CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2017-14

Cristina Bicchieri,
Eugen Dimant & Erte Xiao

November 2017

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

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CEDEX

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

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. In such studies, the enforcement of a behavioral rule to cooperate originates from a single party. This feature may raise concern about the legitimacy of the rule and thereby make it easy for the agents to take a penalty and excuse their selfish behavior. We address the question of punishment legitimacy in our experiment by shedding light upon the importance of social norms and their interplay with punishment mechanisms. We show that the separate enforcement mechanisms of punishment and norms cannot achieve higher cooperation rates. In fact, conformity is significantly increased only in those cases when social norms and punishment are combined, but only when cooperation is cheap. Interestingly, when cooperation is expensive we find that the combination of punishment and empirical information about others conformity can also have traceable detrimental effects on conformity levels. Our results have important implications for researchers and practitioners alike.

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

1. Introduction

A large body of research has examined the effect of punishment on conformity. The standard economic theory of punishment has focused on how incentives can change pay-

[*] We thank Daniel Houser, Gordon Kraft-Todd, 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 Workshop on Social Norms in Epistemic Communities at Monash University, the 2017 ANZWEE annual conference at University of Melbourne, the 7th Biennial Workshop on Social Dilemmas, and the 2017 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 13, 2017

offs and thereby influence outcomes (Becker, 1968). It follows that when punishment is sufficiently severe as to overwhelm the expected benefit of defection, it can prevent opportunistic behavior. However, severe punishment usually requires costly monitoring and can even have undesirable side effects. As a result, punishment in practice is often weak (Tyler, 2006) meaning that the cost of punishment is no higher than the cost of compliance.[1]

Numerous studies show that even weak punishment can also backfire and even lead to the enforcement of bad social norms, though (Gneezy and Rustichini, 2000; Fehr and Rockenbach, 2003; Abbink et al., 2017). For example, Fehr and Rockenbach (2003) show that trustees return less when the investor imposes a weak punishment to enforce the desired return amount. Houser et al. (2008) provide further evidence that this detrimental effect of punishment is significant even when the punishment is naturally imposed in the absence of any negative intention from the investor (for a review see Bowels and Polania-Reyes, 2012). Conversely, there is evidence that weak punishment can promote pro-social behavior when it is properly designed. For example, Tyran and Feld (2006) show that, in a public goods environment, cooperation significantly improves when punishment is imposed by group members rather than exogenously. Xiao and Houser (2011) show that publicly implemented weak punishment can make a social norm salient, and therefore more effectively promotes conformity than privately implemented punishment. The success of weak punishment seems to be due to its social dimension as a group decision or a public event. In both cases the subject of punishment, the violation of shared, established norms, is made salient.

This paper contributes to this literature by noting that in a naturally occurring environment, punishment is typically related to a social norm violation. More precisely, it is often made clear that the behavioral rule enforced by punishment is a shared norm, rather than the self-interested preference of the punisher. We conduct a controlled laboratory experiment to examine how providing the information that the enforced rule is consistent with a shared norm can affect the outcome of punishment. The introduced punishment is not equilibrium shifting which sets us apart from much of the existing research. We focus on the conditions where punishment is weak, in that the cost of punishment is no higher than the cost of compliance, so that monetary incentives cannot be the dominant driver of decisions.

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

Existing empirical research on social norms, both in the laboratory and in the field, shows that one may foster more pro-social behavior by simply providing information about the norm (Goldstein et al. 2008; Bicchieri and Xiao, 2009; Kraft-Todd et al., 2015). Importantly, social norm theory makes a distinction between normative and empirical information (Bicchieri, 2006). Empirical information alone may point to a descriptive norm (what most people do), but not necessarily to a social norm proper (what most people do *and* believe one should do), due to ambiguity as to the underlying normative appropriateness of the behavior. Normative information instead provides a stronger, unequivocal signal that an action is appropriate, thus pointing to a social norm. 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 get only one type of information, it is important to investigate the potentially distinct effects on behavior of such information, especially when accompanied by punishment. By adding normative information about the enforced behavior, it is made clear that any violation is *wrong.* If the enforced rule is seen as consistent with a shared social norm, and thus justified, punishment may increase the psychological cost of violation and enforce compliance. Such an effect, however, may be weaker when the enforced rule is only supported by empirical information and the prescribed behavior is perceived as simply a *deviation* from what others would do. Thus, we hypothesize that people tend to consider punishment as more justified when the enforced behavior is also presented as the right decision rather than when it is simply presented as what others did 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 (a social norm) rather than what a majority *does* or *would* do (a descriptive norm).

Our experiment consists of six treatments. We systematically introduce punishment, normative or empirical information, and the respective combinations thereof within a trust-game setting. Subjects were assigned either to the role of investor or trustee. In the baseline, the investor had to decide whether to transfer any amount of her endowment to the trustee. Any transfer amount was tripled and received by the trustee. The trustee then had to decide how much of the tripled transfer amount to return to the investor. Before the trustee made a decision, the investor could choose to send a request message to the trustee. The request message was in a fixed form, asking the trustee to return at least half of the tripled transfer amount. The message was non-binding in that the trustee was free to return any amount regardless of whether the investor had chosen to send the message.

In three treatments with punishment, the investor's request message becomes binding in that if the trustee returned less than 50%, he/she would receive a penalty of a fixed amount. Since we are only interested in the case when punishment is weak and not equilibrium-shifting, the penalty was designed to be always smaller than the 50% return. The three treatments with punishment vary on whether participants were informed that the request message is consistent with norm information. The norm information either tells the subjects in a descriptive way that in a previous session many trustees returned at least 50% (empirical information) or in a prescriptive way that many participants think the trustees should return at least 50% (normative information). To control for the effect of norm information alone, we also include another two treatments where punishment is absent and only the empirical or normative information is provided.

We find that only the composite effect of normative information and punishment increases conformity significantly, while the separate enforcement mechanisms of punishment and normative information cannot achieve this result by themselves. Interestingly, we find that the combination of punishment and empirical information has traceable detrimental effects on conformity levels.

Our findings contribute to the understanding of the impact of monetary incentives on pro-sociality and cooperation. 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, and sometimes even in combination with social norms, is often less effective or measurably destructive in guiding and changing behavior to the better. Examples include the elimination of FGM in Africa and foot binding in China, banning dueling in Europe, reducing smoking and tax evasion, among others, and is particularly driven by a rift between laws and social norms (Posner, 2002; Benabou and Tirole, 2011; Acemoglu and Jackson, 2016; Bicchieri 2016; Tankard and Paluck, 2016). Research has also suggested that too much transparency about descriptive norms can backfire when social learning of deviant behavior occurs and subsequently leads to imitation of illicit behavior, such as for example corruption (Fisman and Miguel, 2010; Muthukrishna et al., 2017). We offer a new potential explanation for the often-missing positive and sometimes detrimental effect of punishment observed in previous studies. At the center is the argument that individuals may not comply with punishment when they view the prescribed rule as unjustified – unjustified in that it just served the other individual's self-interest. For punishment to be effective, it is important to ensure the perceived legitimacy of the enforced rule. Supporting the rule with normative information can serve such a purpose. In contrast, empirical information does not help. This is in line with the findings by Bicchieri and Marini (2016) who show that the only cases in which a law against FGM was successful are cases in which the government is trusted and there is a norm of legal obedience. The negative result of the combination of punishment with empirical information raises concerns about the recent wave of social norm nudging. It implies that we should carefully distinguish between descriptive and social norms interventions. In our experiment, it is more acceptable to punish *wrongness* than to punish *deviation.* Another important consideration is that empirical information can often be manipulated in a self-serving way. Whereas normative information is unambiguous in pointing out what behavior is socially proper, an empirical message may lend itself to different interpretations if it is not clear who is included in the conforming majority. For this reason, too, we suggest caution in designing norm interventions.

2. Experiment Design and Predictions

We recruited a total of 418 participants across six treatments at the University of Pennsylvania. The experiment is based on a modified version of the investment game (Berg, Dickhaut and McCabe, 1995). At the beginning of the experiment, 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 ECU (2 ECU = $1) and was randomly matched with another participant in a different role.

Treatment	Punishment	Normative Information	Empirical Information	#
Baseline	No	No	No	60
Pun_NoInfo	Yes	No	No	68
NoPun_NormInfo	No	Yes	No	58
Pun_NormInfo	Yes	Yes	No	62
NoPun_EmpInfo	No	No	Yes	94
Pun_EmpInfo	Yes	No	Yes	76

Table 1: Overview over the treatments and number of subjects assigned to each treatment.

Treatments varied by a systematic variation of punishment (absent, present) and norm information (absent, a normative message about what ought to be done, an empirical message about what other participants did) and the combinations thereof. Table 1 depicts the breakdown of participants by treatment.

2.1. Treatments

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

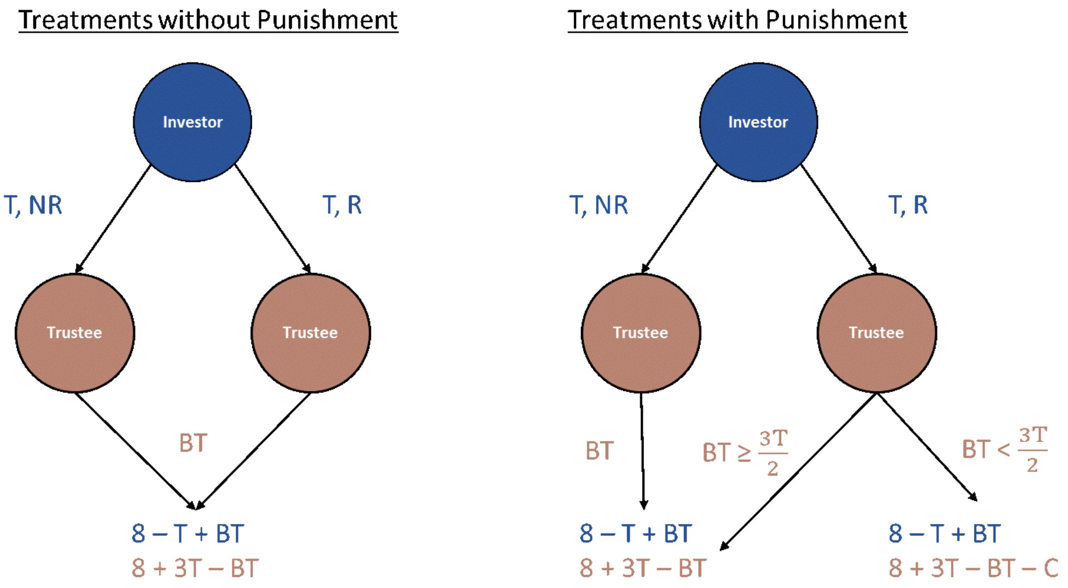


Figure 1: *Sequence of actions and payoff structure in treatments with and without punishment.* ***T:*** *Investor's transfer to trustee.* ***(N)R:*** *Investor's decision to (not) send a return request message to the trustee.* ***BT:*** *Trustee's back-transfer to the Investor.* ***C:*** *Trustee's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

Baseline

At the beginning of each round, the investor had to decide how much to transfer to the trustee. The transfer amount (T) could be either 0 ECU, 4 ECU, or 8 ECU. We limit the action space of the investor to allow differentiation between low and high cost of conformity across all treatment specifications as we will explain in more details below. It was disclosed that the transferred amount was multiplied by a factor of 3 by the experimenter. When deciding how much to transfer, the investor also had to decide whether to send a costless request message (R or NR) to the trustee, indicating whether he/she wanted the trustee to return 50% of the transfer amount. The message was in a fixed form. An example of the message when the transfer amount is 4 ECU is as follows: *"I'd like you to transfer back to* *me at least half of the 12 ECU (i.e. at least 6 ECU)."* All subjects knew that the investor had the choice to (not) send the return request message.

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 from [0, 3T].

To provide clean evidence for the effect of punishment (see below) on the trustees' return decisions, the investors did not know the trustees' return amount in each round until all the 10 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, we can exclude the possibility that the trustees' return behavior in each round might influence investors' transfer decisions in the next round, influencing the trustee's behavior. One round was randomly chosen as the payoff round and participants were paid the amounts they earned in that round.

Punishment Treatments

In the three treatments with the punishment opportunity (Pun_NoInfo, Pun_NormInfo and Pun_EmpInfo), we followed a design structure similar to Houser, Xiao, McCabe and Smith (2008) (also see, Fehr and Rockenbach 2003). Here subjects were also told that if the investor sent a return request message, the trustee would receive a payoff cut of 5ECU 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. That is, in the treatments with punishment, by sending the return request message, the investor also imposed a punishment on the trustee if he/she returned less than half of the tripled transfer amount.

