Carry a big stick, or no stick at all

Punishment and endowment heterogeneity in the trust game

Vicente Calabuig
ERICES, Universidad de Valencia

Enrique Fatas*
University of East Anglia

Gonzalo Olcina
ERICES, Universidad de Valencia

Ismael Rodriguez-Lara
Middlesex University London

Abstract

We investigate the effect of costly punishment in a trust game with endowment heterogeneity. Our findings indicate that the difference between the investor and the allocator's initial endowments determines the effect of punishment on trust and trustworthiness. Punishment fosters trust only when the investor is wealthier than the allocator. Otherwise, punishment fails to promote trusting behavior. As for trustworthiness, the effect is just the opposite. The higher the difference between the investor and the allocator's initial endowments, the less willing allocators are to pay back. We discuss the consistency of our findings with social preference models (like inequality aversion, reciprocity), the capacity of punishment (i.e., the deterrence hypothesis) and hidden costs of punishment (i.e., models of intrinsic and extrinsic motivation). Our results are hardly coherent with the first two (inequality aversion and deterrence), but roughly consistent with the latter.

Keywords: Trust game, endowment heterogeneity, punishment, deterrence hypothesis, crowding-out, intrinsic and extrinsic motivation, experimental economics.

JEL Codes: C91, D02, D03, D69.

Acknowledgement

We thank the comments provided by Timothy Salmon on a previous version of the manuscript. The paper has also benefited from suggestions provided by conference and seminar participants at LUISS Guido Carli (Rome, Italy), Chapman University, Workshop on Trust and Cultural Evolution (Valencia), IMEBE 2012 (Castellón), ESA Meeting New York, III SEET Meeting (Lanzarote), Simposio Analisis Económico (Madrid). Financial support from the Spanish Ministry of Education under the project ECO2014-58297-R is gratefully acknowledged. Enrique Fatas acknowledges the support of the ESRC Network for Integrated Behavioral Science (NIBS).

* Corresponding author. Prof. Enrique Fatas. School of Economics. University of East Anglia. Norwich NR4 7TJ. United Kingdom. Webpage: http://www.uea.ac.uk/~amr11jwu Telephone number: +44 (0) 1603 593415. Email: E.Fatas@uea.ac.uk,

1. Introduction

We study the different effects of punishment on trust and trustworthiness in the presence of endowment heterogeneity. Trust and trustworthiness determine economic performance at the macro level and grease the wheels of governments (Arrow, 1974; Knack and Keefer, 1997; Bachmann and Zaheer, 2006; Gambetta, 1988). At the micro level, they both play an essential role in a variety of strategic environments such as incomplete contracts, the hold-up problem or the principal-agent relationship, which may include sequential investment decisions (e.g., the purchase of a good in which the seller chooses the quality of the product after the buyer bought it).

Starting with Berg et al (1995), trust has been studied as a stylized two party sequential game where a trustor (the investor) unconditionally invests in the first stage of the game. Any amount invested is multiplied before a trustee (the allocator) unilaterally decides how much to give back. Because the sub-game perfect equilibrium under the assumption of purely self-interested subjects is that allocators will return nothing to investors, the investor's decision has been usually identified in the literature as the level of trust, whereas the level of trustworthiness (or reciprocity) has been frequently measured by the allocator's payback.

The determinants of trust and trustworthiness have been extensively studied in the laboratory (see Kagel and Cooper 2009, Johnson and Mislin 2011 and Eckel and Wilson 2011 for recent surveys), albeit most studies do not explicitly investigate how to foster them.[1] Recent theoretical work from anthropology and evolutionary game theory highlights the positive effects of individually costly punishment on cooperation (Boyd et al., 2003; Hauert et al. 2007; Fehr and Fischbacher, 2003; Olcina and Calabuig, 2008). Laboratory experiments do also suggest that sanctions significantly increase contribution to public goods games (Chaudhuri, 2011), even between strangers (Fehr and Gächter, 2002).

Although the potential benefits of sanctions on trust-governed interactions like the ones described above seems straightforward (e.g., by using sticks (punishment) principals might increase the cooperation of their employees, and buyers may get higher quality products from sellers, trying to elude sanctions), the literature on the effects of punishment on trust and trustworthiness is still scarce (see Fehr and Rockenbach, 2003; Fehr and List, 2004; Houser et al. 2008; Rigdon, 2009). Rigdon (2009) is the first paper in which punishment comes at a cost for investors. Punishment significantly increases trust only when it is very effective or relatively cheap for investors; i.e., when investors spending one token results in large losses for allocators.[2] In settings where punishment is costless for investors, and punishment is weak in that the cost of receiving punishment is smaller than the benefit of violation, the experimental evidence

[1] McCabe et al. (2003), Cox (2004), Cox et al. (2014a) or Ashraf et al. (2006) specifically attempt to analyze the different motivations behind trust and reciprocity, including the role of intentions and the importance of altruism and expectations on behavior.

[2] The low-punishment treatment in Rigdon (2009) is such that investors need to spend 1 unit of their endowment to reduce the allocator's payoff in 1 unit, whereas the allocator's payoff is reduced in 3 units in the high-punishment treatment.

suggest that punishment crowds out trustworthiness, as allocators might be less likely to reciprocate when they get threats from investors (see Fehr and Rockenbach, 2003; Houser et al. 2008; Fehr and List, 2004).[3]

Experimental evidence on the effect of endowment heterogeneity on punishment effectiveness is much more scarce. Endowment heterogeneity is a crucial and common feature in most real life settings, like principal-agent relationships or bargaining situations with a monopolistic seller and competitive buyers. One simple example of how endowment heterogeneity may affect the efficacy of costly punishment is to consider the strategic dilemma faced by someone building a new house, and finding the outcome unsatisfactory. In the absence of a budget constraint, hiring a good lawyer to sue the builder is both an available option, and a credible threat. If resources are scarce, the balance between your resources and the builder's company size will most likely determine the extent to which you can effectively punish the builder. Endowment heterogeneity (rather the mere existence of sanctions, and its technology) thus becomes a relevant determinant of the damage you can credibly inflict to the builder.

In this paper, we investigate the effect of endowment heterogeneity on trust and trustworthiness in a controlled laboratory experiment. We give the investor the possibility of punishing the allocator at a cost, and systematically manipulate the distribution of initial endowments. When sanctions are available, trust could increase because the investor may anticipate the allocator will fear sanctions. Arguably, the allocator will be more or less afraid depending on the magnitude of the investor's endowment, relative to the allocator's one. If the investor is much wealthier than the allocator, we say that the investor's capacity of punishment is high because the investor can destroy most of the allocator's payoff investing a relatively small share of her own payoff. Alternatively, the allocator will fear sanctions relatively less if she is wealthier than the investor, and the investor's capacity of punishment of the investor is low.[4] As a result, the effect of sanctions on trust and trustworthiness could very reasonably be mediated by differences in initial endowments. Interestingly enough, endowment heterogeneity also changes other features of the trust game by exogenously introducing an unequal distribution of resources. Inequality aversion may generate strong behavioral reactions to non-reciprocal behavior in some endowment distributions, critically affecting the likelihood of punishment.

We test four different predictions in our experiment. The first one is related to the existence of *social preferences*. Both inequality aversion and reciprocity predict an effect of endowment heterogeneity in the trust game, as we will explain below (Coleman 1990, Xiao and Bicchieri 2010, Smith 2011, Rodriguez-Lara 2015). Secondly, the *deterrence hypothesis* assumes that punishment increases both trust and trustworthiness,

[3] In these setups, investors that are allowed to punish choose the amount they send to the allocator and a desired payback. If the desired payback is very large, this might be perceived as unfair by allocators, and therefore less money is returned. When investors refrain to fine, allocators are more likely to return a higher amount. Thus, the allocators' willingness to fulfill the investors' desired back-transfer depends negatively on the investors' requested amount, as well as on the investors' decision on the fine. Allocators also look at the investor's intention and whether or not the cost of receiving punishment is smaller than the benefit of violation (Houser et al. 2008). We decided not to include a desired-back transfer in our setting with punishment because contracts, even if they are not binding, can foster trust (Abbink et al. 2000).

[4] At the same time, the impact of sanctioning institutions on trust is capped by the necessity of keeping enough resources to make the threat of sanctions credible. When the investor is wealthier than the allocator, the investor can invest relatively more and still keep resources to make the punishment credible. Note that in a linear public good game a cooperator always has the chance of punishing free riders with the returns of her investment in the public good. For the study of punishment in other recent bilateral settings, where punishment is less unequivocal than in the public good game see Abbink et al. (2000), Bolle et al. (2011) or Nikiforakis et al (2013).

with the effect of punishment being non-decreasing with the difference between the investor and the allocator's endowments (i.e., trust and trustworthiness should increase monotonically with the investor's capacity of punishment, as defined above).[5] Our third hypothesis relies on *the hidden costs of punishment*. Fehr and Rockenbach (2003), Fehr and List (2004) and Houser et al. (2008) experimentally show that sanctions may crowd out voluntary reciprocation in the trust game; in other words, punishment may backfire if incentives are not well designed, generating a detrimental effect on trust and trustworthiness.[6] We cautiously follow Gneezy et al. (2011) or Gneezy and Rustichini (2000b) and test a non-linear relationship: a negative effect of punishment if the capacity of punishment is low, and a positive effect on trust and trustworthiness when the investor's capacity of punishment is high.[7] Finally, we use our within-subject design to test *the lasting effect of incentives*, which assumes that eliminating punishment will have a negative effect on trust and trustworthiness, especially when the investor is wealthier than the allocator.

Overall, our findings are hardly coherent with theories of inequality aversion or the deterrence hypothesis, but are roughly consistent with an explanation based on the theory of the hidden cost of punishment (i.e., based on the interaction of intrinsic and extrinsic motivations). In summary, we observe that endowment heterogeneity per se has no effect on the levels of trust and trustworthiness without punishment, even when investors send a smaller proportion of their initial endowment when they receive a large one. When punishment is available, it generates a significant effect on trust only when the investor is wealthier than the allocator. In line with the existence of an interaction between intrinsic and extrinsic motivations, a high capacity of punishment (an endowment favorable to the investor) prevents any crowding-out effect when punishment is available. When it is no longer an option, trust is significantly reduced only when the capacity of punishment is high, in line with our *lasting effect of incentives* hypothesis. Contrary to our predictions, punishment actually increases trustworthiness as a rule with one exception: when the investor's endowment is high. We conclude that endowment heterogeneity may have very different effects on investors and allocators, as their intrinsic and extrinsic motivation may be affected differently. In line with the results schemed in this summary, punishment generates efficiency losses except when the investor's endowment is high.

We contribute to the current literature by studying the different effects of punishment on trust and trustworthiness when endowment is heterogeneous. This is the first paper, to our knowledge, that investigates how the efficacy of punishment is affected by endowment heterogeneity in the trust game.[8]

[5] Although the deterrence hypothesis is mainly used to investigate if crime decreases in deterrent incentives (i.e., the severity and the probability of the punishment), it is presented as a behavioral hypothesis in Gneezy and Rustichini (2000b). In our view, this hypothesis is somehow related to models of social preferences (e.g. Rabin 1993, Fehr and Schmidt 1999, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006), which assume that some subjects in the population are selfish, whereas others have social (or other-regarding) preferences. The latter might display reciprocal behavior, being conditional punishers: they are willing to punish opportunistic behavior from their partners when the cost of punishing is not too high. When punishment is more effective (see Nikiforakis and Normann 2008, Rigdon 2009) the impact of punishers on aggregate behavior is higher, and cooperation in the population increases.

[6] This crowding out effect of incentives on intrinsic motivation has been documented in field and laboratory experiments in different settings (see, among others, Deci et al., 1999; Frey, 1997; Gneezy and Rustichini 2000a, 2000b; Gneezy et al., 2011).

[7] Gneezy et al. (2011) and Gneezy and Rustichini (2000b) argue that incentives may be detrimental only if they are not large enough (e.g., when a fine is imposed to late-coming parents, Gneezy and Rustichini (2000b) find that parents tend to arrive even later, probably because the fine was perceived to be "too low").

[8] See, for example, Korenok et al. (2012) or Chowdhury and Jeon (2014) for evidence in the dictator game, Armantier (2006) for the ultimatum bargaining and Cherry et al. (2005) or Reuben and Riedl (2013) for the public good game.

We also investigate how endowment heterogeneity affects investors' choices when punishment is not allowed. We extend the analysis of Coleman (1990), Xiao and Bicchieri (2010), Smith (2011) and Rodriguez-Lara (2015), as they focus on the role of social preferences in allocators' behavior without punishment. Our paper also contributes to the debate on "pay-enough-or-do-not-pay-at-all", discussed in Gneezy and Rustichini (2000a) and Gneezy et al. (2011) measuring the effect of punishment on intrinsic and extrinsic motivation in a bilateral game where endowment heterogeneity determines the capacity of punishment. Finally, our within-subject design is such that subjects play with and without punishment in each of the sessions. Since we control for the order in which treatments are implemented, we analyze how subjects react when incentives are introduced, and removed.

The rest of the paper is organized as follows. In Section 2, we describe our trust game with punishment and provide a non-formal definition of the capacity of punishment, which relies on the endowment heterogeneity. We discuss the experimental design in Section 3. In Section 4 we discuss the behavioral predictions and present the experimental hypotheses. The results are summarized in Section 5. Section 6 presents final remarks and relates our findings with previous research in other environments, where the effects of incentives on individual behavior are analyzed.

