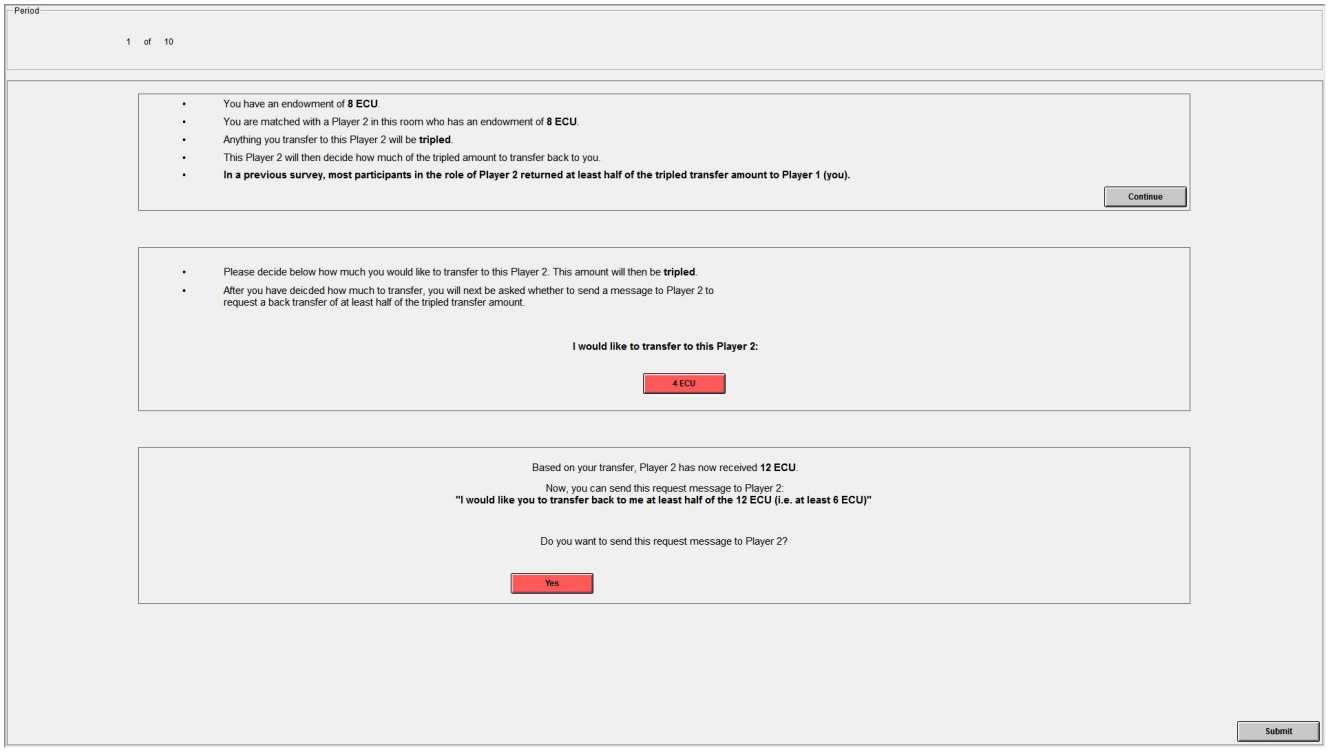
Here, we exemplarily present the screenshots for Treatment 5 (Punishment + Empirical Information). Differences to the other treatments are as previously explained in the experimental instructions. That is, indication of punishment and normative / empirical information was presented where the experimental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