Norm Information Treatments (Normative or Empirical)

We adopt 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 particular, in the treatments with the normative information, the instructions included the following lines: "*In a previous survey, most participants said that Player 2 should return at least half of the tripled transferred amount.*" In the treatments with the empirical information, the instructions included the following lines: "*In a previous survey, most participants in the role of Player 2 returned at least half of the tripled* *transferred amount to Player 1.*"[2]

To summarize, in the baseline condition, subjects play a trust game and the investor could send a non-binding request message asking the trustee to return at least 50%. 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%. The Pun_NormInfo (Pun_EmpInfo) treatments differ from the Pun_NoInfo treatment only in that participants were told that most participants thought trustees should (did) return at least 50%. The NoPun_NormInfo and NoPun_EmpInfo differ from the Pun_NormInfo and Pun_EmpInfo only in that the request message is non-binding without the punishment consequence as in the Baseline. These last two treatments let us to examine whether any difference between the Pun_NormInfo (Pun_EmpInfo) and the baseline can be attributed to the normative (empirical) information alone.

2.2. Procedure

Participants were recruited through the Experiments@Penn software at the University of Pennsylvania. The experiments were conducted at the BeLab. The average duration of a session, which included the experiment and a post-experimental questionnaire, was 45 minutes and the average hourly wage was $18, including a $10 show-up fee. The experiment was programmed using z-Tree (Fischbacher, 2007). Across all treatments, the participants were 22.2 years old and the proportion of female participants was 62.7%, on average.

2.3. Main Predictions

In our experiment, the punishment is always weak in that the required minimum return amount, either 6ECU or 12ECU, is always higher than the fixed fine of 5 ECU. Based on the previous studies (Fehr and Rockenbach, 2003; Houser et al., 2008), we predict the punishment to not have a significant impact on the return rate compared with the case when no punishment is present.

$$\textbf{Prediction 1}\text{: } \text{Return}^{Base} \geq \text{Return}^{Pun\_NoInf}$$

Our main hypothesis is that punishment can be more effective when it is made clear that the enforced rule is consistent with a shared social norm. Furthermore, while the

[2] Like, Bicchieri and Xiao (2009), the data was based a previous pilot session and therefore was truthful. Data available upon request.

normative information directly and explicitly points out that the punished behavior is wrong, the empirical information only indirectly implies the wrongness of the violation and, ultimately, can be vague. Thus, we may hypothesize that normative information is more effective in justifying the punishment. Our hypotheses generate the following prediction[3]:

$$\textbf{Prediction 2}: \text{Return}^{Pun\_NormInfo} > \text{Return}^{Pun\_EmpInfo} \geq \text{Return}^{Base}$$

Note that, independent of the effect of punishment, information alone may potentially achieve a higher return rate than the Baseline. Data from the two information only treatments can shed light on this pure information effect. If the behavioral pattern in Prediction 2 is mainly driven by the information effect alone, we should expect to observe the following return outcome:

$$\text{Return}^{Pun\_NormInfo} \cong \text{Return}^{NoPun\_NormInfo}$$

and

$$\text{Return}^{Pun\_EmpInfo} \cong \text{Return}^{NoPun\_EmpInfo}$$

3. Results

Our hypothesis regards the trustees' return behavior in different conditions. We thus focus on the trustees' average transfer-back behavior in this section.[4] We report the investors' transfer behavior at the end of the section. In the experiment, punishment plays a role only when the investors decide to send the request message. On average, investors sent the return request message in 93% of the cases. To allow for comparability across treatments, our analysis includes only the cases where a return request message was sent.[5]

The compliance cost for the trustees varies with respect to whether the investors send 8ECU or 4ECU. We denote the case when the investor sends 8 ECU as *High Conformity Cost* and the case when the investor sends 4 ECU as *Low Conformity Cost.* In the *High Conformity Cost* condition compliance requires one to return 12 ECU. In the *Low Conformity Cost* condition compliance only requires one to return 6 ECU. Hence, we test the predictions for the two conditions separately.

[3] To allow for comparability across treatments, both our predictions and the results analysis in section 3 includes only the cases where a return request message is sent.

[4] Figure A1 in the Appendix reports trustee return behavior across all rounds and treatments.

[5] Our regression analyses presented in Section 3.4 reveal that our results are robust to the inclusion of the comparably rare instances in which the investor did not send a return request message.

In what follows, we analyze mean differences across treatments.[6] We first report the return outcome in the presence of punishment. We then examine the effect of adding empirical or normative information to the punishment condition. Lastly, we check whether the composite effect of punishment and norm-related information might be fully attributed to the addition of such information. In anticipation of the results, we find that, supporting Prediction 1, punishment alone is not successful in improving return rates, especially in the *High Conformity Cost* condition. Likewise, neither empirical nor normative information alone achieves a higher return rate than the Baseline. The combination of punishment and normative information triggers substantial behavioral change, but only when the compliance cost is relatively low. Interestingly, the combination of punishment and empirical information is not only ineffective when the compliance cost is low, but it has detrimental effects on conformity rates when the compliance cost is high. We will argue that the result that normative information enhances the efficacy of punishment is consistent with Prediction 2. However, the result that such a positive effect is not observed when the compliance cost is high indicates that the benefit of normative information is subject to the cost of conformity. Moreover, the unexpected detrimental impact of empirical information on punishment suggests moral wiggle-room and self-serving bias may arise when the compliance cost is too high.

3.1. Effect of Punishment Alone

Figure 2 reports the average return percentage in the Baseline and punishment treatments. In both the *High Conformity Cost* and the *Low Conformity Cost* conditions, punishment does not significantly increase the return levels of trustees. In the *High Conformity Cost* condition, the return percentage is almost the same when there is a punishment threat as compared to the Baseline (36.8% vs. 36.5%, BSM, p=0.91). In the *Low Conformity Cost* condition, punishment increases the return slightly, but the increase is only marginally significant (34.4% vs. 28.7%, BSM, p=0.10).

Houser et al. (2008) show that punishment leads to bimodal trustee behavior, where the return is either equal to or higher than the requested amount (complete conformity) or

[6] For the analysis of return behavior, we average individual behavior across rounds. We follow Moffat (2015) and employ the bootstrap two-sample t-test method (hereafter BSM) with 9999 replications to analyze mean differences of average return behavior. This has the advantage that we can retain the rich cardinal information in the data without making any assumptions about the distribution. Unless noted otherwise, the use of non-parametric Mann-Whitney-U (hereafter MWU) tests yields results that are in line with our bootstrap approach.

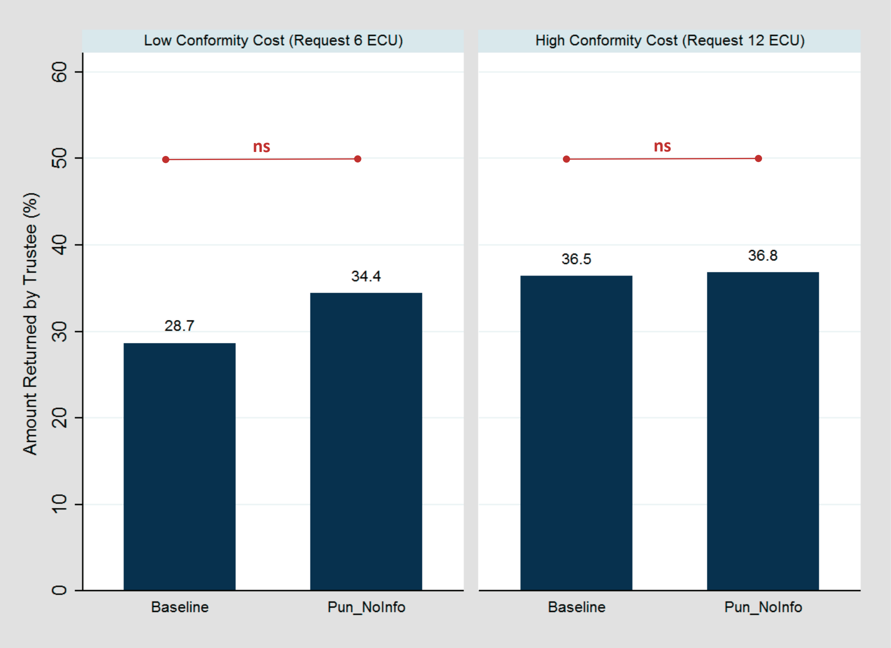


Figure 2: Amount returned by trustees as percentage of amount received from the investor.

where it is zero (complete violation). A further examination of the distribution of return percentages can thus provide additional insights in how punishment affects behavior. In line with this research, we classify the behavior of trustees into three types: *Complete Violation* of trust if returned amount = 0%; *Incomplete Conformity* if 0% <returned amount <50% and *Complete Conformity* if returned amount ≥ 50%.[7] Figure 3 plots the distribution of the three types in each of the four conditions. Using Kolmogorov-Smirnov (hereafter K-S) tests, we find that the distributions in the *Low Cost* condition are significantly different between the Baseline and punishment treatments (K-S, p<0.01). Consistent with Houser et al. (2008), we observe a bimodal return pattern under the punishment condition, and there is a significant decrease in the proportion of *Incomplete Conformity* (40% vs. 66.4%, BSM, p=0.04; 0% vs. 25%, BSM, p<0.01). While the proportion of *Complete Violation* types remained invariant (33.6% vs. 35.0, BSM, p=0.94), punishment significantly increases the

[7] For the analysis of types, we calculate three ratios for high and low conformity costs per participant. The ratios indicate the fraction of complete violation, incomplete conformity, and complete conformity at the individual level across all rounds. This is necessary to account for behavioral changes across all rounds and the fact that participants acted in environments with different conformity costs depending on the amount of money sent by the investor.

proportion of *Complete Conformity.* However, such a positive shift does not translate in a significant change in average return behavior as we reported above. The reason is that many of the *Incomplete Conformity* types in the Baseline were right below the 50% cut-off.

In contrast, in the *High Conformity Cost* condition, the difference between the punishment condition and the Baseline is relatively small and insignificant (K-S, p=0.33). While we observe significantly less *Incomplete Conformity* types in the punishment condition compared to the Baseline (1.7% vs. 18.5%, BSM, p<0.01), the effects of punishment on the other two types are not significant (*Complete Violation*: 29.6% vs. 26.1%, BSM, p=0.41; *Complete Conformity*: 68.7% vs. 55.4%, BSM, p=0.52).

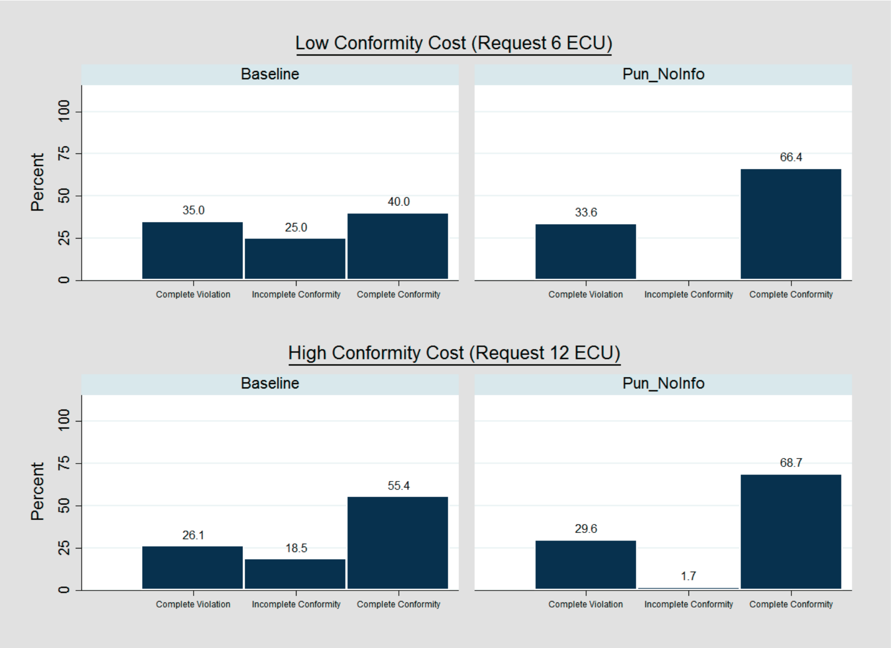


Figure 3: Distribution of return types in Baseline and Pun_NoInfo. In all the figures, significances indicate Kolmogorov-Smirnov distribution tests across treatments.