2. The Trust Game with Punishment

Consider the bilateral trust game (Berg et al., 1995) in which the investor (subject *a*) decides how much of her endowment $e_a \geq 0$ to send to the allocator (subject *b*), who is initially endowed with $e_b \geq 0$. The experimenter triples any amount X in $[0, e_a]$ that the investor sends to the allocator, so that the allocator receives 3X. He can then return to the investor any amount Y in [0,3X]. The subjects' payoffs are obtained as follows:

(1) $\pi^a(e_a,X,Y) := e_a - X + Y \geq 0$

(2) $\pi^b(e_b,X,Y) := e_b + 3X - Y \geq 0$

We extend the game above and introduce a punishment-phase in which we allow the investor, after observing the *interim* payoffs π^a and π^b, to destroy part of the allocator's payoff at a cost $1{:}\lambda$. This means that for every P monetary units that the investor loses from π^a, the allocator loses λP monetary units from π^b. So λ can be interpreted as the effectiveness of punishment (i.e., λ is the factor by which punishment reduces the allocator's payoff). This modification yields the following payoffs:

(3) $\overline{\pi}^a(e_a,X,Y,P) := e_a - X + Y - P = \pi^a - P \geq 0$

(4) $\overline{\pi}^b(e_b,X,Y,P) := e_b + 3X - Y - \lambda P = \pi^b - \lambda P \geq 0$

Note that by assuming that payoffs cannot be negative, $\overline{\pi}^i \geq 0$ for $i \in \{a,b\}$ we impose a constraint on the investor's punishing behavior. If the investor wanted to destroy the allocator's interim payoff completely, she would need to spend an amount $P' \geq 0$ such that:

(5) $\lambda P' = e_b + 3X - Y$

But given the effectiveness of the punishment (λ) and the *interim* payoffs (π^a, π^b), the maximum punishment that the investor can inflict is given by $P^* = \min\{\pi^a, \pi^b/\lambda\}$. Thus, the share of the allocator's interim payoffs π^b that the investor can destroy with her own interim payoff π^a is given by:

(6) $\frac{\lambda \cdot \pi^a}{\pi^b} = \frac{\lambda \cdot (e_a - X + Y)}{(e_b + 3X - Y)}$

which is larger than 1 if the investor can destroy the allocator's payoffs completely (i.e., $P' \geq P^*$).

Nikiforakis and Normann (2008) and Rigdon (2009) studied the central role played by punishment effectiveness in achieving more efficient outcomes. In our setting this translates into fixing the value of the initial endowments (e_a and e_b) to study the effects of $\lambda > 0$ on subjects' behavior. In our experimental design explained below, we shall keep the effectiveness of punishment constant to focus on the effects of endowment heterogeneity.

One aspect that worth noting from equation (6) is the relationship between the different endowments and the maximum share that the investor can destroy. When the investor is wealthier than the allocator, the investor can destroy the allocator's payoff with a small share of her own payoff; i.e., punishing the allocator is relatively cheap. We then say that the investor has a high capacity of punishment when the investor is wealthier than the allocator. In sharp contrast, it is relatively expensive for the investor to punish the allocator when the allocator is wealthier than the investor. We hereafter say that the capacity of punishment of the investor is low when the allocator is wealthier than the investor.[9]

3. Experimental Design and Procedures

Four experimental sessions were run at the Laboratory for Research in Experimental Economics (LINEEX), University of Valencia. We recruited a total of 96 subjects (24 per session) using the electronic recruitment system of the laboratory. All subjects were business and economics undergraduate students with no experience in similar experiments. The experiment was conducted using the z-Tree software (Fischbacher, 2007).

At the beginning of each session, we randomly assigned subjects a fixed role (namely, subject *a* or subject *b*). Each subject went then through a sequence of eight one-shot games with two different treatments: No Punishment (NOPUN) and Punishment (PUN). Each treatment consisted of four rounds in which subjects with different roles were matched in pairs. Subjects interacted with each other using a perfect-stranger protocol within treatments, as they never made more than one decision with the same

[9] The interested reader in a more formal definition of the capacity of punishment can consult Appendix B.

pair in the same treatment, and a partner protocol across treatments (NOPUN and PUN).[10] To see how subjects react to the possibility of punishment, we control for order effects, i.e., subjects played either NOPUN or PUN first in half of the sessions.

Subjects received at the beginning of each treatment (PUN or NOPUN) a printed copy of the experiment instructions.[11] Instructions were read aloud by the session monitor and subjects were allowed to ask any question in private before starting the treatment. We minimized the probability of subjects missing how payoffs were generated with a pre-experimental quiz.

Each round subjects received an initial endowment of $e_i \in \{10{,}40\}$ Experimental Currency Units (hereafter, ECUs), for i$\in\{a,b\}$. After knowing the distribution of endowment for the round $(e_a, e_b) \in \{(10{,}10), (10{,}40), (40{,}10), (40{,}40)\}$, the investor (subject *a*) had to decide the amount of ECUs that she wanted to send (if any) to the allocator (subject *b*). This amount was tripled and then received by the allocator who decided how much to return.

The treatment PUN included a punishment stage in which the investor had the possibility of sending "points" to the allocator after knowing the (interim) payoffs (i.e., value-laden terms were avoided and we did not use the word punishment in the instructions). Sending one point to the allocator cost 1 ECU to the investor and decreased allocator's payoff in 3 ECUs (i.e., λ = 3). This corresponds to the high-punishment treatment in Rigdon (2009). To facilitate the computation of the final payoffs, the investor decided the points to be sent to the allocator using an slider bar that ranged from 0 to P^*, where P^* = min$\{\pi^a, \pi^b/3\}$. By moving the bar, the investor received information about the final distribution of payoffs associated to her choice. The investor could move the sliding bar as many times as she wanted; her decision had to be confirmed by clicking a button at the bottom of the screen.

In both treatments, subjects played the four different distributions of endowment and were informed about their payoffs in the round before being re-matched. We control for the order in which distributions were played so that subjects did not play the distributions in the same order within or across treatments. Instructions informed subjects that they would receive at the beginning of each round "an amount of ECUs that can be either 10 or 40" and that the amount that she gets "does not need to coincide with the amount of ECUs received by the other player", although subjects always knew both amounts before taking their decision. We chose not to inform subjects that they would play the four distributions of endowments so as to avoid that subjects "smoothen out" incentives across distributions.[12]

[10] Each session consisted of 24 subjects, divided in 3 groups of 8 subjects. Within each group, 4 subjects were assigned the role of investors (subject *a*) and 4 subjects were allocators (subject *b*). Subjects from different matching groups never interacted with each other throughout the session.

[11] A translated version of the instructions is available in Appendix A. This contains further details about the experimental design (including some screenshots).

[12] Subjects could not decide to refrain from punishing if their capacity is low to do it more harshly when their capacity was high as subjects only knew that endowments may change across rounds, but were unaware that all of them would play the four distributions.

At the end of the session one of the two treatments (PUN or NOPUN) was randomly selected to pay subjects.[13] Subjects received on average 15 Euros (Exchange rate: 10 ECUs = 1 Euro). Each session lasted around 1 hour and a half, and included a brief questionnaire at the end of the two treatments that was used to collect demographic and other information to be used as control variables in the econometric analysis.

4. Behavioral Predictions and Experimental Hypotheses

We note that the difference between the investor and the allocator's endowment is positive in the distribution (40, 10), in which the investor is wealthier than the allocator. In the distribution (10, 40), the difference is negative because the allocator is wealthier than the investor. In distributions (10,10) and (40,40), the difference is zero because both the investor and the allocator are endowed with the same amount. We use these last two distributions as a natural control for the size of stakes (Johansson-Stenman et al. 2005).

Different behavioral models make different predictions depending on the difference between the investor and the allocator's endowment. The self-interest model, for example, assumes that subjects are exclusively motivated by their material payoff. Trust and trustworthiness will not be sensitive to endowment heterogeneity, as they will be zero in all distributions. Costly punishment will never occur regardless of the difference between the investor and the allocator's endowment. This predicted outcome is not only inefficient but in sharp contrast with the observed behavior in previous laboratory experiments (Cooper and Kagel 2009, Johnson and Mislin 2011, and Eckel and Wilson, 2011 survey these results).

There are other behavioral theories that provide more reasonable predictions on the effects of punishment under endowment heterogeneity. In what follows, we present different hypotheses on the effect of punishment with endowment heterogeneity. The first one in Section 4.1 deals with social preferences, in particular with inequality aversion and reciprocity. Section 4.2 is related to deterrence, and states that punishment should be beneficial for trust and trustworthiness, regardless of the distribution. We discuss the prediction of models on the hidden cost of punishment (i.e., intrinsic and extrinsic motivation) in Section 4.3. Our hypothesis for the order in which treatments are implemented is presented in Section 4.4. In Section 4.5 we present the behavioral hypotheses for the investor's willingness to punish in each of the distributions. Section 4.6 summarizes our testable hypotheses and highlights the differences between all the models discussed in this section.

4.1. Social Preferences

[13] We decide to pay a whole treatment rather than a round to avoid extreme variance in subjects' payoffs -e.g., if round 1 was selected for payment, we would had some subjects who played the distribution (10,10) and subjects who played with (40,40) in that round. By paying one treatment, we ensure that each subject received the money that she accumulated over the four distributions. We could have also paid only for every decision or one random round, although all these methods (including the one we chose) have pros and cons but can be accepted upon different assumptions on preferences. See Azrieli et al. (2014), Cox et al. (2014b), Harrison and Swarthout (2014) for a discussion along these lines.

Models of inequity aversion (e.g., Fehr and Schmidt 1999, Bolton and Ockenfels 2000) predict that endowment heterogeneity will affect the investor and the allocator's behavior in the trust game. In particular, for any possible endowment $e_a \in \{10,40\}$, the inequality-averse investor should send a greater or equal amount when the allocator's endowment is low ($e_b = 10$) compared with the amount sent when the allocator's endowment is high ($e_b = 40$). Along these lines, when the investor is wealthier than the allocator, any amount send will reduce inequality, whereas any amount sent when the allocator is wealthier than the investor will induce more inequality. As a result, higher levels of trust are expected when the investor's endowment increases.[14] As for trustworthiness, these arguments predict that the inequality-averse allocator should return an amount that depends negatively on the difference in the initial endowments (Ciriolo 2007, Xiao and Bicchieri 2010, Smith 2011).[15] When punishment is allowed, the inequality-averse investor will punish more harshly to wealthier allocators to reduce payoff inequality.[16] This behavior, in turn, reinforces the pattern of trustworthiness described above, which is in contrast with the prediction of reciprocity, where the levels of trustworthiness should not be affected by the possibility of punishment or the difference in the initial endowments.[17]

Hypothesis 1. (*Inequality-aversion and endowment heterogeneity*). Inequality-aversion predicts that the level of trust will be decreasing on the allocator's endowment and that the investor sends more when she is wealthier than the allocator. Trustworthiness should decrease when the investor is wealthier than the allocator, both without and with punishment.

4.2. The deterrence hypothesis and the positive effects of punishment

The higher the severity (and the probability) of the punishment the smaller the incentives for an agent to undertake the action that is subject to the sanction. This behavioral hypothesis, referred to as the *deterrence hypothesis* in Gneezy and Rustichini (2000b), is the first one that we want to test in our trust game, where the deterrence incentives are determined by the investors' capacity of punishment (see also Gneezy et al., 2011). This, in turn, can be related in our setting to endowment heterogeneity.

Hypothesis 2 (*The deterrence hypothesis*). a) The levels of trust and trustworthiness are higher with punishment than without punishment. b) Once we allow for punishment, the levels of trust and

[14] Our predictions in this section assume that either the investor or the allocator is inequality-averse. Otherwise, investors will send everything to allocators in all the distributions (to generate efficiency gains) and allocators will then pay back to reduce payoff inequalities.

[15] If we follow Ciriolo (2007) or Xiao and Bicchieri (2010) and assume that the inequity averse allocator wants to equalize payoffs, we can equalize equations (1) and (2) in Section 2 to see that the allocator should return an amount $Y = f(X, e) = 2X + (e_b - e_a)/2$, therefore changing the initial distribution of endowments should affect the allocators' behavior.

[16] If we equalize equations (3) and (4), use the value of $\lambda = 3$ and rewrite the expression, we find that the investor who wants to equalize final payoffs will choose an amount of punishment $\hat{P} = \frac{X}{2} - Y + \frac{(e_b - e_a)}{2} \geq 0$, therefore endowment heterogeneity plays a role in their decision to punish. We note that decisions in the trust game are also relevant for the inequality-averse investor in that the more the investor (allocator) decided to send (return), the more (less) harshly the punishment will be.

[17] As defined in Coleman (1990) or Ciriolo (2007), reciprocity implies that allocators should pay back what they have received from investors $Y = f(X)$, therefore changing the initial distribution of endowments or allowing for punishment should not affect their behavior.

trustworthiness are non-decreasing with the difference between the investor and the allocator's endowment.[18]

4.2. Models of intrinsic and extrinsic motivation: the hidden cost of punishment

In the trust game, the intrinsic motivation for the investor relates to her willingness to trust. Punishment may have a direct *"incentive effect"* by altering subjects' extrinsic motivation, making investors more willing to trust. It may also have an indirect and opposite psychological effect on intrinsic motivation, *crowding out* the incentivized behavior. In line with Gneezy et al. (2011), we expect for the *incentive* effect to dominate the *crowding-out* effect when endowment heterogeneity favors the investor (i.e., if the stick is sufficiently big).[19] Otherwise, punishment will crowd-out intrinsic incentives and will result in a decrease of the level of trust.

Models on intrinsic and extrinsic motivation should operate also for the allocator and his willingness to reciprocate. Explicit incentives might signal distrust in a principal-agent relationship (Benabou and Tirole 2003, Falk and Kosfeld 2006; Ellingsen and Johanhesson 2008). In our game, allocators can perceive as unfriendly the possibility of punishment, what might produce a shift from positive reciprocity to negative reciprocity. This rationale identifies the negative effect of punishment on the allocator's *intrinsic* motivation. There is, however, an *incentive* effect that operates in the opposite direction and predicts that the allocator will be more willing to pay back as the difference between the investor and the allocator's endowment increases (again, when the stick is sufficiently big).

Hypothesis 3 (*Intrinsic and extrinsic motivation*). a) A crowding-out effect may offset any incentive effect linked with a boost of extrinsic motivation, decreasing of trust and trustworthiness b) The incentive effect dominates the crowding-out effect when the difference between the investor and the allocator's endowment (the investor's capacity of punishment) increases.

Thus, models based on the conflicting existence of intrinsic and extrinsic motivation will not predict that punishment always fosters trust and trustworthiness. The relationship between the capacity of punishment and the levels of trust might be non-monotonic.