Investor

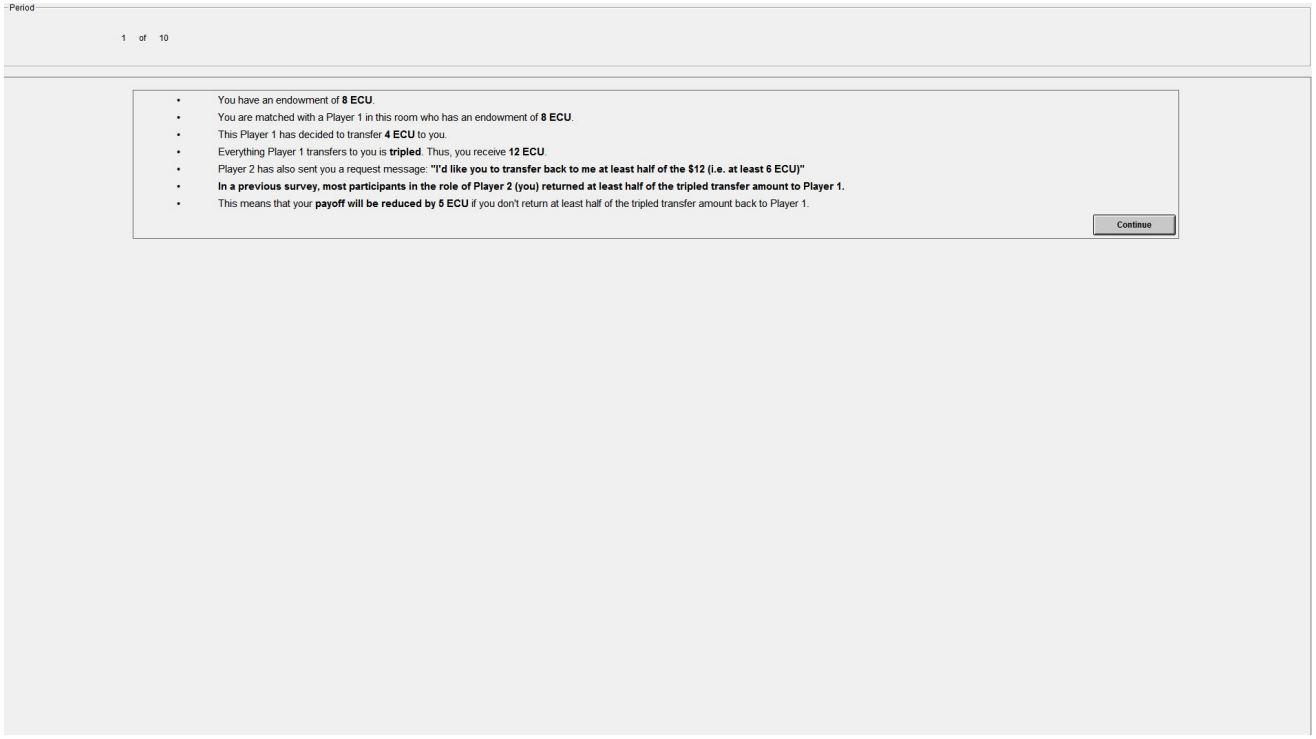


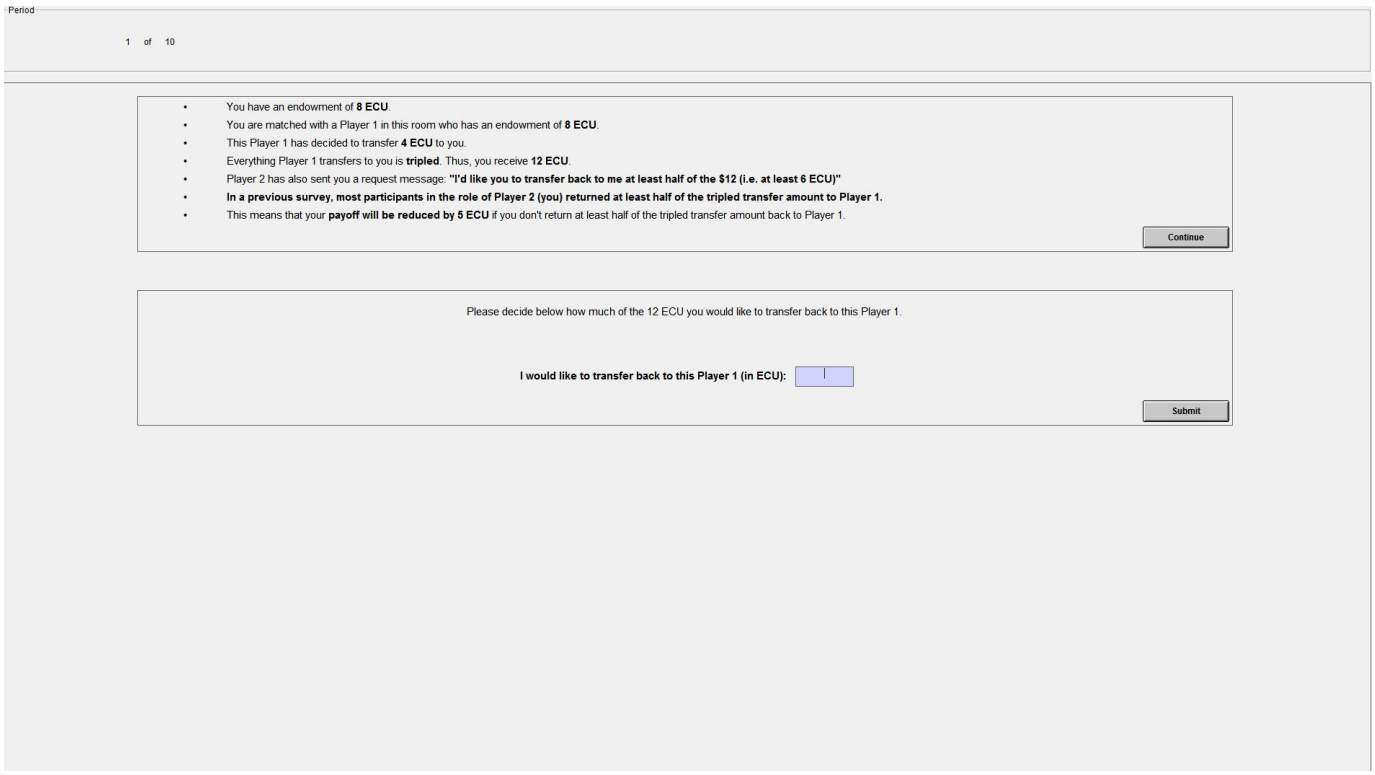

Period
1 of 10
You have an endowment of 8 ECU.
You are matched with a Player 2 in this room who has an endowment of 8 ECU.
Anything you transfer to this Player 2 will be tripled.
This Player 2 will then decide how much of the tripled amount to transfer back to you.
In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transfer amount to Player 1 (you).
Continue
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:
0 ECU
4 ECU
8 ECU


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).
Continue
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
No


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).
Continue
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
Submit

Trustee


Period
1 of 10
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.
Continue


Period
1 of 10
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.
Continue
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):
Submit

End of the round screenshot (Investor and Trustee)