Overall, as in Houser et al. (2008), we observe that with punishment the investors get either what they want or nothing at all. However, Houser et al. (2008) also found a detrimental effect of punishment in that the punishment significantly increases the rate of complete defection (zero return in the trust game) when the required return amount is more than double the penalty amount. We do not observe such a detrimental effect in our experiment even in our *High Conformity Cost* condition. In addition to the potential subject pool differences, one possible explanation is that in our *High Conformity Cost* condition investors transfer all the endowment and request 50%. In Houser et al. (2008), the high compliance cost condition is defined purely based on the requested amount. The transfer amount can be small or high and the requested amount can also be small or high (e.g. more than 50%).

Next, we examine whether showing that the investor's request is consistent with, and is justified by, a norm alters the effect of punishment on the return levels of the trustees.

3.2. Effect of Punishment and Norm Information Combined

Figure 4 plots the average return in the Baseline, Pun_NoInfo and the two treatments in which punishment was combined with either normative information (Pun_NormInfo) or empirical information (Pun_EmpInfo).

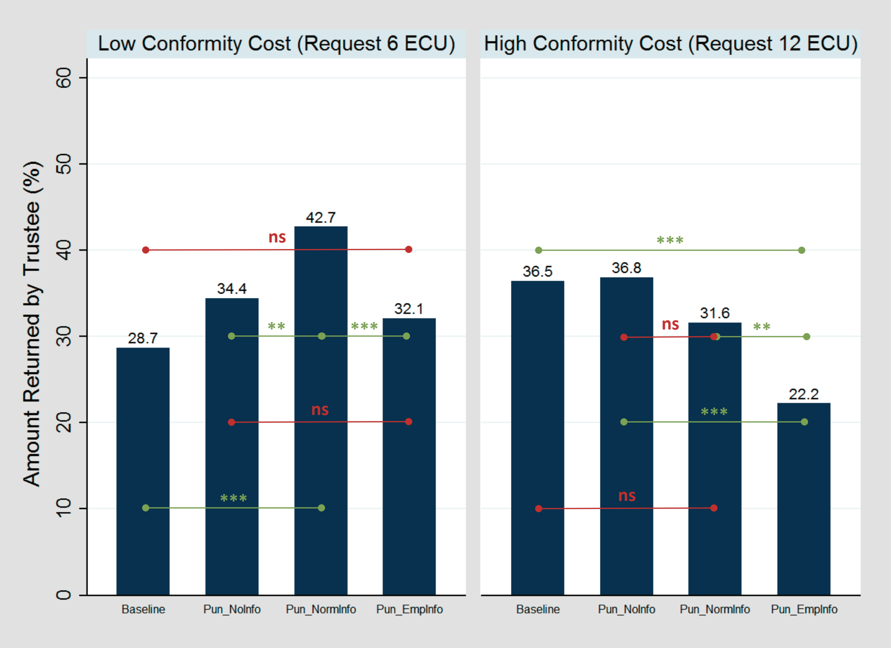


Figure 4: Amount returned by trustees as percentage of amount received from the investor.

When the cost of conformity is low (Low Conformity Cost), the combination of punishment and normative information leads to a significant increase in trustees' return behavior (42.7% vs. 28.7%, BSM, p<0.01), well above the 5.7% increase in the no-punishment condition. The return percentage is also significantly higher than the Pun_NoInfo treatment (42.7% vs. 34.4%, BSM, p=0.02). However, when the information is instead about what others do (empirical), the return in the Pun_EmpInfo treatment is only slightly higher than the Baseline and the difference is not significant (32.1% vs. 28.7%, BSM, p=0.25). The return in the Pun_EmpInfo condition is also significantly lower than the Pun_NormInfo treatment (32.1% vs. 42.7%, BSM, p<0.01). It should also be noted that the return in the Pun_EmpInfo is very close to the return of the punishment-only condition (32.1% vs. 34.4%, BSM, p=0.43). These results support Prediction 2 that punishment is more effective when it is made clear that the punished behavior is socially disapproved, but the positive effect is less likely when the punished behavior is perceived as simply a deviation from the majority behavior.

However, Prediction 2 does not seem to hold when the cost of conformity is high (*High Conformity Cost* condition). First, the return in the Pun_NormInfo is not significantly different in the Baseline and the Pun_NoInfo treatments (31.6% vs. 36.5%, BSM, p=0.18; 31.6% vs. 36.8%, BSM, p=0.18). Moreover, adding the empirical information seems to be even counterproductive in that the return is significantly lower than it was in either the Baseline or the Pun_NoInfo treatments. (22.2% vs. 36.5%, BSM, p=0.01; 22.2% vs. 36.8%, BSM, p=0.01).

These results suggest that the cost of conformity and the kind of norm information (social or descriptive norm) influence the benefit of making the alignment of the enforced rule and a norm salient. As we hypothesized, normative information is helpful, but its effect is subject to the cost of conformity. When the cost is too high adding normative information does not help improve the efficacy of punishment. Empirical information does not help when the cost of conformity is low or when it is high. In fact, the effect of empirical information can even be negative when the conformity cost is high. The detrimental effect of empirical information is a novel finding that was not considered in our hypothesis. A further examination of the return distribution reported below (see Figure 5) reveals that the decrease in average return in the Pun_EmpInfo treatment mainly comes from the shift to complete violation.

First, we examine the *Low Conformity Cost* condition. As shown in Figure 5, the distribution of behavior in the Pun_NormInfo and Pun_EmpInfo treatments are significantly different from the Baseline treatment (K-S, p<0.01). Similar to the Pun_NoInfo treatment that we reported above, both the Pun_NormInfo and the Pun_EmpInfo treatments generate a bimodal pattern of the return distribution. This is represented by a significant decrease in *Incomplete Conformity* types in both treatments (2.3% vs. 25.0%, BSM, p<0.01; 2.9 vs. 25%, BSM, p<0.01).

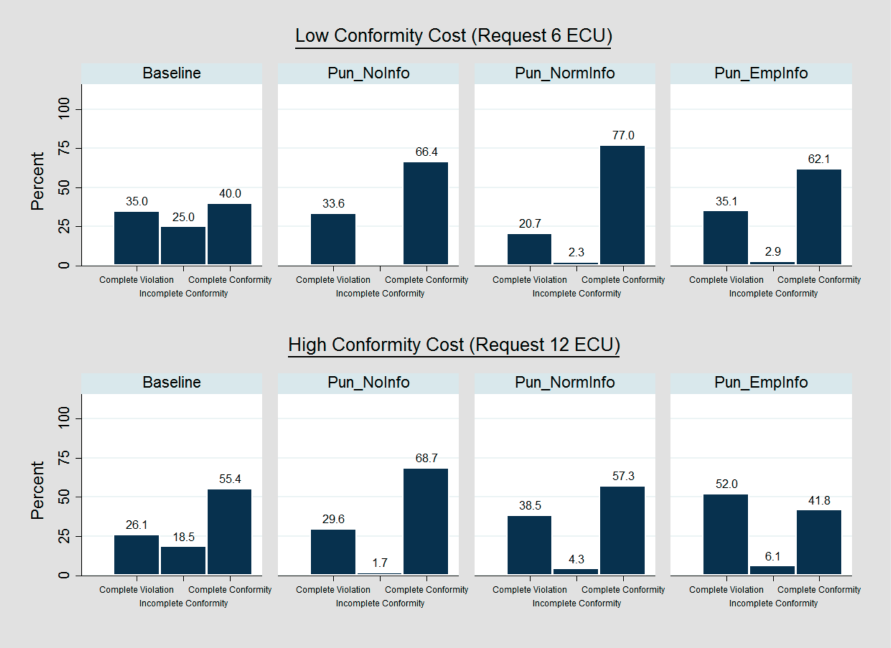


Figure 5: Distribution of return types by treatment.

Compared with the Baseline treatment, we find a significant increase in *Complete Conformity* types (77.0% vs. 40.0%, BSM, p<0.01) and a substantial decrease of *Complete Violation* types (20.7% vs. 35.0%, BSM, p<0.01) in the Pun_NormInfo treatment. Such a significant shift in Pun_NormInfo cannot be attributed to punishment alone. To see this, we compare Pun_NormInfo and the Pun_NoInfo treatments and observe that Pun_NormInfo achieves 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 not only the rate of complete compliance, but also by decreasing the rate of complete violation. Such a positive effect, however, is absent when adding the empirical information to the punishment.

The distribution data in Figure 5 further reveals that while Pun_EmpInfo significantly increases the percentage of *Complete Conformity* types (62.1% vs. 40.0%, BSM, p<0.01), such an effect seems to be mainly due to punishment itself. That percentage is very close to what we observe in the Pun_NoInfo (62.1% vs. 66.4, BSM, p=0.48). We find no significant change in the *Complete Violation* types compared to the Baseline (35.1% vs 35.0%, BSM, p=0.74), which is also close to the Pun_NoInfo treatment (35.1% vs. 33.6%, BSM, p=0.66).

Thus, we conclude that when the cost of conformity is low the return patterns across treatments are consistent with Prediction 2. 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 information is normative. The empirical information that signals only a descriptive norm, however, provides a much weaker justification for punishment. As a result, it is much less successful than the normative information in making punishment effective.

Next, we turn to the treatment comparisons of the return distributions in the *High Compliance Cost* condition.

First, we observe the bimodal return pattern remains in both the Pun_NormInfo and the Pun_EmpInfo treatments due to the significant decrease of *Incomplete Conformity* compared with the Baseline (4.3% vs. 18.5%, BSM, p<0.01; 6.1% vs. 18.5%, BSM, p=0.04). As we have 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 the empirical information is even detrimental. Figure 5 further reveals that the detrimental effect observed in the Pun_EmpInfo treatment is mainly driven by the significant increase of *Complete Violation* (52.0% vs. 26.1%, BSM, p<0.01). At the same time, we only observe a slightly significant increase of *Complete Violation* in Pun_NormInfo compared to the Baseline (38.5% vs. 26.1%, BSM, p=0.06). *Complete Conformity*'s frequency is also marginally less significant in Pun_EmpInfo than in the Baseline (41.8% vs. 55.4%, BSM, p=0.06). Such a negative shift does not occur in the Pun_NormInfo treatment (*Complete Conformity*: 57.3% vs. 55.4%, BSM, p=0.58).

Figure 3 demonstrates that the significant negative shift of the conformity types is not observed in the Pun_NoInfo treatments. These results suggest that the detrimental effect is mainly due to adding the empirical information to punishment rather than just the punishment itself. The significantly higher *Complete Violation* rate suggests that either the empirical information is too weak to justify punishment in a high cost context, thus leading to resentment, or that it may provide an incentive to form self-serving beliefs to justify non-compliance (Bicchieri and Dimant, 2017). We discuss this result in the Discussion section. Next, we examine whether the observed composite effect of norm information and punishment in the Pun_NormInfo and Pun_EmpInfo can be attributed to the norm information alone.

3.3. Effect of Norm Information Alone

We find the norm information alone does not have a significant impact on the return. Figure 6 plots the average return percentage in the Baseline, the NoPun_NormInfo and NoPun_EmpInfo treatments. In both the Low and the High Conformity Cost conditions, the differences in the average return between the Baseline and the two information only treatments are small and insignificant. (low cost: 23.7% vs. 28.7%, BSM, p=0.17, and 23.3% vs. 28.7%, BSM, p=0.11; high cost: 30.5% vs. 36.5%, BSM, p=0.11, and 30.9% vs. 36.5%, BSM, p=0.11).

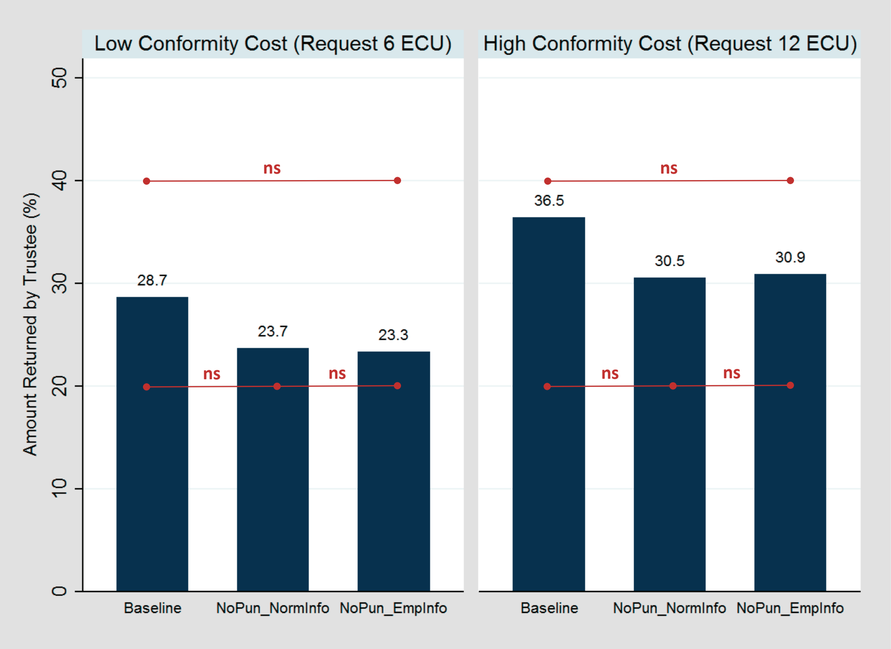


Figure 6: Amount returned by trustees as percentage of amount received from the investor.