4.4. The order effect: the lasting effects of incentives

The literature on intrinsic and extrinsic motivation highlights that behavior may be affected not only by current incentives but also by the incentives offered previously (see Gneezy and Rustichini 2000b). Our experimental design allows us to study possible crowding out and lasting effects caused by the

[18] In a dynamic theoretical model, Olcina and Calabuig (2008, 2015) show that the levels of trust and trustworthiness increase with the capacity of punishment. In their setting, this capacity of punishment is given by the value of λ, which is exogenous given in that the maximum proportion of the allocators' payoff that investors can destroy is not determined by differences in initial endowments but by the legal system of the country.

[19] In our game, there exists an additional feature that reinforces this prediction: the credibility of punishment is partially determined by trusting behavior. Because the investor has to build up her own capacity of punishment, she may want to reduce the amount sent in order to make the punishment more credible. This *credibility effect* would reduce, and would naturally be weaker, when the investor is wealthier than the allocator.

introduction or the extinction of the punishment mechanism. The idea is that "*once incentives are removed, people may pursue the desired outcome less eagerly*" (page 192, Gneezy et al. 2011). In our setting, any increase in extrinsic motivation due to the possibility of punishment might have an additional lasting or long run effect decreasing future intrinsic motivation, apart from the possible short-run crowding-out effect on intrinsic motivation. The intensity of incentives indeed matters to determine how intrinsic and extrinsic motivation are affected (Gneezy et al. 2011). The stronger is the crowding out effect due to high-powered incentives, the larger is the lasting effect on intrinsic motivation. Because incentives are related to endowment heterogeneity through the capacity of punishment in our setting, we expect that removing punishment will negatively affect the investor and the allocator's behavior, especially when the investor is wealthier than the allocator.

Hypothesis 4 (*Lasting effect on intrinsic motivation*)**.** When punishment is removed, the level of trust and trustworthiness will decrease, especially if the difference in endowment is in favor of the investor.

4.5. Punishment likelihood

Previous models provide some useful predictions on the effect of endowment heterogeneity on the investor's willingness to punish. Models of inequality aversion, for example, propose that investors will be more willing to punish when allocators are wealthier than themselves to reduce payoffs inequality. In sharp contrast, investors' willingness to punish may depend on its implementation cost. Investors then should punish more when wealthier than allocators. Finally, reciprocal investors should not consider endowment heterogeneity but the levels of trust and trustworthiness to determine their behavior. Reciprocal investors should punish more if what they receive back is a small amount relative to what they sent, regardless of the relative level of endowments.

Hypothesis 5 (*Punishing behavior*)**.** Inequality aversion predicts that punishment will be more pervasive when the allocator is wealthier than the investor. Models that rely on the capacity of punishment, however, predict the opposite. As for reciprocal investors, their decision to punish should not be affected by endowment heterogeneity.

4.6. Testable predictions: Summary

Our experimental design allows us to test two inequality-aversion clear-cut predictions on the effect of endowment heterogeneity: (i) allocators will return less when investors are wealthier, and (ii) investors will punish more when allocators are wealthier.

The experimental results can also be used to disentangle between the competing predictions of the deterrence hypothesis and models on intrinsic and extrinsic motivation in two different ways: (i) Whereas the deterrence hypothesis predicts a positive effect of punishment on trust and trustworthiness, models on intrinsic and extrinsic motivation allow for non-monotonicity (e.g., punishment backfires), and (ii) introducing punishment should (should not) decrease trust according to models on intrinsic and extrinsic motivation (the deterrence hypothesis).

5. Results

In this section, we present our main experimental results. We study the investor's behavior in Section 5.1. The allocator's behavior is presented in Section 5.2. We discuss the efficiency of punishment, and briefly comment on the investor's punishing behavior in Section 5.3.

Since investors might differ in their endowment and in order to make meaningful comparisons on their behavior, we use the investment ratio (*x*) as the measure of trust (see Smith, 2011; Johnson and Mislin, 2011). This is defined as the proportion of the initial endowment that the investor sends to the allocator (X/e_a). The level of trustworthiness (*y*) is defined in our paper by the return ratio; i.e., the proportion of the received amount that the investor decides to return (Y/X).[20]

5.1. The Investor's Behavior

5.1.1. The effects of endowment heterogeneity on trust

We examine in this section how the investor behaves by examining the effect of punishment on trust. We present an overview of our data in Figure 1, where we plot the distribution of trust in both treatments by considering each possible distribution of endowments separately. We group the data considering investors who transfer between [0-10%], (10-25%], (25-50%] and (50-100%] of their initial endowment. The descriptive statistics are given in the table below the figure, where we report the average level of trust and the standard deviation (in brackets). The table includes the results of the t-test and the Wilcoxon signed-rank that compare the level of trust in each distribution with and without punishment.

[Figure 1 around here]

We can use the investor's behavior in the absence of punishment test inequality aversion. If investors were inequity averse, they would send more when the allocator's endowment is low, which does not seem to be the case. There is no statistical difference between the amounts sent in (10, 10) and (10, 40) (p-value = 0.372) or between the amounts sent in (40, 10) and (40, 40) (p-value = 0.942), contrary to inequality aversion that predicts less (more) trust when the allocator's endowment increases. Indeed, our findings indicate that investors trust more when allocators are wealthier than they are. For example, investors are more likely to send nothing in the (40,10) distribution, compared with the (10, 40) distribution (p-value = 0.032).[21]

[20] This is a good measure of reciprocity in that a return ratio smaller than 1 (larger than 1) indicates that the allocator gives back to the investor less than (more than) what he has received. We note, however, that all our findings are robust if we consider instead the proportion of the generated surplus that allocators return (Y/3X) as a measure for reciprocity (Ashraf et al. 2006, Chaudhuri and Gangadharan, 2007).

[21] One plausible explanation is that investors anticipate that allocators will be inequality-averse (which does not seem to be case, as it is shown in Section 5.2). It is also possible to assume that other factors such as risk aversion or attitudes to betrayal may be at

Although we reject the prediction based on inequality aversion models, we find support for the existence of an endowment effect. Without punishment, trust is smaller if the investor's endowment is high (40 ECUs), relative to the case in which her endowment is low (10 ECUs). This result is in line with previous findings suggesting that trust decreases as the size of inverstors' endowment increases (e.g., Johansson-Stenman et al., 2005, 2008).

Interestingly, the *endowment effect* vanishes with a high capacity of punishment. If punishment is available, trust in the (40,10) distribution (19.95%) is roughly the same as in distributions in which the investor's endowment is low (21.81% and 25.83%). But, trust in the (40,40) distribution is significantly smaller.[22]

Result 1. (Endowment effect) *Without punishment, we find no support for an endowment effect, and trust decreases as the investor's endowment increases. With punishment, only a high difference between the investor and the allocator's endowment prevents the appearance of an endowment effect.*

Note that our Result 1 suggests that trust follows the investor's endowment (i.e., endowment effect) more than endowment differences, in sharp contrast with inequality aversion predictions.

5.1.2. The effects of punishment on trust

Figure 1 strongly suggests that the distribution of trust is roughly the same with and without punishment, except in the (40,10) condition, in which the average level of trust nearly doubles (11.2% versus 19.9%). The t-test and the non-parametric Wilcoxon signed-rank test reject the null hypothesis that punishment does not affect the level of trust in the distribution (40,10), but they fail to reject the same hypothesis in the rest of the distributions.[23]

Result 2. (Effect of punishment on the level of trust) *Punishment fosters the level of trust only if the investor is wealthier than the allocator. Otherwise, the level of trust is not affected by the possibility of punishment.*

This result seems to reject the deterrence hypothesis, which predicts a positive effect of punishment on the level of trust, regardless of the difference between the investor and the allocator's endowment. To further discriminate between the deterrence hypothesis and models on intrinsic and extrinsic motivation, we next investigate investors' behavior by considering the order in which treatments are implemented.

5.1.2. Order effect: Trust and crowding-out

In our experimental design, subjects played either with or without punishment in the first part of the session. Figure 2 presents changes in the percentage of trust for each distribution, both when the

stake in the trust game (see Bohnet and Zeckhauser 2004, Hong and Bohnet 2007). Thus, investors may believe it is riskier to send money to allocators who are at a disadvantage position because they would be more likely to betray them.

[22] Table 2.B. in the appendix reports the p-values of pairwise comparisons between the four different distributions using the Wilcoxon signed-rank test and provide statistical support for our findings in this section.

[23] Our results hold when we control for the order in which treatments are implemented. Section 5.1.3 presents the results of a random effects model that also controls for unobserved individual heterogeneity and take into account the order in which treatments are implemented (see also Table 1.B in Appendix B, where we estimate a random effect model for each distribution of endowments separately).

punishment is introduced (solid line) and removed (dotted line) in the second part of the experiment. The effect of punishment in the pooled data is presented in bars as a percentage of the level of trust without punishment. Any positive (negative) value of this percentage should then be interpreted as trust being higher (smaller) with punishment than without.

[Figure 2 around here]

The deterrence hypothesis predicts that introducing the possibility of punishment should foster trust. Figure 2*,* however, suggests the opposite: the introduction of punishment has a detrimental effect on trust (consistent with a *crowding-out effect*), except if the investor is wealthier than the allocator. This is the only case in which punishment does not significantly change the proportion sent by investors (t = 0.62, p-value = 0.54). Following the same deterrence hypothesis, the extinction of the institution should be detrimental for the level of trust, we find that the effect is only significant when the investor is wealthier than the allocator: trust was 70% higher with punishment, going from 26.9% to 7.7% (t = 3.25, p-value = 0.003). In the rest of the distributions, changes on trusting behavior are also in the positive domain but are never significant (p-values > 0.208).[24]

Result 3. (Order effect and trust) *The introduction of punishment crowd-outs the level of trust by decreasing the relative amount sent, unless the investor is wealthier than the allocator. The extinction of punishment is detrimental for the level of trust only when the investor is wealthier than the allocator.*

In a nutshell, our data reject the idea of inequality aversion for the investor's behavior. As for the effect of punishment, our findings provide evidence against the deterrence hypothesis and seem to support models on the hidden cost of punishment (i.e., intrinsic and extrinsic motivation). In the next subsection, we rely on an econometric analysis so as to provide further evidence in favor of our findings. Because all our findings are in line with our previous discussion, readers may decide to skip this subsection.

5.1.3. Trust and punishment: Econometric analysis

Table 1 presents the maximum-likelihood estimates of a random effects model that controls for unobserved individual heterogeneity. In our specification, we follow Smith (2011) and Johnson and Mislin (2011), and consider the proportion of the endowment that the investor sends to the allocator (X/e_a) as the dependent variable. The set of independent variables include the period in which the decision is made (Period), the investor's net earnings in the previous round (Earnings t-1) and dummy variables for the existence of punishment (PUN), for the possibility of punishment being the first treatment to be implemented in the session (PUNFIRST) and for the possibility of subjects having a high level of endowment (i.e, the variable e_i^H for i$\in \{a,b\}$ takes the value 1 if e_i = 40 for i$\in \{a,b\}$, being 0 otherwise).

[24] Note that the solid (dotted) line is in the negative (positive) region indicating that level of trust is always lower in the second treatment to be implemented. The period and the order in which treatments are implemented might have an effect when the investor chooses how much out of her endowment to send. We account for this possibility in the regression analysis presented in the section 5.1.3, where we show that our results are robust to these features.

We include the interaction of some dummies as well as the data collected in the questionnaire such as the investor's age, the investor's gender, or the answer to the attitudinal survey question from the General Social Survey (GSS): "Generally speaking, would you say that most people can be trusted or that you cannot be careful in dealing with people?" Column 1 presents the estimates with all the controls. Columns 2 and 3 exclude the order in which treatments are implemented and the variables collected in the questionnaire respectively.[25] In each column, the reported standard errors (in brackets) take into account matching group clustering.

[Table 1 around here]

Specification 1 that controls for the order of the treatments and the demographic variables is our preferred specification. The baseline model considers that both subjects are endowed with a low level of endowment (10 ECUs) and there is no punishment. In that context, the estimated effect of punishment is negative and significant, what provides evidence for a *crowding-out effect* in the baseline distribution (10,10). The χ^2-test suggests that introducing punishment has a detrimental effect also in distributions (10,40) and (40,40) (p-values = 0.089 and 0.039, respectively). There is not a crowding-out effect when the capacity of punishment is high as the estimated effect of $PUNe_a^H$ is positive and sufficiently high to compensate the negative effect of *PUN* (p-value = 0.677). When we study the effect of removing the punishment, the χ^2-tests suggest that it is detrimental for the level of trust if the investor is wealthier than the allocator (p-value < 0.001), while the opposite is not true (p-values > 0.187). If we do not control for the order in which treatments are implemented (specification 2), we find that PUN is no longer significant in the baseline distribution; i.e., punishment does not have any effect if both subjects have 10 ECUs. Punishment fosters trust only if the capacity of punishment is high (p-value = 0.036). All these findings confirm Results 2 and 3 and provide further evidence against the deterrence hypothesis.[26]

Finally, we find that previous net earnings do not affect the level of trust, which decreases over time as already occurs in other experiments that involve repetition (e.g., public good experiments). In line with previous research, women in our experiment are estimated to trust less than men (e.g., Chaudhuri and Gangadharan, 2007; Buchan et al., 2008; Eckel and Grossman, 2008; Rigdon, 2009; Garbarino and Slonim, 2009). We also find that those investors who believe that people can be trusted are likely to send a higher proportion of their endowment.[27]

[25] The interested reader can see further details on the demographics in Appendix B. In our data, there is no correlation between gender and age (r = -0.06, p-value = 0.42). The answer to the GSS question is positively correlated with the investor's age (r = 0.18, p-value = 0.011). The correlation with the investor's gender is also positive but weakly significant (r = 0.12, p-value = 0.092).