Figure 7 reports the distribution of the three types as defined above in the two information only treatments and the Baseline specification. We do not observe the extremely low frequency of the *Incomplete Conformity* type as we observed in the Pun_NormInfo and Pun_EmpInfo. Moreover, none of the pairwise distribution comparisons between the information only treatments and the Baseline are significant. These results suggest that the composition effect of norm information and punishment cannot be attributed to the norm information alone, hence supporting the robustness of our previous findings.

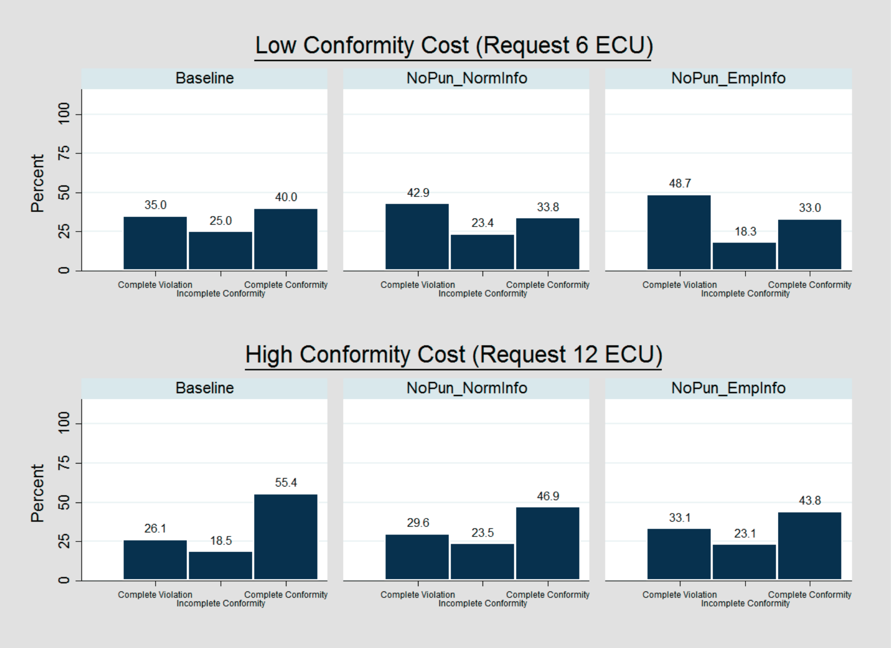


Figure 7: Distribution of return types by treatments.

3.4. Regression Results

In what follows, we harmonize our previous results by analyzing our data through the lens of multivariate regressions.[8] We do this by employing different variants of regression models to assess the robustness of our results. As our results in Table 2 indicate, 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. The results are as follows:

Result 1: Neither punishment nor information can significantly affect return behavior in isolation. This holds true regardless of the magnitude of 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 amounts to return rates that are between 11-13% higher compared to the Baseline specification in which neither punishment nor norms were present. As presented in Table

[8] In all cases, we employ random effects panel regressions with standard errors clustered at the participant level.

A1 in the Appendix, a further analysis reveals that when compliance is expensive, the combination of punishment and normative information leads to a substantially significant drop in return rates compared to behavior when the costs for compliance are low.

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 some 10-13% relative to the Baseline specification. A further examination in Table A1 in the Appendix reveals that return rates are also significantly lower compared to the combination of punishment and empirical information when conformity costs are low.

The coefficients from our controls suggest that return behavior declines over time, and that participants with higher self-control (as measured by our scale taken from Tangney et al., 2004) have higher return rates. We find no significant gender heterogeneity.[9] In conclusion, our regression results emphasize the robustness of our previous mean behavior analysis.

3.5. Investor Behavior

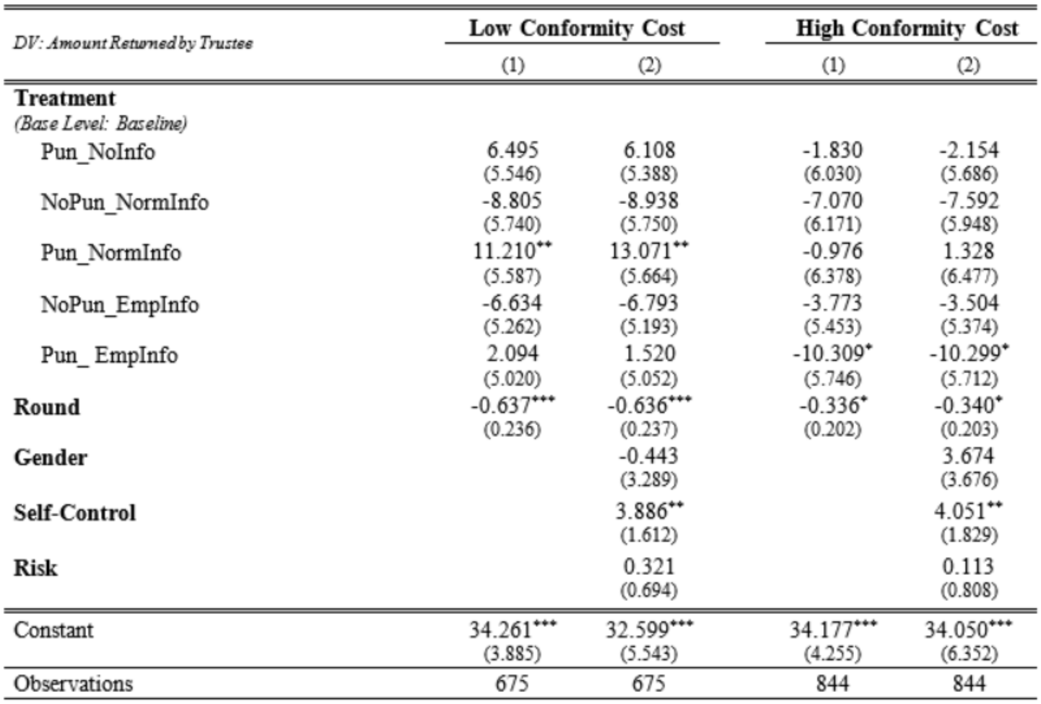
Finally, we investigate investor behavior to understand whether investors anticipated the behavior of trustees in the different conditions. Our main finding is that the investors seem to have anticipated the negative reaction of the trustees towards the Pun_EmpInfo condition when the conformity cost was high. While this is already supported by the dip in investment behavior in the Pun_EmpInfo condition[10] (see Figure A2 in the Appendix), the most interesting insights are generated from examining the behavior conditional on the investor's action space.

When we look at the investment amount (0, 4ECU or 8ECU) we observe substantial heterogeneity of investment behavior and a striking difference of investment behavior in the NoPun_Empinfo and the Pun_EmpInfo conditions compared to the other conditions (Figure 8).[11] When there is empirical information, but no punishment (NoPun_EmpInfo) 57.6% of investors transfer all of their 8ECU; only 26.1% are willing to do so when the

[9] Note that all results are robust to the inclusion of the 7% of data in which investors did not send a return request message, as shown in Table A2 in the Appendix. We provide a more detailed analysis of the drivers of trustee behavior across treatments in Table A3 in the Appendix.

[10] Significantly lower than investments in the NoPun_EmpInfo and Pun_NormInfo condition at the 1% level. More detailed analysis is skipped for brevity, but is available upon request from the authors.

[11] Here, we focus only on investments that were accompanied by a return request message. By design, only this makes sense since the return request message evokes the implementation of punishment where it was available (Pun_NoInfo, Pun_NormInfo, Pun_EmpInfo).



DV: Amount Returned by Trustee	**Low Conformity Cost**		**High Conformity Cost**	
	(1)	(2)	(1)	(2)
Treatment				
(Base Level: Baseline)				
Pun_NoInfo	6.495	6.108	-1.830	-2.154
	(5.546)	(5.388)	(6.030)	(5.686)
NoPun_NormInfo	-8.805	-8.938	-7.070	-7.592
	(5.740)	(5.750)	(6.171)	(5.948)
Pun_NormInfo	11.210**	13.071**	-0.976	1.328
	(5.587)	(5.664)	(6.378)	(6.477)
NoPun_EmpInfo	-6.634	-6.793	-3.773	-3.504
	(5.262)	(5.193)	(5.453)	(5.374)
Pun_ EmpInfo	2.094	1.520	-10.309*	-10.299*
	(5.020)	(5.052)	(5.746)	(5.712)
Round	-0.637***	-0.636***	-0.336*	-0.340*
	(0.236)	(0.237)	(0.202)	(0.203)
Gender		-0.443		3.674
		(3.289)		(3.676)
Self-Control		3.886**		4.051**
		(1.612)		(1.829)
Risk		0.321		0.113
		(0.694)		(0.808)
Constant	34.261***	32.599***	34.177***	34.050***
	(3.885)	(5.543)	(4.255)	(6.352)
Observations	675	675	844	844

Table 2: Note: Random effects model with robust standard errors (in parentheses) clustered on the participant level. 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). Significance levels: $^{*}p$ <0.10, $^{**}p$ <0.05, $^{***}p$ <0.01

empirical information is accompanied by punishment (Pun_EmpInfo). This difference is highly statistically significant (BSM, p<0.01). Information that a majority of trustees returned half in a previous game would presumably make investors more confident about the return of high investment, as confirmed by the highest frequency (57.6%) of sending 8ECU in the NoPun_EmpInfo treatment. Yet when that information is accompanied by punishment, that confidence is shattered. Investors are more cautious and the majority now gives half of their endowment or nothing at all (46.3% and 27.7%, respectively). Consequently, distribution of behavior in Pun_EmpInfo is significantly different from the distribution of behavior in any other treatment (Kolmogorov-Smirnov tests, p<0.01).

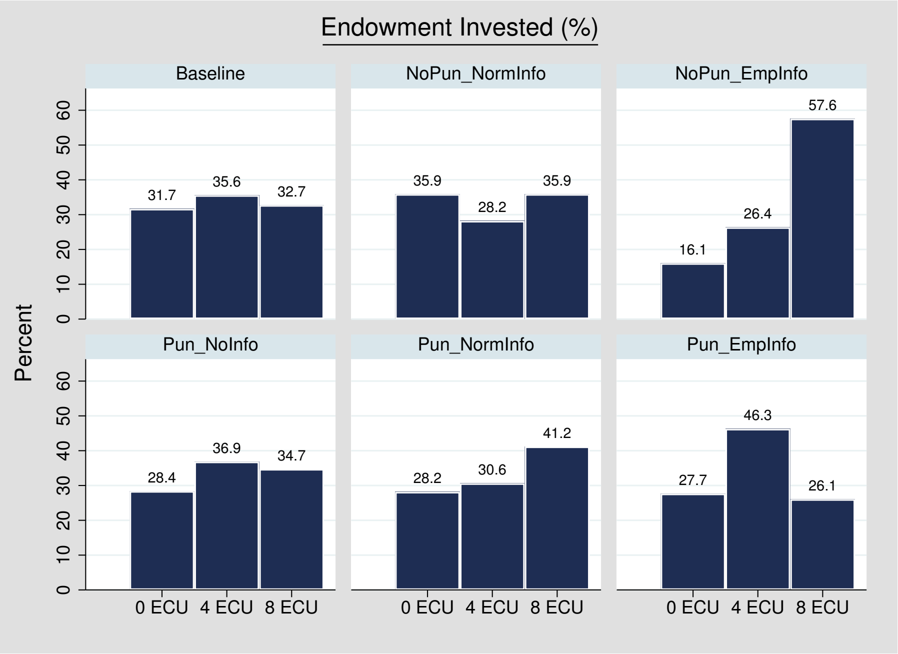


Figure 8: Amount invested by trustors when return request message was sent.

We conclude that the trustor's investment choices are suggestive of their (correct) anticipation that punishment could be detrimental in this context, but only when it is accompanied with empirical information. We have no evidence for a similar line of reasoning when punishment is combined with normative information.

4. Discussion and Conclusion

We find a positive impact of normative information on the efficacy of punishment, subject to the compliance cost. In contrast, providing empirical information does not help regardless of the compliance cost, and its presence can even backfire. We provide further evidence that punishment, empirical information, or normative information alone are insufficient to explain the effect of the combination of punishment and norm information. These results suggest that norm information affects behavior by changing individuals' perception of the legitimacy of the rule enforced by punishment, especially when the information points to a social norm.