[26] The estimated effects of the high level of endowments (e_a^H and e_b^H) and the interaction of the dummy variables do also support the findings summarized in Result 1 regarding the existence of an endowment effect that is compensated by the capacity of punishment when investors are allowed to punish (see Appendix B, Table 2.B). Additional results in the order effect are presented in Appendix B, Table 3.B. We also note that our results are robust to other specifications such as including *homegrown* trust (i.e., the amount sent by investors in the previous round) or the previous level of trustworthiness (i.e., the share received by investors in the previous round). We thank the referee for proposing these robustness checks, which are available upon request.

[27] This latter result differs from the reported data in Glaeser et al. (2000), where the GSS question does not have any predictive power on the level of trust. For a discussion about "trust questions" in the World Value Survey and behavior in the trust game, see Capra et al. (2008) and Johnson and Mislin (2012).

5.2. The Allocator's Behavior

5.2.1. The effects of endowment heterogeneity on trustworthiness

In this section we explore trustworthiness.[28] Figure 3 presents the average amount allocators send back (Y), the average return ratio ($y = Y/X$) and the frequency of trustworthy allocators.

[Figure 3 around here]

Without punishment, allocators roughly return 72% of what they receive, consistent with Johnson and Mislin (2011)'s meta-study. Inequality aversion predicts that allocators will be less likely to reciprocate when their endowment is lower than the investor's one. Panel B, however, suggests that allocators' behaviors is quite consistent across distributions, returning a fraction of what is generated. This, in turn, provides little support for the existence of inequality aversion (see Rodriguez-Lara 2015).

Result 4. *In the absence of punishment, endowment heterogeneity does not affect the return ratio. In particular, allocators return a proportion of what is generated that is consistent across distributions.*

Result 4 thus suggests that allocators are more reciprocal than inequity-averse. One possible explanation for this behavior is that allocators do not held themselves responsible for the exogenously determined initial distribution of endowments. They rather focus on what may be distributed when making their choices.

5.2.2. The effects of punishment and endowment heterogeneity

Figure 3 indicates that the possibility of punishment fosters trustworthiness, except if investors are wealthier than allocators. In the distribution (40,10), the fear of punishment does not significantly increase the amount that allocators return (8.409 versus 9.000), the return ratio (0.660 versus 0.711), or the frequency of allocators returning a positive amount (0.727 versus 0.742). While allocators were expected to reciprocate more if the investor's capacity of punishment was high, our findings show that any positive effect of punishment on trustworthiness decreases on the difference between the investor and the allocator's endowment.

Result 5. (Effect of punishment on the level of trustworthiness) *Punishment increases the return ratio and the likelihood of returning a positive amount, except if the investor is wealthier than the allocator.*

Our rationale for Result 5 is that when allocators receive a higher transfer with punishment (as it actually happens when investors are wealthier than allocators), they might perceive the larger offer as being associated to the possibility of being sanctioned, rather than to the investor's intrinsic motivation, being less willing to reciprocate by increasing the return ratio. To provide some statistical content on this conjecture, we rely on the measure of reciprocity in Berg et al. (1995) and compute the correlation between the amount allocators receive and the return ratio. The Pearson's correlation coefficient suggests

[28] We only consider allocators who received a positive transfer from investors, as those receiving nothing cannot return any amount. Hereafter, we drop all observations in which allocators received no money from investors for a meaningful analysis.

that the return ratio is negatively correlated with the amount received by allocators when there is punishment (r = -0.19, p-value=0.032). The correlation is not significant when there is not (r = -0.08, p-value=0.437).[29, 30]

5.2.3. Order effect: Trustworthiness and crowding-out

Figure 4 depicts the percentage changes in the return ratio when punishment is introduced (solid line) and when it is removed (dotted line). The effect of the punishment in the pooled data (that is decreasing with the capacity of punishment) appears as grey bars.

[Figure 4 around here]

Introducing punishment has a positive effect on trustworthiness, albeit it vanishes as the capacity of punishment increases. When punishment is removed, the return ratio goes down by roughly 40%, except if the investor is wealthier than the allocator. In that case, the extinction of the punishment is beneficial for the return ratio, although the effect is not significant.[31]

Result 6. (Order effects and trustworthiness). *The introduction of punishment fosters the return ratio, except when the investor is wealthier than the allocator. The extinction of punishment is detrimental for the return ratio only if the investor is wealthier than the allocator.*

These findings provide evidence against Gneezy et al. (2011) hypothesis as described above (the introduction of punishment should decrease (increase) the intrinsic (extrinsic) motivation to reciprocate, and foster trustworthiness when the capacity of punishment is sufficiently high). In sharp contrast, we find that allocators tend to pay back more with punishment than without punishment, except when the investor is wealthier than the allocator. The econometric analysis in the next subsection supports our previous findings. In our analysis, neither the allocator's gender nor the GSS question have a predictive power on the return ratio.

5.2.4. Trustworthiness and punishment: Econometric analysis

Table 2 provides the results of a panel random-effect regression where the dependent variable is the return ratio (y). The set of independent variables includes the amount of ECUs received from the

[29] Although previous studies do not allow for punishment, our results in the absence of punishment are consistent with Berg et al. (1995). Chaudhuri and Gangadharan (2007) find "substantial evidence in favor of positive reciprocity in the sense that receivers do return money to the senders and the amount returned is positively correlated with the amount received" (page 960). The same conclusion is found in Rigdon (2009). Our data do also provide a significant and positive correlation between the amount sent by investors and the amount returned by allocators at the 1% significance level (NOPUN: $r_s = 0.37$, p-value<0.0001; PUN: $r_s = 0.38$, p-value<0.0001). We do not rely on these correlations to analyze reciprocity because "absolute amounts sent and amounts received will bias the correlation statistic upwards, i.e., low amounts sent preclude some high returns" (Berg et al. 1995, page 131).

[30] As it was the case for investors, a detailed analysis of the allocator's behavior can be found in Section 5.2.3. The results are in line with the descriptive statistics presented in Figure 3, and suggest a positive effect of punishment in the return ratio, except if the capacity of punishment is high.

[31] Our regression analysis in Appendix B supports these findings and suggest that there is no order effect in the sense that punishment is always higher with punishment, except in the (40,10) distribution, and this is true regardless of the order in which the PUN treatment is implemented.

investor, a dummy for the possibility of having been punished in the previous round (Punished t -1) and the controls for the period, the previous net earnings, the level of endowments, the possibility of punishment and the demographic variables.

[Table 2 around here]

Specification 1 that controls for the order of the treatments and the demographic variables is again our preferred specification. The baseline model considers that both subjects are endowed with a low level of endowment (10 ECUs) and there is no punishment. In that context, endowments (e_a^H and e_b^H) do not have any effect on the return ratio in the absence of punishment, thus we fail to support the hypothesis of inequality aversion. With punishment, however, we find that investors tend to return less as the investors' capacity of punishment increases. Thus, we find significant differences in behavior using the χ^2-test when we compare the return ratio in the distribution (10, 40) and the distribution (40,10) (p-value < 0.07).[32]

Although there is no effect of punishment in the baseline distribution, we find that the effect of punishment on the return behavior depends on endowment heterogeneity. Take, for example, the distributions (10,40) and (40,10). We find that introducing punishment in the former distribution has a positive effect on the return ratio (p-value = 0.031), whereas eliminating punishment has a significant negative effect on the return ratio (p-value = 0.026). There is no effect, however, of introducing or eliminating punishment in the distribution (40,10) (p-values = 0.451 and 0.540 respectively).

Our estimates indicate that the amount that allocators received from investors, the allocator's net earnings in the previous rounds or the fact that allocators were punished in the previous round, do not have any predictive value when estimating the return ratio. Note that this latter result is perfectly in line with our experimental design as we are using a perfect-stranger protocol. Finally, our results indicate that the return behavior is gender invariant as already suggested in Rigdon (2009), and that the GSS question has no predictive power on the allocator's behavior.

5.3. Efficiency of punishment and punishing behavior

In the trust game, investors' decision is linked to efficiency gains. Our results in Section 5.1 indicate that investors send a higher proportion of their endowment when they are wealthier than allocators. Given that punishment is costly for both the investor and the allocator, we do control for the cost of punishment to assess any net efficiency gain. We use the sum of initial endowments as a reference point (0% in the horizontal axe) and we measure any efficiency gains (EG) relative to the same game played without punishment. Figure 5 plots changes in earnings for every distribution and treatment, as a percentage of the surplus obtained in the trust game using the measure EG = $(\pi^a + \pi^b)/(e_a + e_b)$ - 1.

The first column plots the percentage of generated surplus in the corresponding treatment without punishment as a natural benchmark. The second column plots the interim payoffs; that is, the surplus

[32] We report the value of the χ^2-tests for pairwise comparisons across distributions in Appendix B, Table 4.B. We find that () show that the order in which treatments are implemented is not important to explain the allocator's behavior in Appendix B, Table 5.B.

generated as a percentage net of punishment costs (i.e., the efficiency gains *before* investors employ the punishment). The third column shows the ratio between final payoffs and initial endowment once punishment has taken place, and its cost has been accounted for. Errors bar reflect one standard error.

[Figure 5 around here]

As already suggested by Result 2, we find no interim efficiency gains associated to punishment, except if the investor's capacity of punishment is high (p-value = 0.017).[33]

Final earnings with punishment (column 3) are not significantly different from initial endowments (the 0% value in the vertical axe), except in the distribution with high capacity of punishment, where final payoffs are larger than the sum of initial endowments (p-value = 0.048). Comparing columns 2 and 3, we see that investors do not refrain from punishing in any distribution (p-values < 0.002). Punishment generates significant efficiency losses in every distribution relative to the baseline (with no punishment, first column) (p-values < 0.018). The only distribution in which final earnings with and without punishment do not differ is again the one in which the capacity of punishment is high (p-value = 0.653).[34]

Result 7. (Efficiency of punishment) *Investors punish allocators in all the distributions and destroy any (possible) efficiency gains. Only if the investor is wealthier than the allocator, the sum of the final payoffs with punishment is larger than the sum of the initial endowments. In the rest of distributions the final payoffs with punishment are not significantly different from the initial endowments.*

As for punishing behavior, inequality aversion, reciprocity and the capacity of punishment posit different predictions depending on endowment heterogeneity and decisions in the trust game. Table 3 presents the maximum-likelihood estimates of a random effects model that controls for unobserved individual heterogeneity. Our first specification provides some insights into whether or not investors decided to punish (i.e., the dependent variable is a dummy that takes the value 1 if the investor punished the allocator).[35] We consider the return ratio ($y = Y/X$) as a measure of reciprocity. The set of independent variables include the period in which decision is made, earnings in the previous round, the levels of endowments, the order in which punishment was implemented and the demographic variables. Column 2 replicates the analysis when the dependent variable is the amount *P* that investors devote to punish. Column 3 reports the estimates when we constraint our analysis to investors who decided to punish. The reported standard errors in each column (in brackets) take into account matching group clustering.

[Table 3 around here]

Our results highlight that investors care about reciprocity in that the return ratio affects both the likelihood of punishment and the amount devoted to punish. In particular, the higher the return ratio is,

[33] The reported test statistics in this section are based on the t-test for difference in efficiency gains. All tests reported use a two-sided alternative. Using a non-parametric analysis does not change the statistical results.

[34] Overall, these results suggest that investors decide to punish and generate efficiency losses, as already found in public good experiments (Chaudhuri, 2011). The interested reader can consult Appendix B (Table 6.B) to see how investors and allocators' earnings are affected by the possibility of punishment. Our results indicate that punishment decreases investors' payoffs only if the investor is wealthier than the allocator, while allocators' earnings are lower with punishment than without punishment in all distributions except in the one in which the investor is wealthier than the allocator.

[35] We focus the econometric analysis on investors who decided to trust as they are the ones allowed to punish. The interested reader on the Spearman correlation coefficients can see Appendix B, Table 7.B.

the less likely (and less harshly) investors punish. Interestingly enough, the levels of endowment do not seem to affect punishment behavior except for the likelihood of punishment, which depends negatively on the allocator's endowment. This result suggests that investors are less likely to punish when allocators are wealthier than they are what, in turn, provides no much evidence in favor of inequality aversion. Arguably, the capacity of punishment seems to be the main determinant of punishing behavior.

Result 8. (Punishing behavior) *Investors are more likely to punish allocators if the return ratio is low (consistent with the existence of reciprocity concerns). Invertors do not seem to use punishment to equalize payoffs, so that inequality-aversion does not explain their behavior. The capacity of punishment, however, seems to play an important role as investors punish less frequently when allocators are wealthier than they are.*

6. Discussion and concluding remarks

In this paper, we have studied how the endowment heterogeneity influence trust and trustworthiness in a simple but well-known strategic setting: the trust game (Berg et al., 1995). Using a within-subjects design, we manipulated both the investor and allocator's endowment and find that endowment heterogeneity plays no role in explaining the level of trust or trustworthiness in the trust game without punishment, except in that the proportion sent by investors decrease with her endowment (see Johansson-Stenman et al., 2005, 2008). The endowment effect disappears with punishment. These findings provides little support for models of social inequality-aversion (see Bereby-Meyer and Niederle (2005), Chowdhury and Jeon (2014), Rodriguez-Lara (2015), for similar results), suggesting that behavior in the trust game may be "altruism plus" as already pointed out in Cox (2004, 2014a).[36]

Interestingly, endowment heterogeneity does play a major role in explaining changes in behavior when we allow investors to punish allocators. If the investor is wealthier than the allocator, punishment fosters trust, although it fails otherwise. As for trustworthiness, the effect is just the opposite. Allocators are less willing to pay back with punishment than without punishment when investors are wealthier. In the other distributions, the possibility of punishment increases the return ratio. Punishment also generates a crowding out effect in trust across all distributions, except when the investor is wealthier than the allocator. Symmetrically, the extinction of punishment is detrimental for the return ratio in all distributions, except in the distribution in which the investor is wealthier than the allocator.