While the ineffectiveness of the empirical information is consistent with our hypothesis, as well as, previous studies, its detrimental effect has not been previously considered. One possible explanation suggests that people view punishment as illegitimate when the reason for punishment is perceived as the divergence between theirs and others' behavior, and not the wrongness of the action. The consequent resentment towards punishment is consistent with the shift from complete conformity to complete violation. Yet, this interpretation begs the question: Why does the resentment towards punishment occur only under the condition of high compliance cost?

One possibility is that the message conveyed by the empirical information allows for substantial wiggle-room. Recall that in our experiment the message says the following: "In a previous survey, most participants in the role of Player 2 returned at least half of the tripled transferred amount to Player 1." The message did not specify what the behavior looked like in the *Low Conformity Cost* and *High Conformity Cost* situations separately. Since the reference group for this behavior remained unspecified in the instructions, participants might have had an incentive to form self-serving beliefs, choosing to believe that the high compliance information mostly referred to the low conformity cost condition. Such a self-serving belief may further lead trustees to view punishing deviant behavior under the high conformity cost condition as illegitimate, because it asks trustees to return 50% only when others do so under the low compliance cost condition.

This conjecture is consistent with recent experimental evidence that suggests that when empirical information is ambiguous individuals tend to choose self-serving behavior more often due to easier justifiability (i.e., Konow, 2000; Dana et al., 2007; Spiekermann and Weiss, 2016; Bicchieri and Dimant, 2017). Importantly, such a line of reasoning does not apply to our punishment plus norm information condition, since normative information is much less equivocal. Normative information tells us what the right thing to do is and not what others in fact do. Indeed, it would be hard to argue that the obligation to reciprocate is only valid when the cost of compliance is sufficiently low.

Numerous studies on punishment demonstrate its potential counterproductive effect. We point out that one potential explanation for the failure of punishment is the perception that the enforced rule is illegitimate or unjustifiable. Data from our experiment shows that providing normative information emphasizing that violations are socially disapproved of enhances the efficacy of punishment as long as compliance is not sufficiently costly. This result demonstrates that punishment itself may not be sufficient for rule enforcement. It is important to highlight the social desirability of the enforced behavior.

The finding of the detrimental effect of empirical information is novel and requires further investigation to help identify the underlying causal mechanisms that result in said effect. There has recently been mounting interest in applying social norm methods to nudge behavior (OECD, 2015). As we need to differentiate carefully between empirical and normative information, our results suggest caution. We show people react negatively to punishment against deviations. We speculate that such a negative reaction may be caused by the ambiguity of data provided by the empirical information. It would be valuable to further study whether resolving this ambiguity eliminates the detrimental effect.

5. References

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

Acemoglu, Daron and Jackson, Matthew O. (2016). Social Norms and the Enforcement of Laws. MIT Department of Economics Working Paper No. 14-16.

Becker, Gary S. (1968). Crime and punishment: an economic approach. *Journal of Political Economy, 76*(2), 169-217.

Benabou, Roland, and Jean Tirole (2011). Laws and Norms. NBER Working Paper 15579

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Bicchieri, C., & Dimant, E. (2017). It's not a lie if you believe it. An experimental analysis of lying behavior in ambiguous norm environments. Mimeo.

Bicchieri, C., & Marini, A. (2016) Ending female genital cutting: The role of macro variables (Working Paper). Behavioral Ethics Lab: University of Pennsylvania.

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

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

Cialdini, R. B., Reno, R. R., & 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.

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

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

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

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

Fisman, R. & Miguel, E., 2010. Economic gangsters: Corruption, violence, and the poverty of nations. Princeton University Press

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

Goldstein, N., Cialdini, R., & 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. doi:10.1086/586910

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

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., & Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, 3, 96-101.

Moffatt, P. G. (2015). Experimetrics: Econometrics for experimental economics. Palgrave Macmillan.

Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). *Corrupting cooperation and how anti-corruption strategies may backfire. Nature Human Behaviour, 1*(7), s41562-017.

OECD (2015). Behavioral Insights and New Approaches to Policy Design. The Views from the Field. International Seminar Report.

Posner, Eric (2002) Laws and Social Norms. Cambridge: Harvard University Press

Spiekermann, K., & 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.

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

Tankard, M., & Paluck, E.L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review, 10*(1), 181-211.

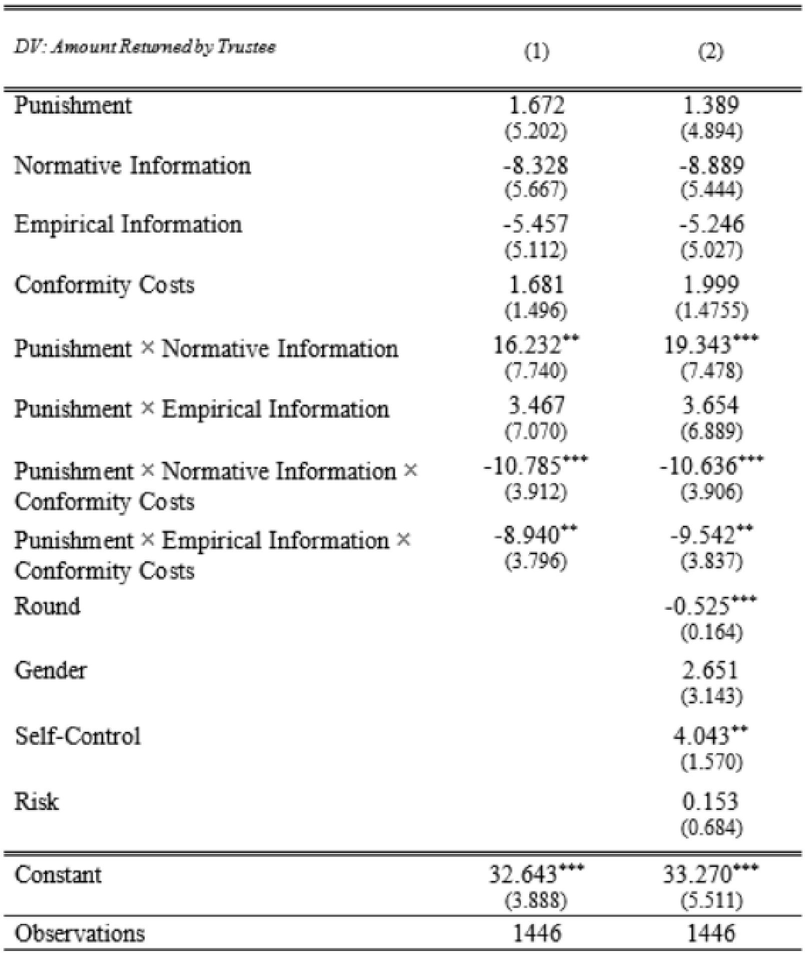
Tyler, T.R. (2006) Why People Obey the Law. Princeton University Press

Tyran, J. R., & Feld, L. P. (2006). Achieving Compliance when Legal Sanctions are Non-Deterrent. *The Scandinavian Journal of Economics, 108*(1), 135-156.