Overall, these findings are hardly coherent with the *deterrence hypothesis* or *models of inequality aversion*, but are more consistent with an explanation based on the hidden cost of punishment (i.e., intrinsic and extrinsic motivation). This literature highlights two distinct effects of sanctions. First, the *incentive effect* makes the incentivized behavior more attractive (as in Prendergast 1999), increasing the extrinsic motivation to give and reciprocate. Second, the introduction of punishment has an opposite psychological effect on intrinsic motivation, causing a *crowding-out* effect (see Deci et al., 1999; Frey, 1997; Gneezy et al., 2011). This later effect is absent in the *deterrence hypothesis* that would never predict a negative effect of punishment on the

[36] Cox (2004) compares investors' giving when allocators are endowed and when they are not so as to isolate the effect of trust from the one of altruism.

level of trust and trustworthiness. Although this possibility is not excluded by models based on the hidden cost of punishment, the incentive effect should dominate the crowding-out effect when the stick is sufficiently high, that is, when investors are wealthier than allocators.

In our paper, investors' behavior is in line with the "pay-enough-or-do-not-pay-at-all" idea, as described in Gneezy and Rustichini (2000a) and Gneezy et al. (2011), which would plead for increasing the differences in endowments in favor of the investors to foster the level of trust.[37] When we analyze the level of efficiency, we indeed observe that sanctions wipe out the potential benefits of trust in almost every combination of endowments. Only if the investor is wealthier than the allocator, the sum of the final payoffs with punishment is larger than the sum of the initial endowments. In the rest of distributions the final payoffs with punishment are not significantly different from the initial endowments.

Our results also suggest that allocators do not respond to incentives as investors do. When sanctions are available, the positive *incentive effect* of sanctions on trustworthiness is dominated by the *crowding-out effect* when the investors' capacity of punishment is *too* high. Interestingly, this seems to indicate that the "pay-enough-or-do-not-pay-at-all" does not always operate for allocators. One possible explanation for allocators' behavior relies on the fairness of sanctions. Previous evidence in public good games concluded that sanctions might backfire if perceived as unfair (e.g., Denant-Boemont et al. 2007, Nikiforakis 2008). In the trust game, the negative effect of sanctions has always been reported in contexts in which investors punished for free (Fehr and Rockenbach 2003; Fehr and List, 2004; and Houser et al., 2008). In our experiment, the negative effect of costly punishment on trustworthiness only survives when the capacity of punishment is high.

[37] In our setting, there is an additional effect on behavior due to the fact that trust determines, together with the endowment heterogeneity, the credibility of sanctions. When the capacity of punishment is low, that is when allocators are wealthier than investors, the later ones may want to keep money in their pocket to make sanctions credible. This *credibility* effect is naturally weaker when the capacity of punishment is high; i.e., when investors are wealthier than allocators. This idea is consistent with our results. Punishment fosters trust only when the *stick* is big because the *incentive effect* dominates the *crowding effect* (as models on intrinsic and extrinsic motivation predict), but also because investors are not prisoners of the necessity of making sanctions credible, by sending less, when the capacity of punishment is high.

References

Abbink, K., Irlenbusch, B., Renner, E., 2000. The moonlighting game: An empirical study on reciprocity and retribution. Journal of Economic Behavior and Organization 42, 265- 277.

Armantier, O., 2006. Do wealth differences affect fairness considerations. International Economic Review 47, 391–429.

Arrow, K., 1974. The limits of organization. New York, Oxford University Press.

Ashraf, N., Bohnet, I., Piankov, N., 2006. Decomposing trust and trustworthiness. Experimental Economics 9, 193-208.

Azrieli, Y., Chambers, C. P., Healy, P. J., 2014. Incentives in experiments: A theoretical analysis, mimeo.

Bachmann, R., Zaheer, A., 2006. Handbook of trust research, Cheltenham; Edward Elgar.

Benabou, R., Tirole, R., 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70, 489-520.

Bereby-Meyer, Y. Niederle, M., 2005. Fairness in bargaining. Journal of Economic Behavior and Organization 56, 173-186.

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity and social history. Games and Economic Behavior 10, 122–142.

Bolle, F., Breitmoser, Y., Schlächter, S., 2011. Extortion in the laboratory. Journal of Economic Behavior and Organization 78, 207 - 218.

Bolton, G. E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. American Economic Review 90, 166-193.

Bohnet, I., Zeckhauser, R., 2004. Trust, risk and betrayal. Journal of Economic Behavior and Organization 55, 467-485.

Boyd, R., Gintis, H. Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proceedings of the National Academy of Sciences (USA) 100, 3531-3535.

Buchan, N., Solnick, S., Croson, R., 2008. Trust and gender: An examination of behavior, biases and beliefs in the investment game. Journal of Economic Behavior and Organization 68, 466-476.

Capra, C. M., Lanier, K., Meer, S., 2008. Attitudinal and behavioral measures of trust: A new comparison. Working Paper Emory University.

Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics 14, 47–83.

Chaudhuri, A., Gangadharan, L., 2007. An experimental analysis of trust and trustworthiness. Southern Economic Journal 73, 959–985.

Cherry, T., Kroll, S., Shogren, J., 2005. The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. Journal of Economic Behavior and Organization 57, 357–365.

Chowdhury, S. M., Jeon, J. Y., 2014. Impure Altruism or Inequality Aversion? An experimental investigation based on income effects. Journal of Public Economics 118, 143-150.

Ciriolo, E., 2007. Inequity aversion and trustees' reciprocity in the trust game. European Journal of Political Economy 23, 1007-1024.

Coleman, J. (1990). Foundations of social theory. Cambridge, MA: Belknap Press.

Cooper, D., Kagel, J., 2009. Other regarding preferences: A selective survey of experimental results. Forthcoming in J. H. Kagel, and A. Roth (Eds.), The handbook of experimental economics (Vol. 2). Princeton University Press.

Cox, J., C. 2004. How to identify trust and reciprocity. Games and Economic Behavior 46, 260-281.

Cox, J. C., Kerschbamer, R., Neururer, D., 2014a. What is trustworthiness and what drives it? Working Paper. University of Innsbruck. Department of Public Finance.

Cox, J. C., Sadiraj, V., Schmidt, U., 2014b. Paradoxes and mechanisms for choice under risk. Experimental Economics 18, 215-250.

Deci, E., Koestner, R., Ryan, R., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627-668.

Dufwenberg, M., Kirchsteiger, G., 2004. A Theory of Sequential Reciprocity. Games and Economic Behavior 47, 268-98.

Eckel, C., Grossman, P., 2008. Differences in the economic decisions of men and women: Experimental evidence. In Handbook of Experimental Economics Results, 1, Ed. C. Plott and V. Smith (509-519), New York, Elsevier.

Eckel, C., Wilson, R., 2011. Trust and social exchange. In the Handbook of Experimental Political Science, edited by J. Druckman, D. Green , J. Kuklinski and A. Lupia, Boston: Cambridge University Press, 243–257.

Ellingsen, T., Johannesson, M., 2008. Pride and Prejudice: The Human Side of Incentive Theory, American Economic Review 98, 990-1008.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293-315.

Falk, A., Kosfeld, M., 2006. The hidden cost of control. American Economic Review 96, 1611-1630.

Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425, 785-791.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Fehr, E., List, J. A., 2004. The hidden costs and returns of incentives—trust and trustworthiness among CEOs. Journal of the European Economic Association 2, 743-771.

Fehr, E., Rockenbach, B., 2003. Detrimental effects of sanctions on human altruism. Nature 422, 137–140.

Fehr, E., Schmidt, K., 1999. A Theory of Fairness, Competition and Cooperation. Quarterly Journal of Economics 114, 817-868.

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

Frey, B., 1997. Not just for the money. An economic theory of personal motivation. Cheltenham, UK and Brookfield, USA. Edward Elgar.

Gambetta, D., 1988. Can we trust?, in Gambetta, D. (Ed.), Trust: Making and Breaking Cooperative Relations, New York: Blackwell. 213-237.

Garbarino, E., Slonim, R., 2009. The robustness of trust and reciprocity across a heterogeneous U.S. population. Journal of Economic Behavior and Organization 69, 226–240.

Glaeser, E., Laibson, D., Scheinkman, J., Soutter, C., 2000. Measuring trust. Quarterly Journal of Economics 115, 811-846.

Gneezy, U., Rustichini., A., 2000a. Pay enough or don't pay at all. Quarterly Journal of Economics 115, 791–810.

Gneezy, U., Rustichini, A., 2000b. A fine is a price. Journal of Legal Studies 29, 1–18.

Gneezy, U., Meier, S., Rey-Biel, P., 2011. When and why incentives (don't) work to modify behavior. Journal of Economic Perspectives 25, 191–210.

Harrison, G.W., Swarthout, J. T., 2014. Experimental Payment Protocols and the Bipolar Behaviorist. Theory and Decisions 77(3), 423–438.

Hauert, C., Traulsen, A., Nowak, M. A., Brandt, H. H., Sigmund, K., 2007. Via freedom to coercion: the emergence of costly punishment. Science 316, 1905-1907.

Hong, K., Bohnet, I., 2007. Status and distrust: The relevance of inequality and betrayal aversion. Journal of Economic Psychology 28, 197- 213.

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

Johansson-Stenman, O., Mahmud, M., Martinsson, P., 2005. Does stake size matter in trust games? Economic Letters 88, 365–369.

Johansson-Stenman, O., Mahmud, M., Martinsson, P., 2008. Trust and religion: Experimental evidence from Bangladesh. Economica 76, 462-485.

Johnson, N. D., Mislin, A., 2011. Trust games: A meta-analysis. Journal of Economic Psychology 32, 865–889.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? A cross-country investigation. Quarterly Journal of Economics 112, 1251-1288.

Korenok O., Millner, E., L., Razzolini, L., 2012. Are Dictators Averse to Inequality? Journal of Economic Behavior and Organization 82, 543-547.

McCabe, K. A., Rigdon, M. L., Smith, V. L., 2003. Positive reciprocity and intentions in trust games. Journal of Economic Behavior and Organization 52, 267-275.

Nikiforakis, N., Normann, H.T., 2008. A comparative statics analysis of punishment in public-good experiments. Experimental Economics 11, 358-369.

Nikiforakis, N., Oechssler, J., Shah, A. 2013. Hierarchy, coercion, and exploitation: An experimental analysis. Journal of Economic Behavior and Organization 97, 155-168.

Olcina G., Calabuig, V., 2008. Cultural Transmission and the Evolution of Trust and Reciprocity in the Labor Market. Documentos de Trabajo, 11, Fundación BBVA.

Olcina G., Calabuig, V., 2015. Coordinated Punishment and the Evolution of Cooperation. Journal of Public Economic Theory 17, 147–173.

Prendergast, C., 1999. The provision of incentives in firms. Journal of Economic Literature 37, 7–63.

Rabin, M., 1993. Incorporating Fairness into Game Theory and Economics. American Economic Review 83, 1281-1302.

Reuben, E., Riedl, A., 2013. Enforcement of contribution norms in public good games with heterogeneous populations. Games and Economic Behavior 77, 122–137

Rigdon, M., 2009. Trust and reciprocity in incentive contracting. Journal of Economic Behavior and Organization 70, 93–105.

Rodriguez-Lara, I., 2014. Equal distribution or equal payoffs? Reciprocity and inequality aversion in the investment game, MPRA Paper No. 63313.

Smith, A., 2011. Income inequality in the trust game. Economics Letters 111, 54–56

Xiao, E., Bicchieri, C., 2010. When equality trumps reciprocity. Journal of Economic Psychology 31, 456-470.

Table 1

Table 1. The determinants of trust: random effects regressions

Dependent Variable: Investment ratio (x)			
	Model (1)	Model (2)	Model (3)
Period	-0.025**	-0.025**	-0.025**
	(0.011)	(0.011)	(0.011)
Net earnings t – 1	0.002	0.002	0.002
	(0.003)	(0.003)	(0.003)
PUN	-0.116**	-0.029	-0.118***
	(0.046)	(0.045)	(0.045)
PUNFIRST	-0.061		-0.022
	(0.067)		(0.064)
PUN PUNFIRST	0.176***		0.176***
	(0.055)		(0.054)
e_a^H	-0.148***	-0.148***	-0.145***
	(0.046)	(0.045)	(0.047)
e_b^H	-0.004	-0.004	-0.001
	(0.035)	(0.035)	(0.034)
$e_a^H e_b^H$	0.060	0.0592	0.051
	(0.044)	(0.045)	(0.044)
PUN e_a^H	0.128***	0.129***	0.130***
	(0.039)	(0.038)	(0.038)
PUN e_b^H	0.0425	0.043	0.044
	(0.0474)	(0.047)	(0.047)
PUN e_a^H e_b^H	-0.138*	-0.140*	-0.141*
	(0.080)	(0.080)	(0.079)
Women	-0.173***	-0.179***	
	(0.053)	(0.052)	
Age	-0.004	-0.005	
	(0.008)	(0.007)	
GSS	0.093**	0.101***	
	(0.041)	(0.036)	
Constant	0.500***	0.504***	0.320***
	(0.182)	(0.156)	(0.057)
σ_u	0.163	0.160	0.163
σ_e	0.201	0.207	0.201
ρ	0.397	0.374	0.396
R-square	0.156	0.132	0.072
Wald test (p-value)	< 0.0001	< 0.0001	< 0.0001
Number of obs.	288	288	288

Notes. The set of independent variables include the period in which the decision is made, the investor's net earnings in the previous round and dummy variables for the existence of punishment (PUN), for the possibility of punishment being the first treatment to be implemented in the session (PUNFIRST) and for the possibility of subjects having a high level of endowment (i.e, the variable e_i^H for $i\epsilon\{a,b\}$ takes the value 1 only if $e_i^H = 40$ for $i\epsilon\{a,b\}$, where a (b) stands for the investor (allocator). We include the interaction of some dummies as well as the data collected in the questionnaire such as the investor's age, the investor's gender, or the answer to the attitudinal survey question from the General Social Survey (GSS): "Generally speaking, would you say that most people can be trusted or that you cannot be careful in dealing with people?" The robust standard errors take into account matching group clustering and are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2

Table 2. The determinants of trustworthiness: random effects regressions

Dependent variable: return ratio (Y/X)

	Model (1)	Model (2)	Model (3)
Amount received (X)	-0.001	-0.005	-0.003
	(0.009)	(0.009)	(0.009)
Period	0.065	0.074	0.066
	(0.077)	(0.076)	(0.076)
Net earnings t – 1	0.0004	-0.0001	-0.0004
	(0.004)	(0.004)	(0.004)
Punished t – 1	0.240	0.280	0.208
	(0.194)	(0.188)	(0.193)
PUN	0.470*	0.397	0.435
	(0.275)	(0.245)	(0.270)
PUNFIRST	-0.240		-0.206
	(0.215)		(0.220)
PUN x PUNFIRST	-0.033		-0.009
	(0.253)		(0.250)
e_a^H	-0.111	-0.091	-0.101
	(0.266)	(0.265)	(0.265)
e_b^H	-0.068	-0.129	-0.112
	(0.252)	(0.244)	(0.249)
$e_a^H e_b^H$	0.291	0.324	0.277
	(0.354)	(0.352)	(0.352)
PUN e_a^H	-0.252	-0.264	-0.253
	(0.339)	(0.337)	(0.335)
PUN e_b^H	0.152	0.191	0.206
	(0.334)	(0.328)	(0.329)
PUN $e_a^H e_b^H$	0.001	-0.030	-0.024
	(0.478)	(0.475)	(0.471)
Women	0.054	0.062	
	(0.158)	(0.164)	
Age	0.0459*	0.041*	
	(0.024)	(0.025)	
GSS	-0.005	-0.062	
	(0.211)	(0.216)	
Constant	-0.510	-0.473	0.548*
	(0.613)	(0.628)	(0.296)
σ_u	0.316	0.354	0.375
σ_e	0.700	0.695	0.691
ρ	0.169	0.206	0.227
R-square	0.162	0.138	0.127
Wald test (p-value)	0.037	0.034	0.033
Number of obs.	153	153	153

Notes. The set of independent variables include the amount that allocators receive, the period in which the decision is made, the investor's net earnings in the previous round and dummy variables for the possibility of being punished the previous round, for existence of punishment (PUN), for the possibility of punishment being the first treatment to be implemented in the session (PUNFIRST) and for the possibility of subjects having a high level of endowment (i.e, the variable e_i^H for $i\in\{a,b\}$ takes the value 1 only if e_i^H = 40 for $i\in\{a,b\}$, where *a* (*b*) stands for the investor (allocator). We include the interaction of some dummies as well as the data collected in the questionnaire such as the investor's age, the investor's gender, or the answer to the attitudinal survey question from the General Social Survey (GSS): "Generally speaking, would you say that most people can be trusted or that you cannot be careful in dealing with people?" The robust standard errors take into account matching group clustering and are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3

Table 3. The determinants of punishment: random effects regressions

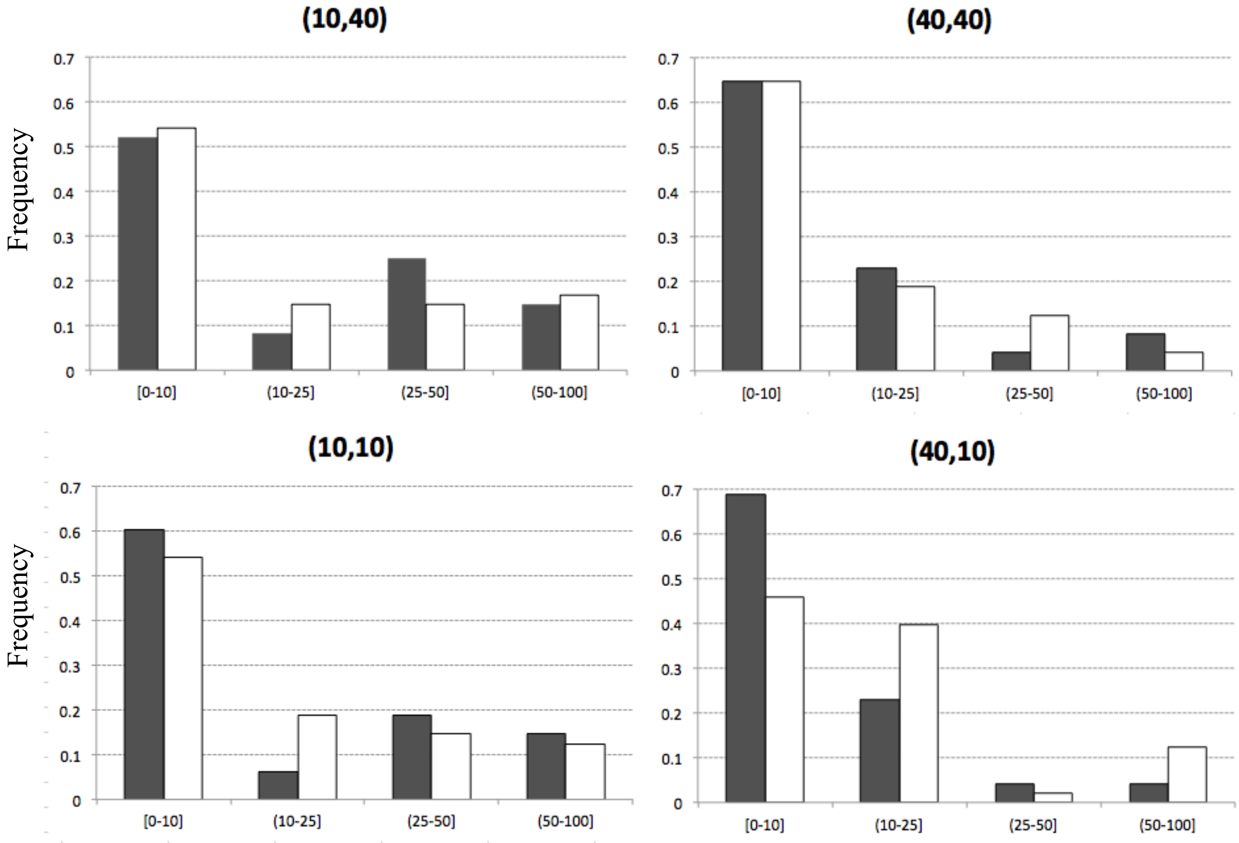
Dependent variables: Probability of punishment (Model (1)) and amount devoted to punish (Models (2) and (3)).

	Model (1)	Model (2)	Model (3)
Return ratio (Y/X)	-0.157***	-1.236**	-2.965**
	(0.06)	(0.51)	(1.32)
Period	-0.096*	-0.456	-1.243
	(0.05)	(0.59)	(0.85)
Earnings t – 1	0.001	-0.042	-0.057
	(0.004)	(0.03)	(0.05)
PUNFIRST	0.012	-1.197	0.802
	(0.12)	(1.318)	(2.51)
e_a^H	-0.032	1.268	1.365
	(0.15)	(1.15)	(2.10)
e_b^H	-0.197*	-0.051	0.421
	(0.11)	(0.74)	(1.30)
$e_a^H e_b^H$	0.157	0.270	1.469
	(0.17)	(1.70)	(3.20)
Women	0.133	-0.932	-1.937
	(0.15)	(1.29)	(2.52)
Age	0.067**	0.285	0.147
	(0.03)	(0.18)	(0.90)
GSS	-0.341*	-0.911	2.289
	(0.18)	(1.14)	(5.68)
Constant	-0.595	-0.048	7.343
	(0.89)	(4.33)	(18.85)
σ_u	0.239	2.116	4.163
σ_e	0.363	3.382	2.343
ρ	0.301	0.281	0.759
R-square	0.253	0.234	0.367
Wald test (p-value)	0.000	0.000	0.000
Number of obs.	82	82	32

Notes. The first model estimates the likelihood of punishment. The second model estimates the amount devoted to punishment. The third one replicates this analysis when we constraint to those investors who decided to punish a positive amount. The set of independent variables include the return ratio, the period in which the decision is made, the investor's earnings in the previous round and dummy variables for the possibility of punishment being the first treatment to be implemented in the session (PUNFIRST). We control for endowment heterogeneity with dummy variables (e.g, the variable e_i^H for $i\in\{a,b\}$ takes the value 1 only if $e_i^H = 40$ for $i\in\{a,b\}$, where *a* (*b*) stands for the investor (allocator), respectively). We include the data collected in the questionnaire such as the investor's age, the investor's gender, or the answer to the attitudinal survey question from the General Social Survey (GSS): "Generally speaking, would you say that most people can be trusted or that you cannot be careful in dealing with people?" The robust standard errors take into account matching group clustering and are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Figure 1

Figure 1. The effect of punishment on the level of trust: relative transfer in each distribution



	Distribution			
Treatment	**(10,40)**	**(40,40)**	**(10,10)**	**(40,10)**
■ NOPUN	0.266 (0.32)	0.151 (0.26)	0.243 (0.34)	0.112 (0.22)
□ PUN	0.258 (0.34)	0.141 (0.21)	0.219 (0.29)	0.199 (0.29)
t-test (t)	0.887	0.830	0.637	0.017
Wilcoxon test (Z)	0.737	0.950	0.612	0.050

Figure 2. The effect of introducing and eliminating punishment on the investor's behavior.

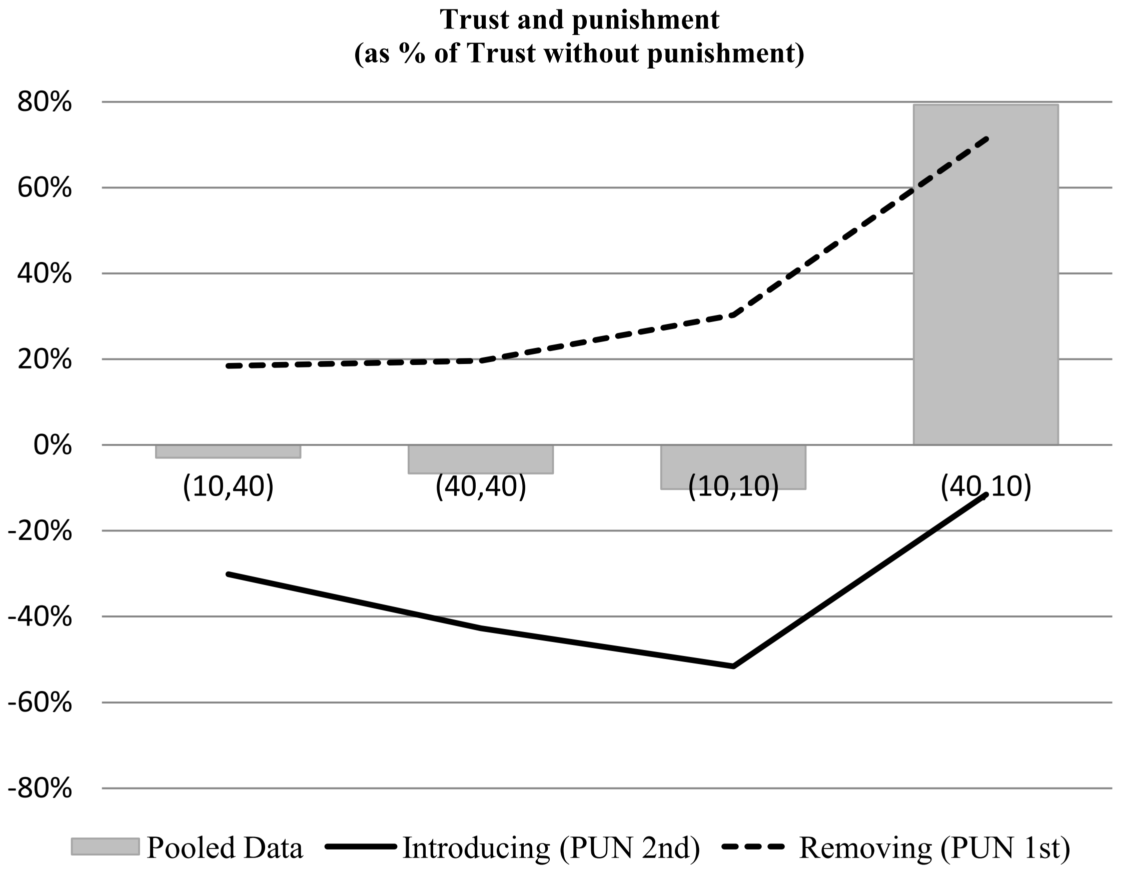


Figure 3

Figure 3. The allocator’s behavior in each treatment for each distribution

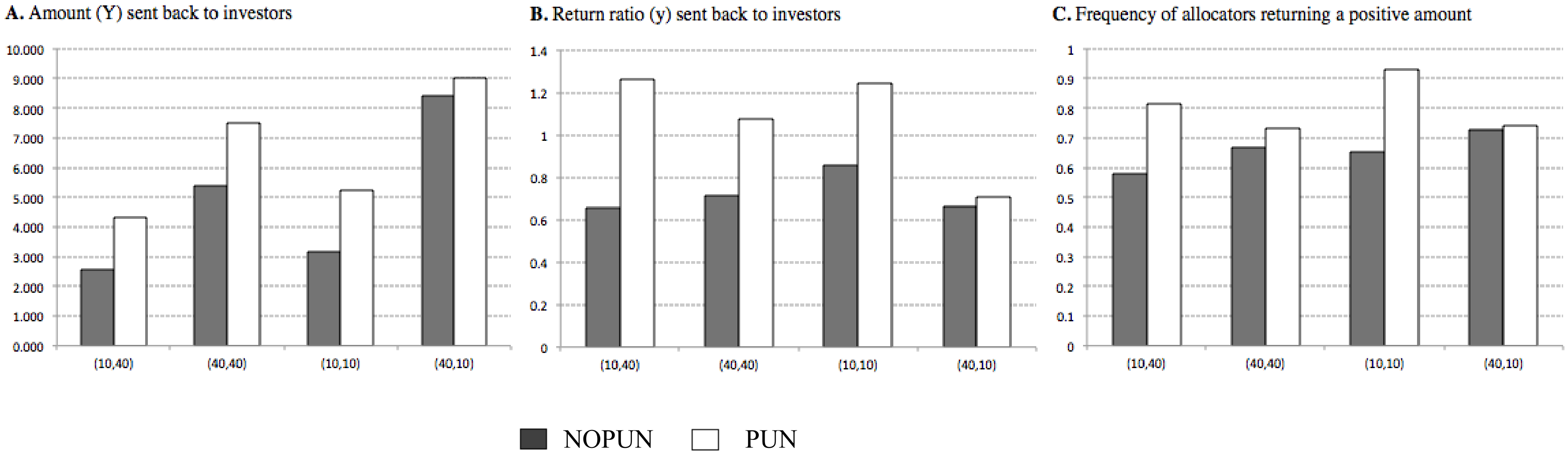


Figure 4

Figure 4. The effect of introducing and eliminating punishment on the allocator's behavior.