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

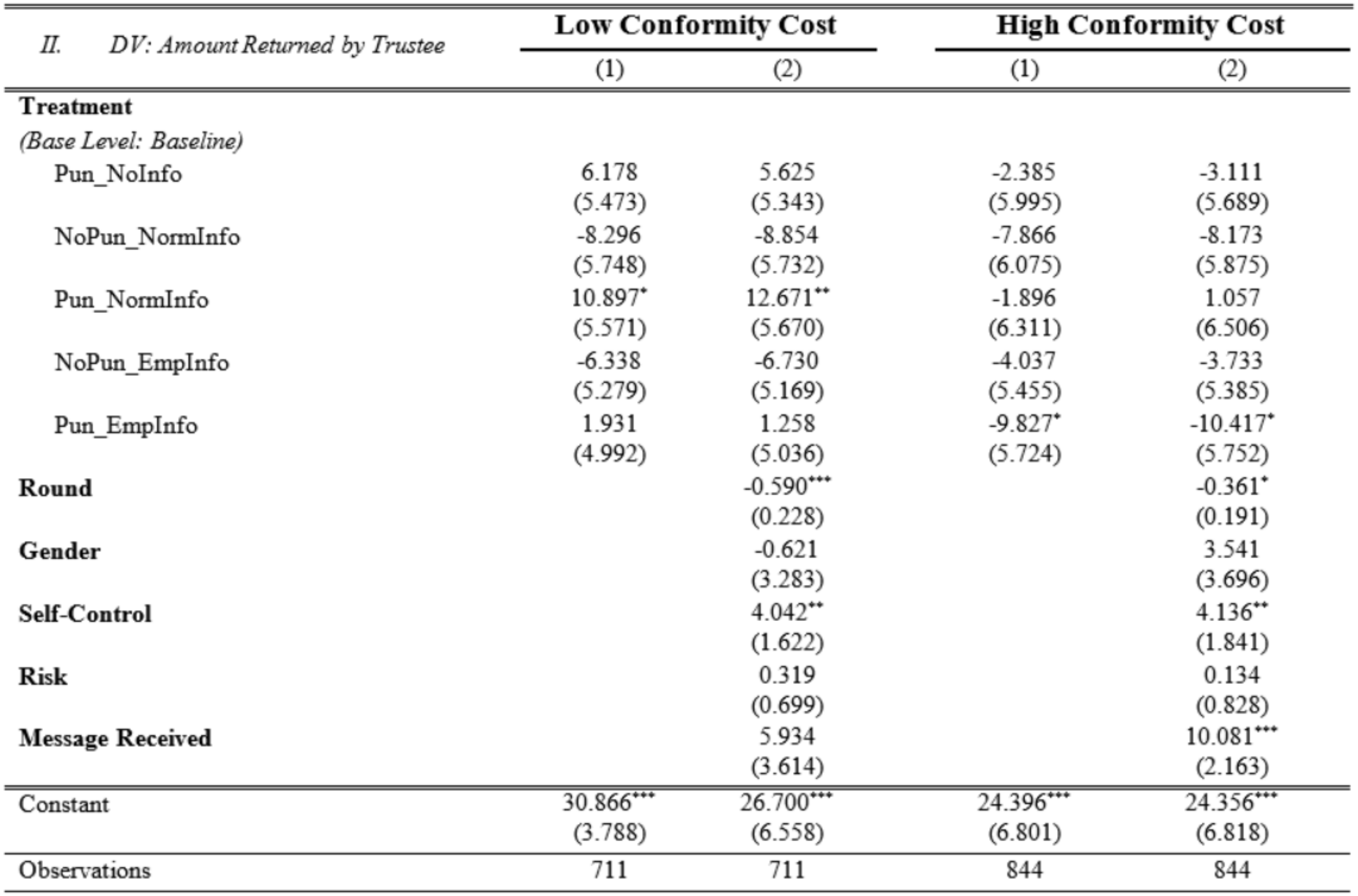
6. Appendix

6.1. Robustness Checks: Additional Tables and Figures



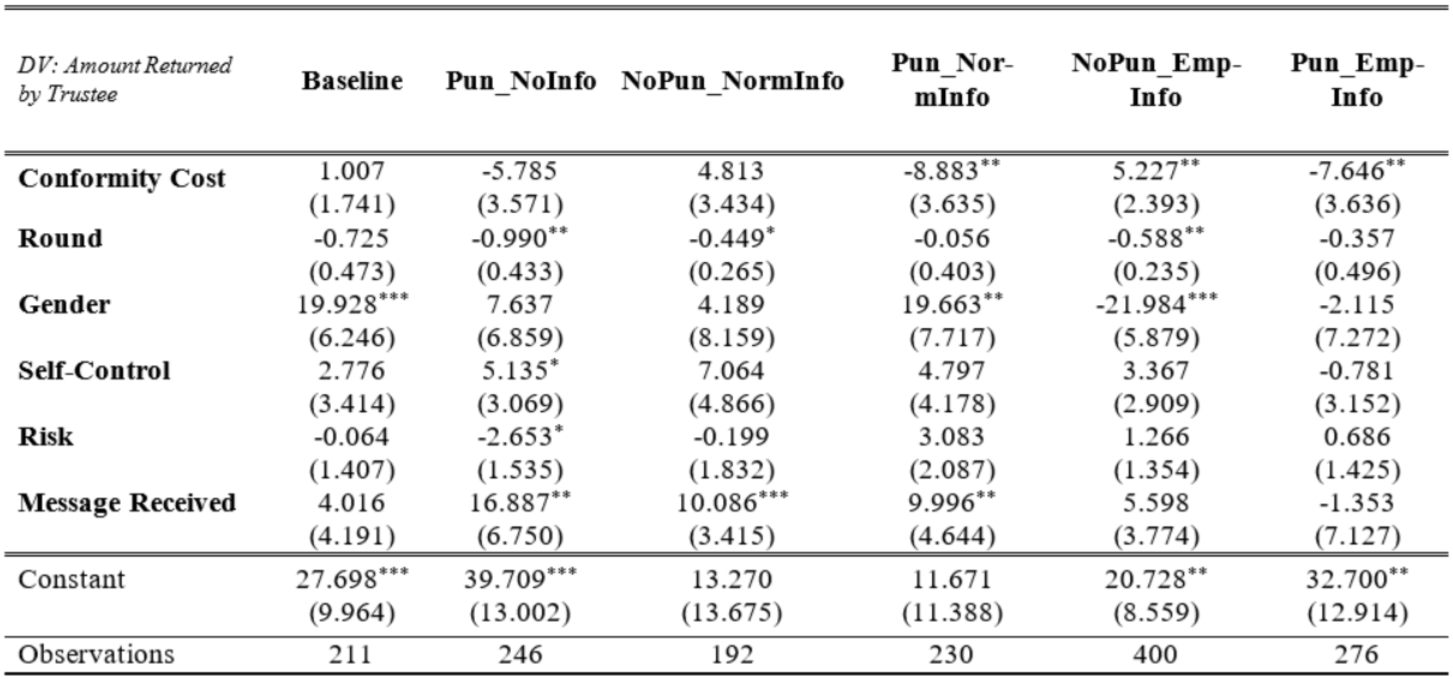
DV: Amount Returned by Trustee	(1)	(2)
Punishment	1.672 (5.202)	1.389 (4.894)
Normative Information	-8.328 (5.667)	-8.889 (5.444)
Empirical Information	-5.457 (5.112)	-5.246 (5.027)
Conformity Costs	1.681 (1.496)	1.999 (1.4755)
Punishment × Normative Information	16.232** (7.740)	19.343*** (7.478)
Punishment × Empirical Information	3.467 (7.070)	3.654 (6.889)
Punishment × Normative Information × Conformity Costs	-10.785*** (3.912)	-10.636*** (3.906)
Punishment × Empirical Information × Conformity Costs	-8.940** (3.796)	-9.542** (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*** (3.888)	33.270*** (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



II. DV: Amount Returned by Trustee	**Low Conformity Cost**		**High Conformity Cost**	
	(1)	(2)	(1)	(2)
Treatment				
(Base Level: Baseline)				
Pun_NoInfo	6.178	5.625	-2.385	-3.111
	(5.473)	(5.343)	(5.995)	(5.689)
NoPun_NormInfo	-8.296	-8.854	-7.866	-8.173
	(5.748)	(5.732)	(6.075)	(5.875)
Pun_NormInfo	10.897*	12.671**	-1.896	1.057
	(5.571)	(5.670)	(6.311)	(6.506)
NoPun_EmpInfo	-6.338	-6.730	-4.037	-3.733
	(5.279)	(5.169)	(5.455)	(5.385)
Pun_EmpInfo	1.931	1.258	-9.827*	-10.417*
	(4.992)	(5.036)	(5.724)	(5.752)
Round		-0.590***		-0.361*
		(0.228)		(0.191)
Gender		-0.621		3.541
		(3.283)		(3.696)
Self-Control		4.042**		4.136**
		(1.622)		(1.841)
Risk		0.319		0.134
		(0.699)		(0.828)
Message Received		5.934		10.081***
		(3.614)		(2.163)
Constant	30.866***	26.700***	24.396***	24.356***
	(3.788)	(6.558)	(6.801)	(6.818)
Observations	711	711	844	844

Table A2: Note: Random effects model with robust standard errors (in parentheses) clustered on the participant level. Coding of control variables the same as in Table 2. Significance levels: * p <0.10, ** p <0.05, *** p <0.01



DV: Amount Returned by Trustee	**Baseline**	**Pun_NoInfo**	**NoPun_NormInfo**	**Pun_Nor-mInfo**	**NoPun_Emp-Info**	**Pun_Emp-Info**
Conformity Cost	1.007	-5.785	4.813	-8.883**	5.227**	-7.646**
	(1.741)	(3.571)	(3.434)	(3.635)	(2.393)	(3.636)
Round	-0.725	-0.990**	-0.449*	-0.056	-0.588**	-0.357
	(0.473)	(0.433)	(0.265)	(0.403)	(0.235)	(0.496)
Gender	19.928***	7.637	4.189	19.663**	-21.984***	-2.115
	(6.246)	(6.859)	(8.159)	(7.717)	(5.879)	(7.272)
Self-Control	2.776	5.135*	7.064	4.797	3.367	-0.781
	(3.414)	(3.069)	(4.866)	(4.178)	(2.909)	(3.152)
Risk	-0.064	-2.653*	-0.199	3.083	1.266	0.686
	(1.407)	(1.535)	(1.832)	(2.087)	(1.354)	(1.425)
Message Received	4.016	16.887**	10.086***	9.996**	5.598	-1.353
	(4.191)	(6.750)	(3.415)	(4.644)	(3.774)	(7.127)
Constant	27.698***	39.709***	13.270	11.671	20.728**	32.700**
	(9.964)	(13.002)	(13.675)	(11.388)	(8.559)	(12.914)
Observations	211	246	192	230	400	276

Table A3: Note: Random effects model with robust standard errors (in parentheses) clustered on the participant level. Coding of control variables the same as in Table 2. Significance levels: * p <0.10, ** p <0.05, *** p <0.01

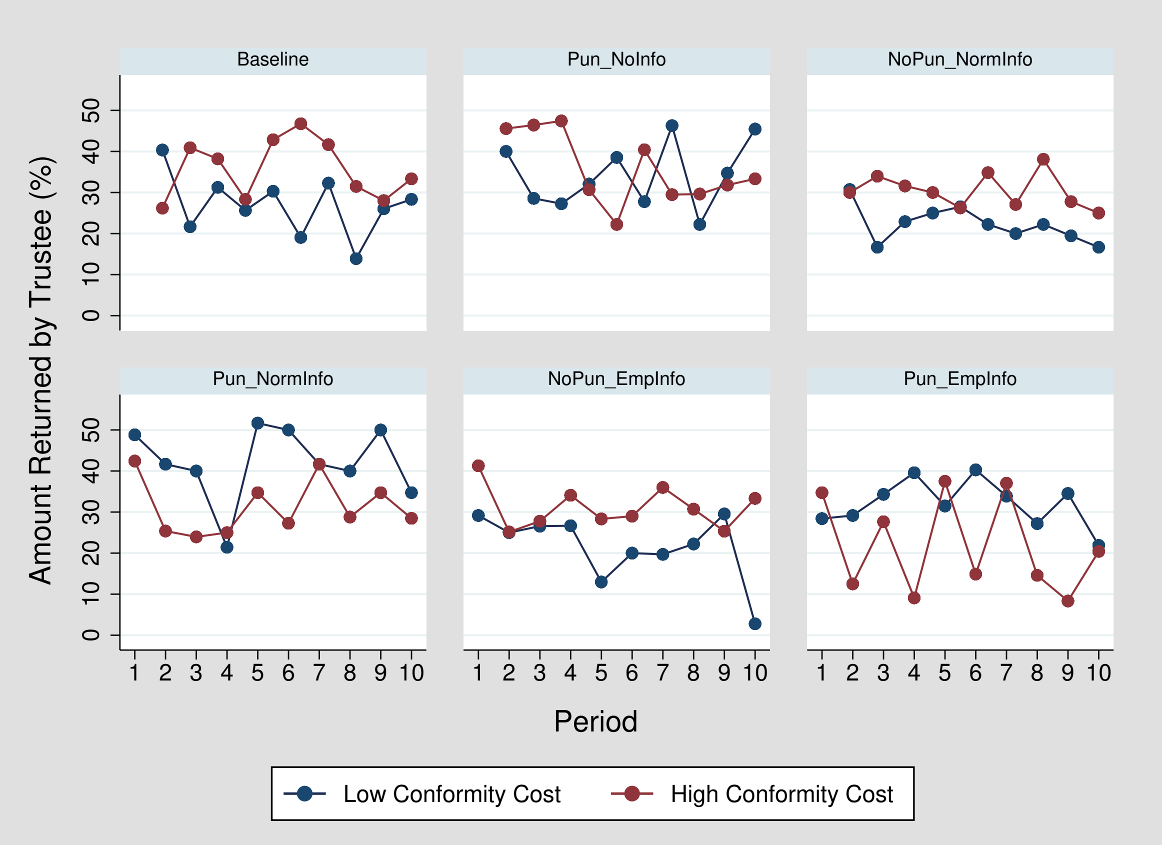


Figure A1: Amount returned by trustees as percentage of amount received from the investor over rounds.

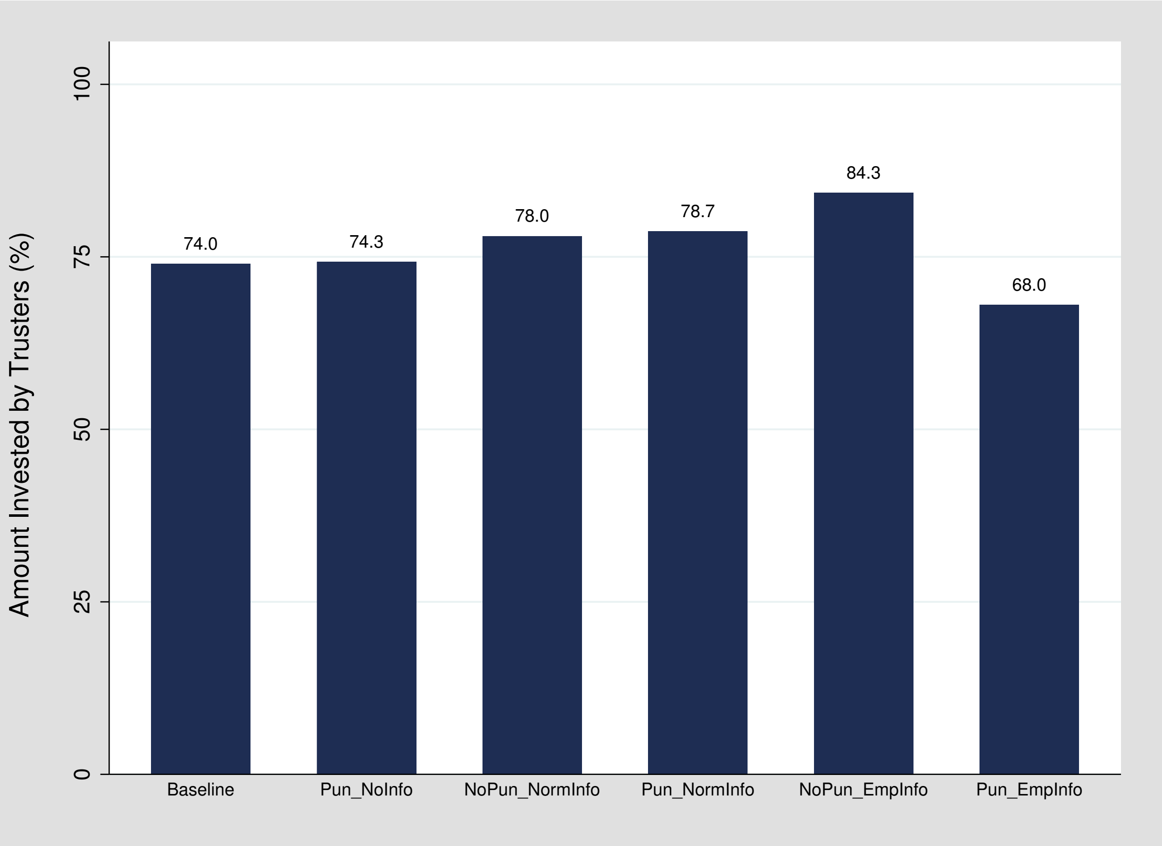


Figure A2: Amount invested by trustors when a return request message was sent (93% of the cases).

6.2. Experimental Instructions

Subsequently, we present the instructions exemplarily for Treatment 5 (Punishment + Empirical Information). Differences with our other treatments are highlighted in the text. More specifically, the part highlighted yellow was presented only in this treatment and in Treatment 4 (No Punishment + Empirical Information) to the participants. In Treatment 2 (No Punishment + Normative Information) and Treatment 3 (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. Therefore, it was presented (absent) only in treatments 1, 3, and 5 (Baseline, 2, and 4).

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 •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.