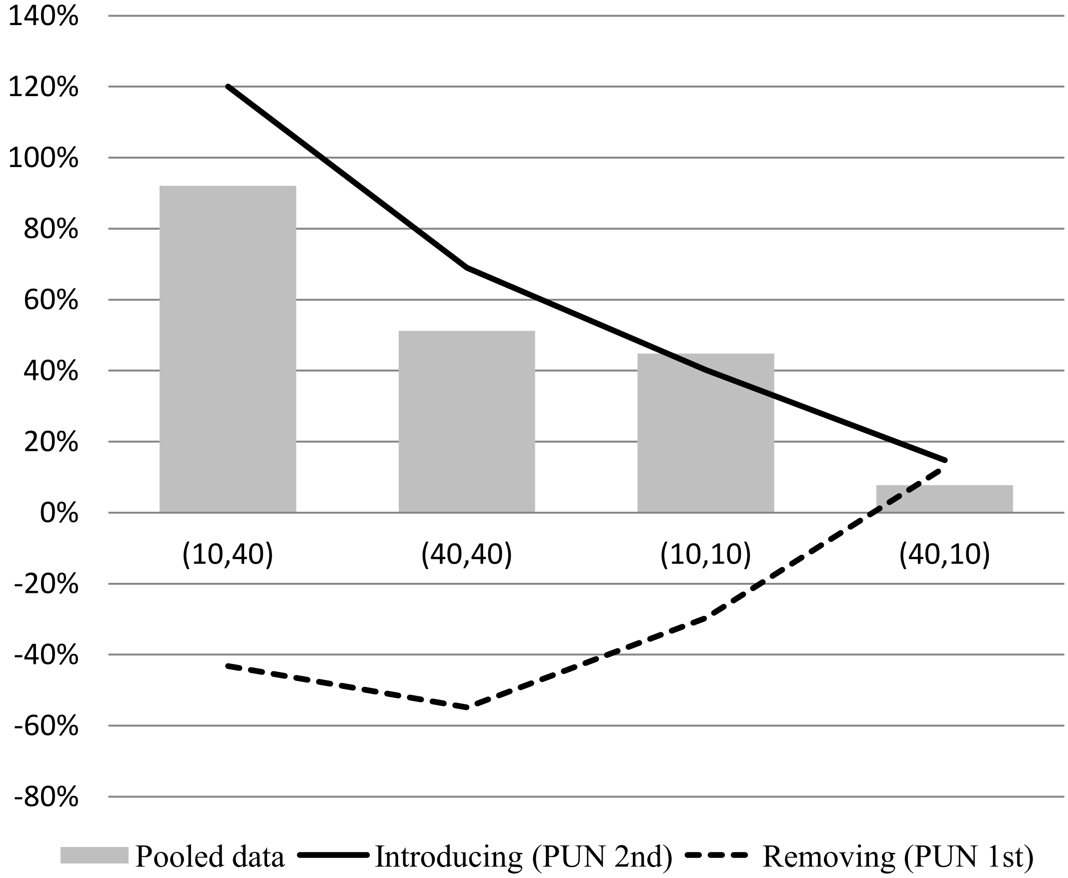
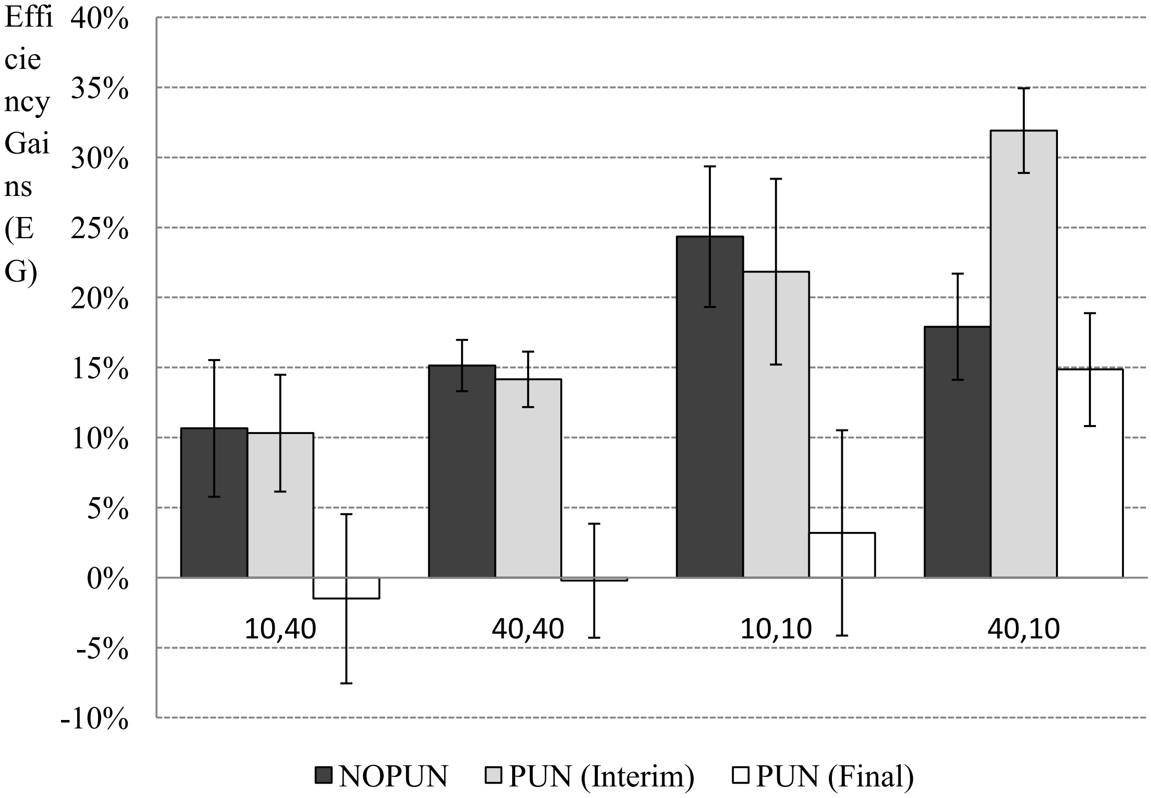


Figure 5

Figure 5. Efficiency gains in each distribution comparing the sum of the final payoffs and the sum of the initial endowments



Carry a big stick, or no stick at all

Punishment and Endowment Heterogeneity in the Trust Game

Supplementary Material

This supplementary material is divided in two sections. The first one (Appendix A) presents the experimental instructions and some screenshots (in Spanish) of our experiment. The second one (Appendix B) contains the derivatives for our lemma and supplementary econometric analyses of our data for investors and allocators, which support the findings discussed in our manuscript.

APPENDIX A:

INSTRUCTIONS [*]

This is an experiment to study decision-making. The instructions are simple and if you follow them carefully you will get an amount of money in cash at the end of the experiment in a confidential manner. All through the experiment you will be treated anonymously. Neither the experimenters nor the people in this room will ever know your particular choices or the amount of money that you get. Talking is forbidden during the experiment. If you have any questions, raise your hand and remain silent. You will be attended as soon as possible.

The experiment has 8 rounds, divided into 2 blocks of 4 rounds. These instructions explain how the experiment unfolds in the first block. At the beginning of the second block, you will be provided with new instructions. At the end of the experiment, one of the two blocks will be randomly selected to pay you. We will convert your gains in ECUs (Experimental Currency Units) during that block to Euros using the rate of 10 ECU= 1€.

In this experiment there are two types of players: A and B. Before starting the experiment, you will be randomly selected either as player A or player B and this type will be kept all through experiment.

In each round, you will be matched with one of the subjects of the other type (i.e., you will be matched with a player B if you are player A, and you will be marched with a player A if you are player B). In each block, you will never be matched with the same person twice. It means that in each block you will take decisions with a different person in each round.

At the beginning of each round, you will get an amount of ECUs that can be either 10 or 40. The amount that you get does not need to coincide with the amount of ECUs received by the other player you are matched with, although you will always know both amounts before taking your decision.

If you are player A, you have to decide the amount of ECUs (if any) to send to player B. The amount of ECUs that you send will be deducted from your initial ECUs and will be triplicated (i.e., we will multiply this amount by 3). The amount of ECUs that you don't send to player B will be yours.

If you are player B, you will get three-times the amount of ECUs that player A sent you. After you know this amount, you have to decide the amount of ECUs (in any) that you want to return to player A. You will keep the ECUs that you do not send to player A plus your initial ECUs.

So, in this block, your gains in each round depend of your decisions in the following way:

Final payoff player A = Initial ECU of A – ECU sent to B + ECUs received from B

Final payoff player B = Initial ECU of B + 3* ECU received from A - ECU sent to A

To check that you have understood the instructions, we ask you to look at the computer screen. First, you will see the logic of the experiment through a numerical example. Next, you will need to compute the final payoffs of an example in which in which the computer chooses numbers randomly the ECUs send by player A and the ECUs returned by player B.

[*] This appendix contains the instructions for the sessions in which the possibility of punishment is introduced in the second part of the experiment. Instructions are originally in Spanish.

INSTRUCTIONS

This is an experiment to study decision-making. The instructions are simple and if you follow them carefully you will get an amount of money in cash at the end of the experiment in a confidential manner. All through the experiment you will be treated anonymously. Neither the experimenters nor the people in this room will ever know your particular choices or the amount of money that you get. Talking is forbidden during the experiment. If you have any questions, raise your hand and remain silent. You will be attended as soon as possible.

This second block has a total of 4 rounds, in which you keep being player A or B. In each round, you will be matched with a person of the other type that changes across rounds. Thus, if you are player A (B), you will be matched in each round with a different player B (A). As in the first block, at the beginning of each round you will get an amount of ECUs that can be 10 or 40 ECUs.

Each round in this block hast two stages. The **first stage** is identical to the first block. If you are player A, you have to decide the amount of ECUs (if any) to send to player B. The amount of ECUs that you send will be deducted from your initial ECUs and will be triplicated (i.e., we will multiply this amount by 3). The amount of ECUs that you don't send to player B will be yours.

If you are player B, you will get three-times the amount of ECUs that player A sent you. After you know this amount, you have to decide the amount ECUs (in any) that you want to return to player A. You will keep the ECUs that you do not send to player A plus your initial ECUs.

These decisions determine your provisional payoffs.

Provisional payoff player A = Initial ECU of A – ECU sent to B + ECUs received from B

Provisional payoff player B = Initial ECU of B + 3* ECU received from A - ECU sent to A

In the **second stage** of the round, and after being informed of the provisional playoffs, the player A will be asked to take a second decision. This second decisions consists in choosing the number of **points** (if any) to send to player B. Each point that player A sends to player B will reduce the player A's payoff in 1 ECU. Per each point that player B receives from player A, the player B's payoffs will be reduced in 3 ECUs.

Your **final payoffs** will be then computed as follows:

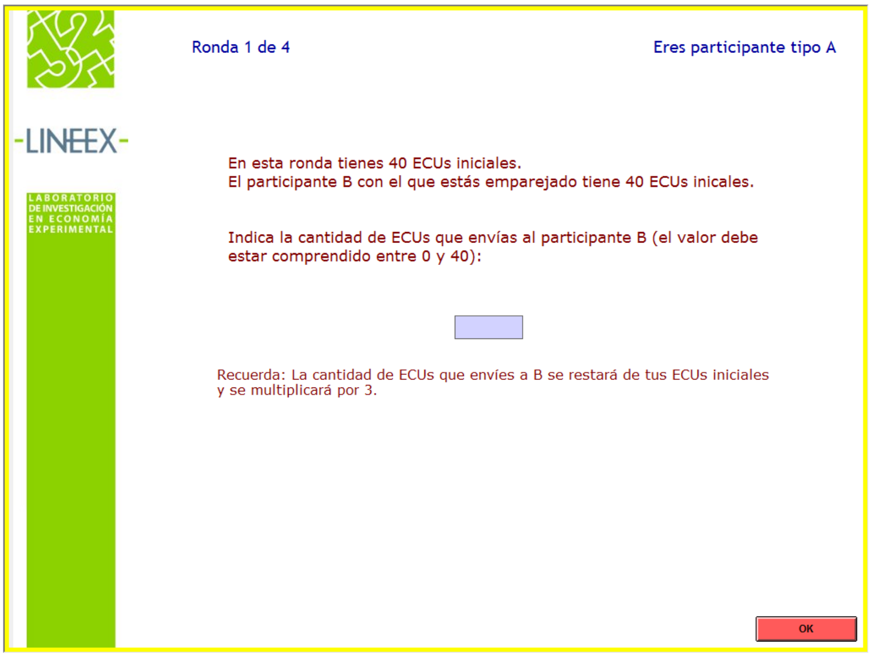
Final payoff player A = Provisional payoff player A – points sent by A

Final payoff player B = Provisional payoff player B– 3*points sent by A

To check that you have understood the instructions, we ask you to look at the computer screen. First, you will see the logic of the experiment through a numerical example. Next, you will need to compute the final payoffs of an example in which in which the computer chooses numbers randomly the ECUs send by player A and the ECUs returned by player B.