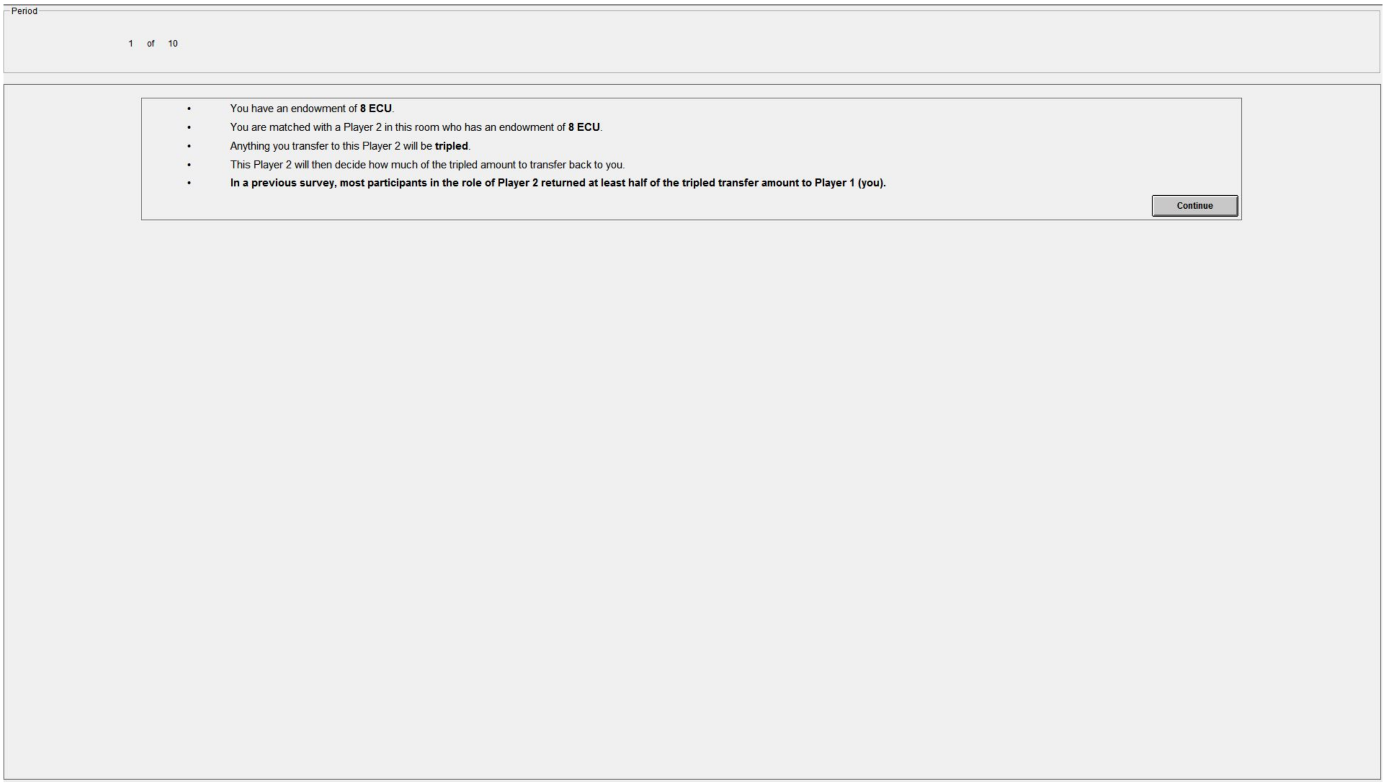
Player 1 will not know the result of each round (i.e. Player 1 will not know Player 2's decision in each round) until all the 10 rounds have finished. After all the 10 rounds have finished, each Player 1 will learn the matched Player 2's decision and the payoff outcomes in each round. Each Player 2 will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

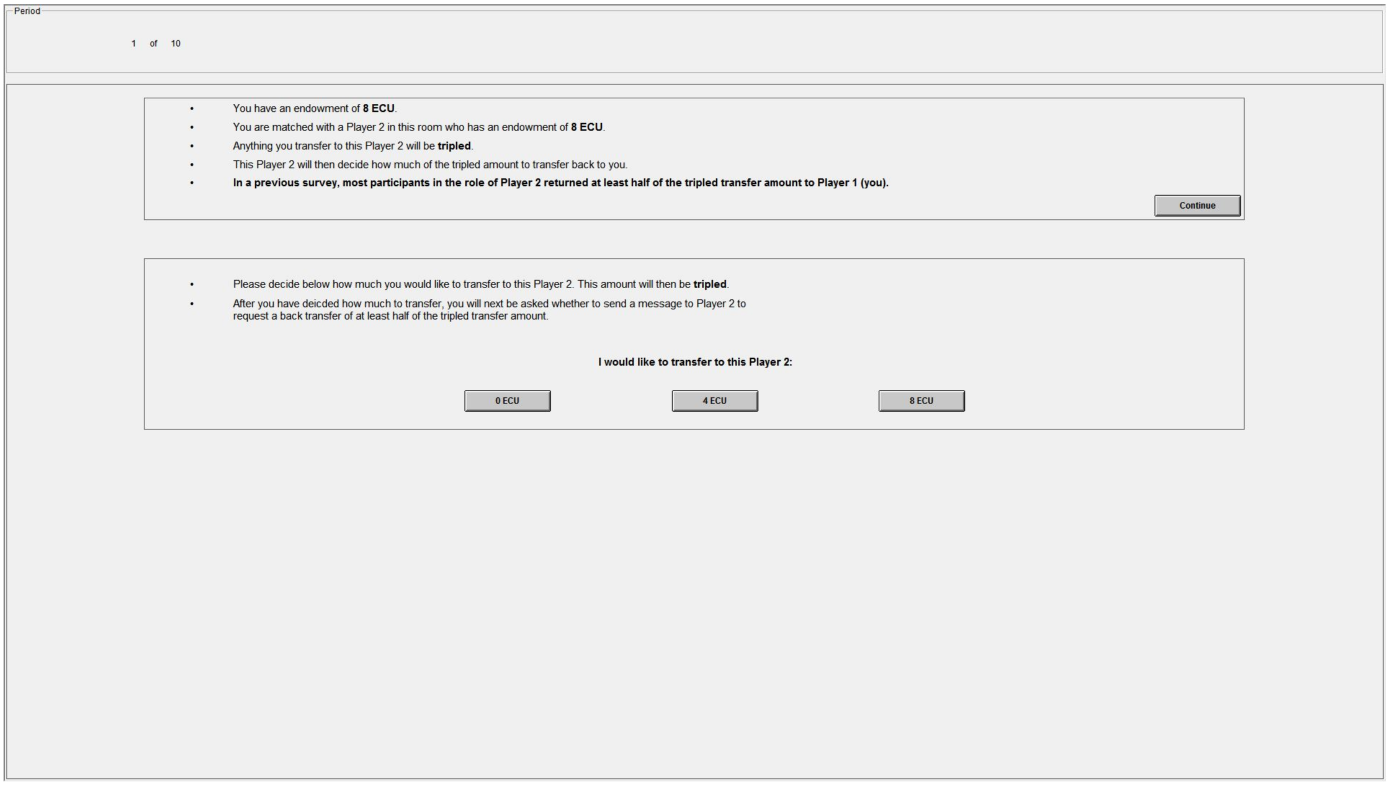
6.3. Screenshots of Experimental Procedure

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.

<u>Investor</u>



- 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).**



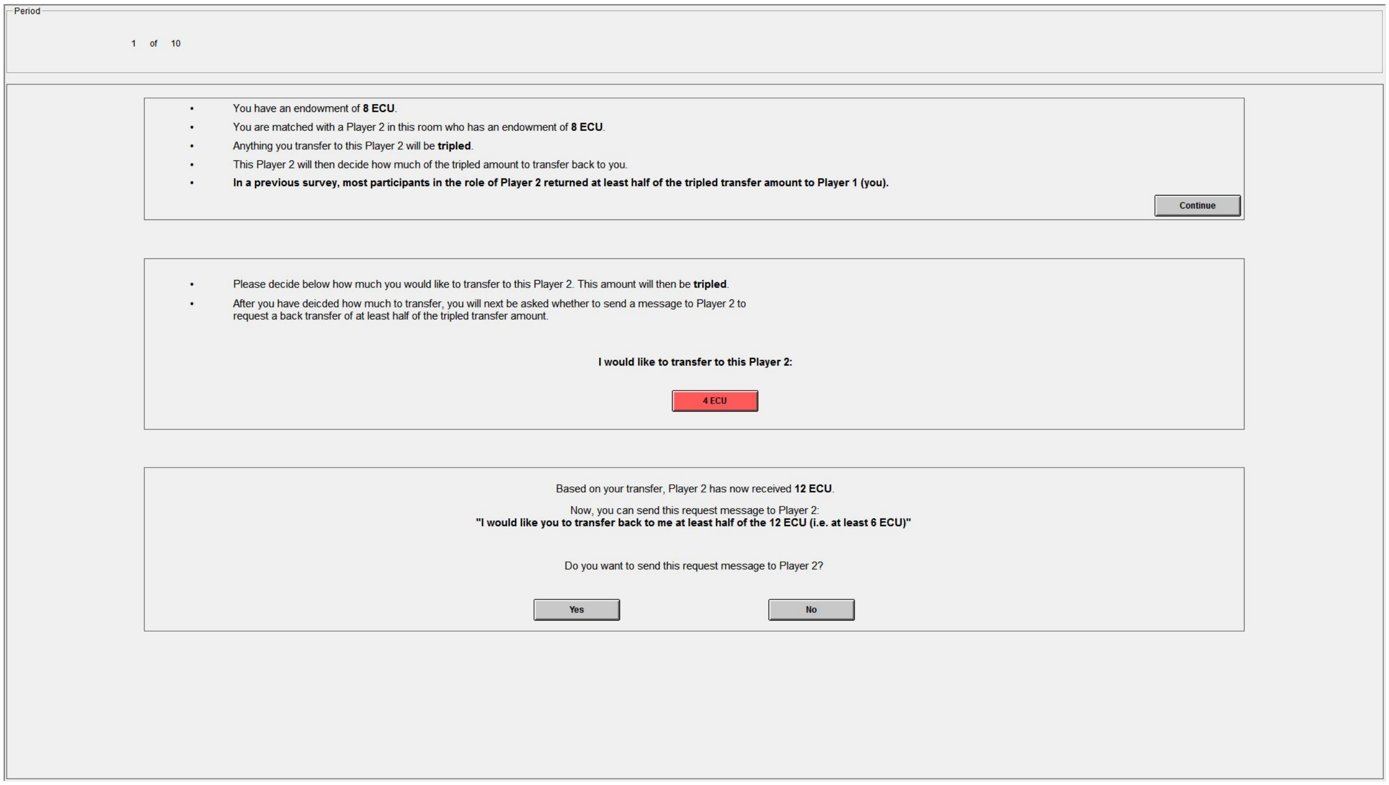
Period

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:



- 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).**

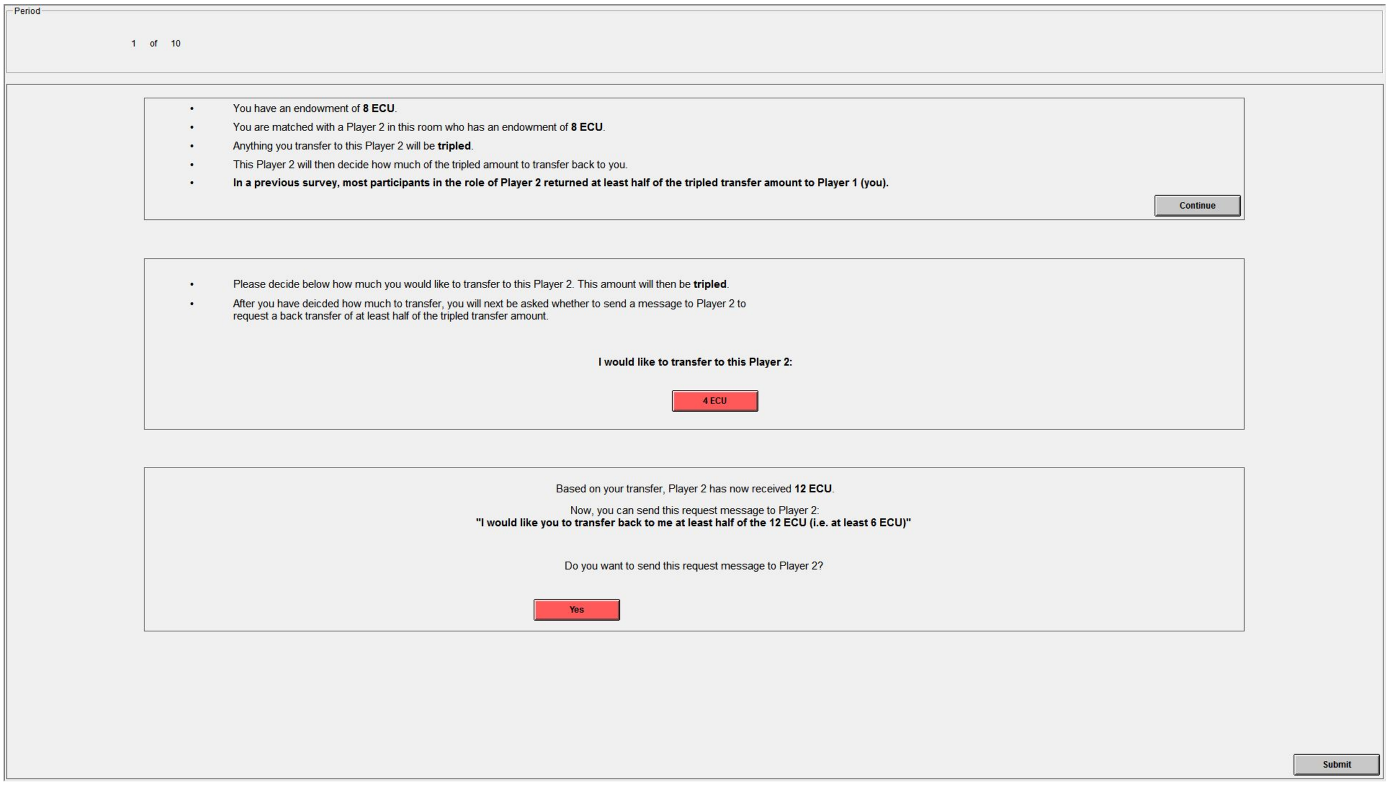
- Please decide below how much you would like to transfer to this Player 2. This amount will then be **tripled**.
- After you have deicded how much to transfer, you will next be asked whether to send a message to Player 2 to request a back transfer of at least half of the tripled transfer amount.

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?



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.

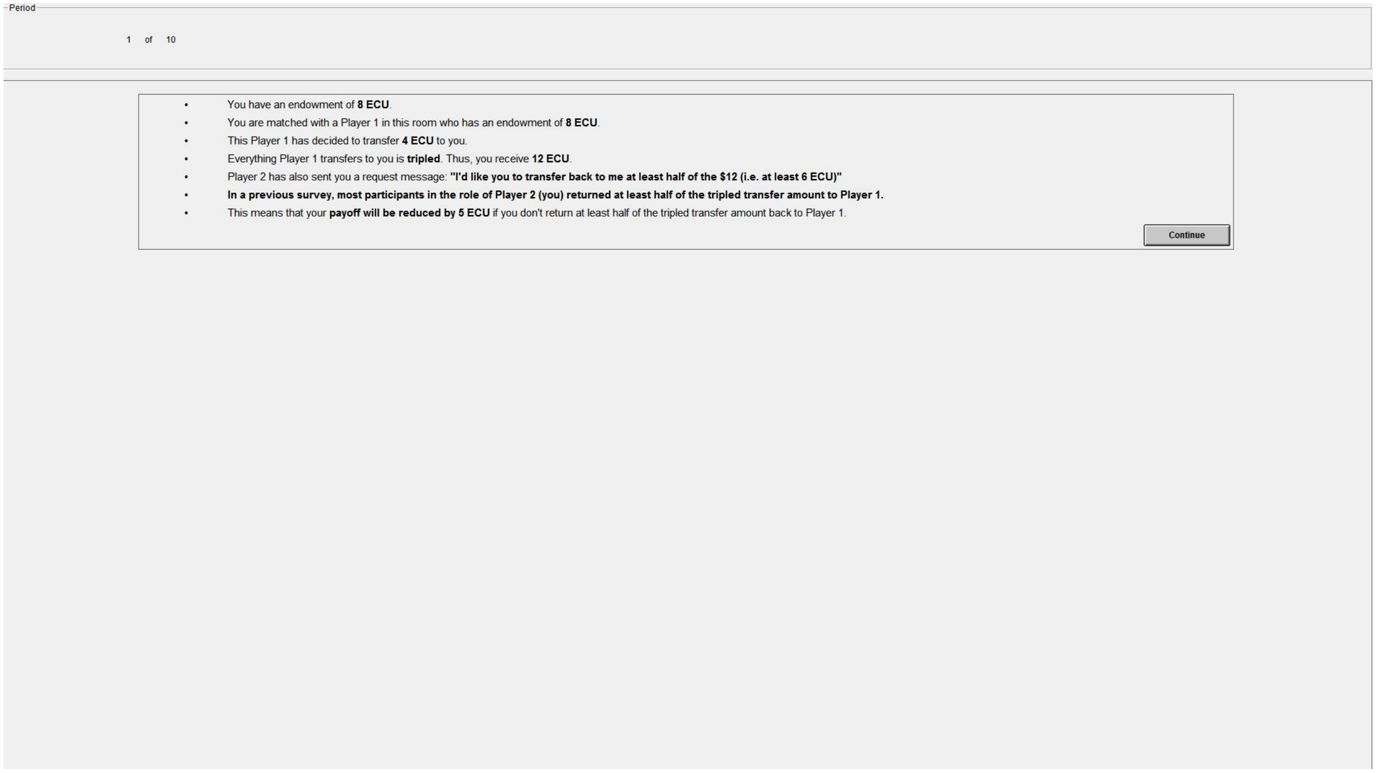
I would like to transfer to this Player 2:

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

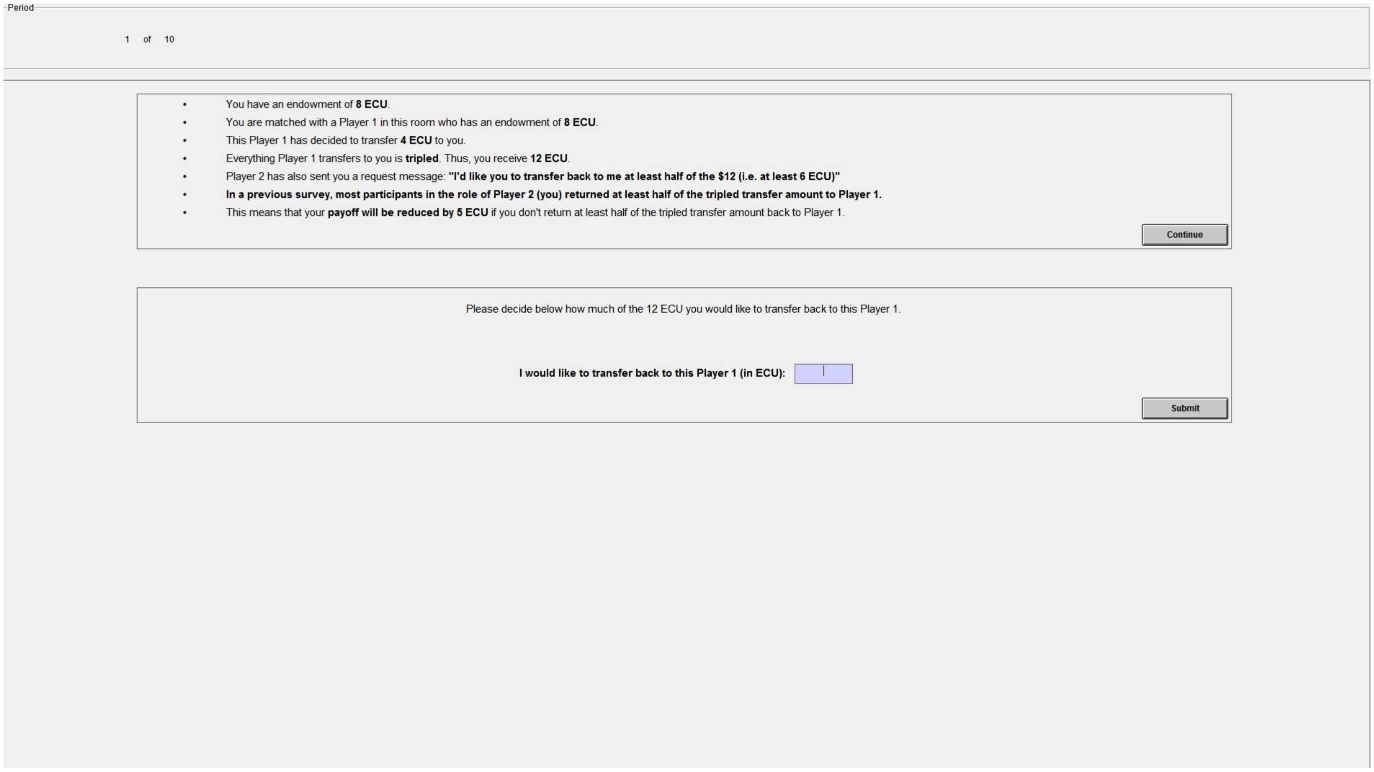
Now, you can send this request message to Player 2:
"I would like you to transfer back to me at least half of the 12 ECU (i.e. at least 6 ECU)"

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

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.



Period

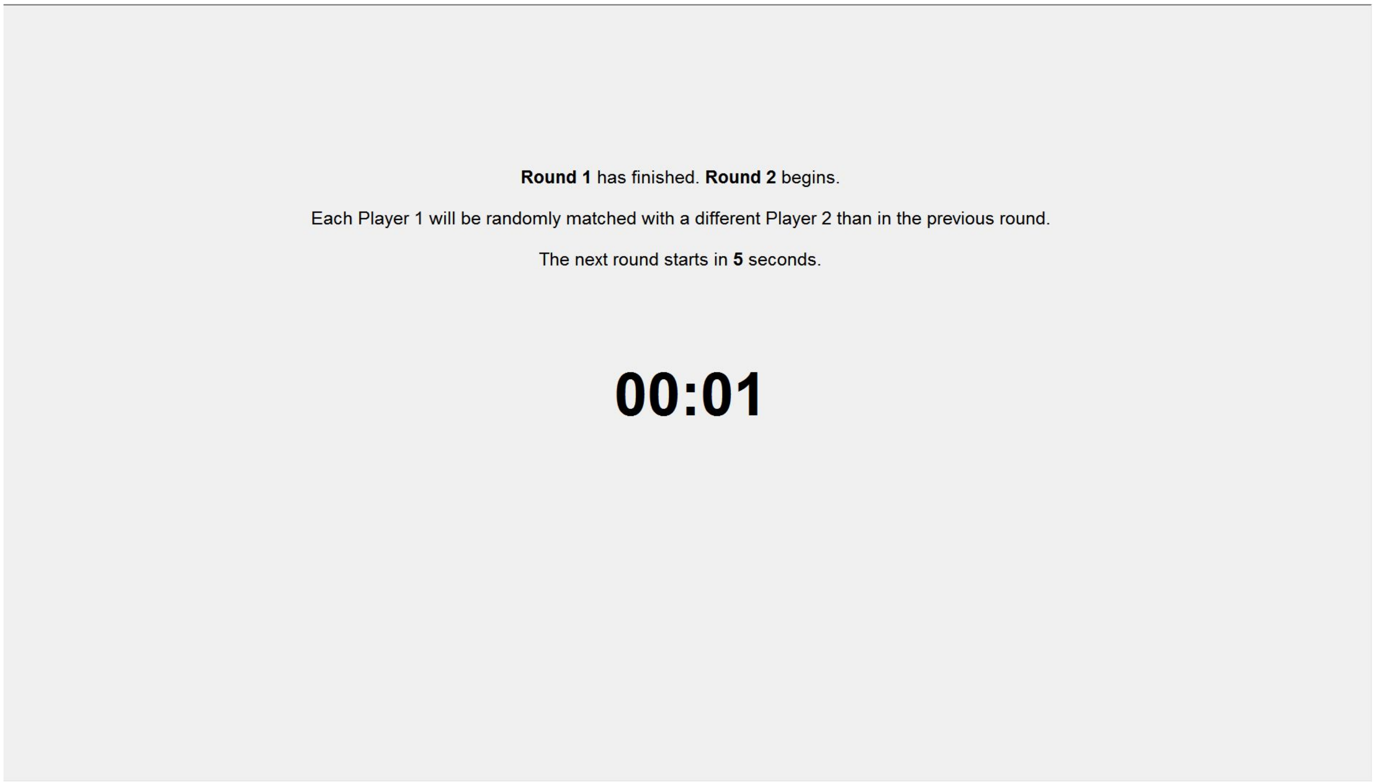
1 of 10

- 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.

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