SCREENSHOTS

I. The investor's Behavior: Trust



Ronda 1 de 4

Eres participante tipo A

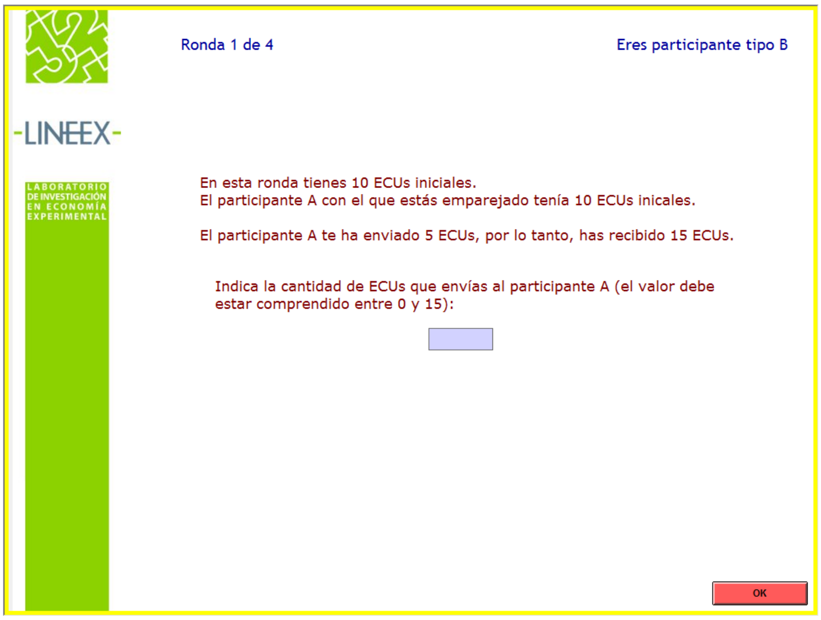
En esta ronda tienes 40 ECUs iniciales.
El participante B con el que estás emparejado tiene 40 ECUs inicales.

Indica la cantidad de ECUs que envías al participante B (el valor debe estar comprendido entre 0 y 40):

Recuerda: La cantidad de ECUs que envíes a B se restará de tus ECUs iniciales y se multiplicará por 3.

Investors were informed on this screen: "In this round you have 40 ECUS. The player you are matched with has 40 ECUS". Then, investors had to "Indicate the amount of ECUs to send to player B (the amount must be between 0 and 40)". Investors chose the desired transfer using the blue box. The text below the box reminds subjects that "the amount that you send will be reduced from your initial ECUs and multiplied by the 3"

II. The allocator's behavior: Trustworthiness



Ronda 1 de 4

Eres participante tipo B

En esta ronda tienes 10 ECUs iniciales.
El participante A con el que estás emparejado tenía 10 ECUs inicales.

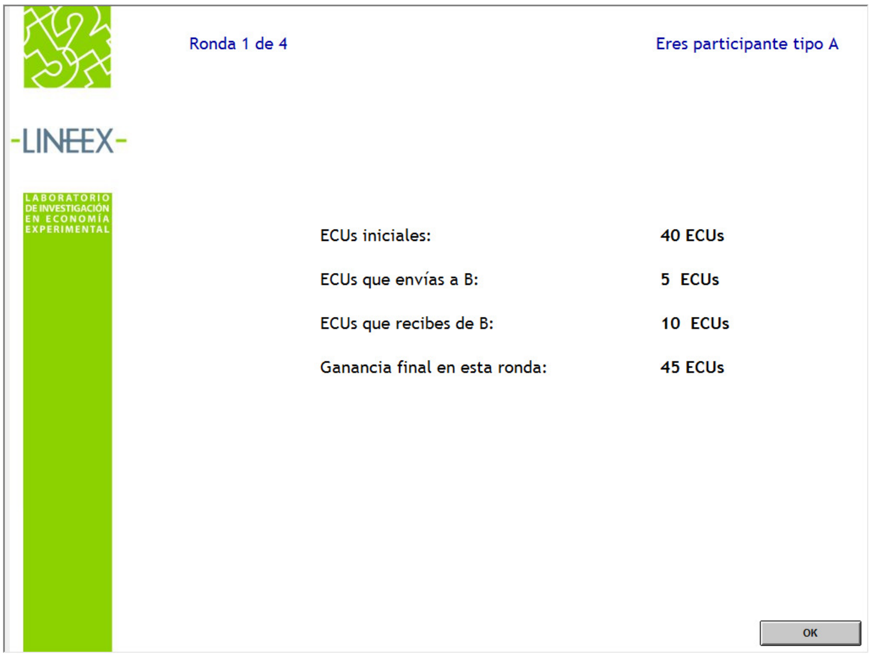
El participante A te ha enviado 5 ECUs, por lo tanto, has recibido 15 ECUs.

Indica la cantidad de ECUs que envías al participante A (el valor debe estar comprendido entre 0 y 15):

Investors were informed on this screen about the initial endowments (as explained for the case of investors). In the third line, the text states: “Player A sent you 5 ECUs, therefore you have received 15 ECUs. Indicate the amount that you want to send to player A (the amount must be between 0 and 15)”.

III. Earnings

The screenshot below informed player A about their initial ECUs, the amount sent to B, the ECUs received and the final earnings in that round. Player B faced a similar screen.



Ronda 1 de 4

Eres participante tipo A

ECUs iniciales:	40 ECUs
ECUs que envías a B:	5 ECUs
ECUs que recibes de B:	10 ECUs
Ganancia final en esta ronda:	45 ECUs

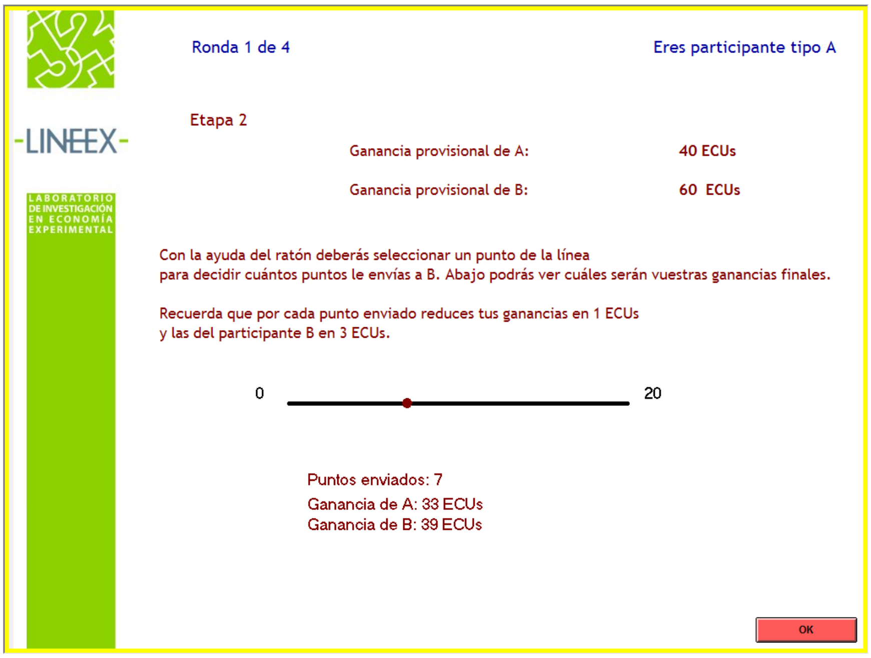
We decided to inform subjects about their earnings at the end of each round because in the punishment treatment, this information must be available for investors to decide whether to punish or not. With our design, we wanted to avoid that subjects received more information (feedback) in the treatment with punishment.

IV. The investor's behavior: Punishment

In the punishment treatment subjects were first informed about the amount that they had earned during the trust game (i.e., before the punishment-phase was played).

The screen was very similar to the one in section III, with the exception that the last sentence concerned “provisional earnings in that round” (instead of “final earnings in this round”)

Once subjects receive this information, investors were allowed to send “points” to allocators, as it is shown below:



Ronda 1 de 4

Eres participante tipo A

Etapa 2

Ganancia provisional de A:	40 ECUs
Ganancia provisional de B:	60 ECUs

Con la ayuda del ratón deberás seleccionar un punto de la línea
para decidir cuántos puntos le envías a B. Abajo podrás ver cuáles serán vuestras ganancias finales.

Recuerda que por cada punto enviado reduces tus ganancias en 1 ECUs
y las del participante B en 3 ECUs.

Puntos enviados: 7
Ganancia de A: 33 ECUs
Ganancia de B: 39 ECUs

To facilitate the computation of the final payoffs, the investor decided the points to be sent to the allocator using an slider bar that ranged from 0 to P*, as explained in the main text of the paper. By moving the bar, the investor received information about the final distribution of payoffs associated to her choice. The investor could move the sliding bar as many times as she wanted; her decision had to be confirmed by clicking the button “ok” at the bottom of the screen.

APPENDIX B

This appendix provides a more formal definition for the capacity of punishment and presents the partial derivatives of the capacity of punishment with respect to the level of endowments. We also provide further results on the econometric analysis. The investor' s behavior is analyzed in Section I (Table 1.B, Table 2.B, Table 3.B), and the allocator's behavior in Section II (Table 4.B, Table 5.B, Table 6.B). Results on efficiency are reported in Section III (Table 7.B).

Capacity of Punishment

Equation (6) in the Section 2 of the paper defines the maximum the share of the allocator's interim payoffs that the investor can destroy:

$$(6)\,\frac{\lambda\cdot\pi^a}{\pi^b}=\frac{\lambda\cdot(e_a-X+Y)}{(e_b+3X-Y)}$$

In our analysis, we shall use the investment ratio ($x := X/e_a$) as the measure of trust (see Smith, 2011; Johnson and Mislin, 2011), and the return ration (y := Y/X) as the measure for trustworthiness. If we rewrite equation (6) in terms of these variables we may obtain a formal definition for the *capacity of punishment* (CP).

Definition. *The* capacity of punishment *refers to the maximum share of allocator's interim payoff* π^b *that the investor can destroy after she trusts by sending a proportion x of her endowment and receives back a return ratio y from the allocator.*

$$(7)\,CP(e,\lambda;x,y)=\frac{\lambda\cdot\pi^a}{\pi^b}=\frac{\lambda\cdot(1-x+yx)}{(e_b/e_a+3x-yx)}$$

Notice that the inverse of the capacity of punishment captures how costly is for the investor to destroy the allocator's payoff completely; i.e., the value of ($\pi^b/\lambda\pi^a$) determines the share of the interim payoffs π^a that the investor would need to make $\pi^b=0$. In that vein, our measure for the capacity of punishment can be related to its cost and credibility. When the capacity of punishment is high, the investor can destroy the allocator's payoff with a small share of her own payoff, therefore the threat of punishment is much more credible.

Our formula for the capacity of punishment shows that the level of endowments (e_a and e_b), the level of trust (x) and the return ratio (y) are important variables at stake. By simply taking derivatives we can see that the investor will reduce the credibility of her punishment by trusting, but a higher return ratio will make *cheaper* for her to destroy the allocator's payoff completely. For any given (*x,y*) what crucially determines the capacity of punishment is the level of endowments.

Lemma. *Consider two distributions of endowments* $e = (e_a, e_b)$ *and* $e' = (e'_a, e'_b)$. *If the level of trust (x) and the level return ratio (y) are the same in both cases then:*

$$CP(e, \lambda; x, y) \gtreqless CP(e', \lambda; x, y) \text{ if and only if } (e_a/e_b) \gtreqless (e'_a/e'_b)$$

This lemma allows us to rank different capacities of punishment depending on the level of endowments. To show the result, we simply take derivatives in equation (7).

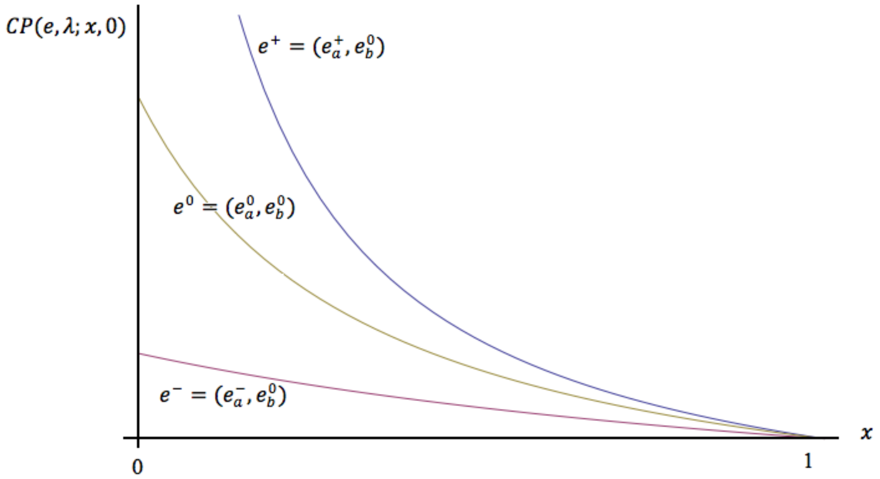
$$\frac{\partial CP(e, \lambda; x, y)}{\partial e_a} = \frac{-\lambda\ (1 + (y-1)x)}{e_b\ (e_b/e_a + 3x - yx)^2} > 0$$

$$\frac{\partial CP(e, \lambda; x, y)}{\partial e_b} = \frac{e_a\ \lambda\ (1 + (y-1)x)}{{e_b}^2\ (e_b/e_a + 3x - yx)^2} < 0$$

We then observe that the investor's capacity of punishment increases (decreases) with the investor's (allocator's) endowment, certeris paribus.

To illustrate this graphically, consider the worst possible scenario for the investor in which she trusts sending x but receives nothing back form the allocator (y = 0). The next figure plots the investors' capacity of punishment for each possible value of x in $[0, 1]$. We consider three different distributions of endowment satisfying $e_a^- < e_a^0 = e_b^0 < e_a^+$.

Figure. Capacity of punishment for different level of endowments



It is not difficult to see that for any level of trust x, it is always the case that the higher the value of e_a compared with e_b, the higher is the proportion of the allocator's endowment the investor can destroy.

Demographics and Data breakdown

	Investors	Allocators	Total
Women	0.54	0.56	0.55
	(0.50)	(0.50)	(0.50)
Age	22	21.94	21.97
	(2.57)	(3.37)	(2.99)
Trust	0.17	0.17	0.17
	(0.37)	(0.37)	(0.37)
N	48	48	96

The subject's gender is a dummy variable that takes the value 1 for women. The subject's age vary between 18 and 36 years. The GSS variable refers to the attitudinal survey question: "Generally speaking, would you say that most people can be trusted or that you cannot be careful in dealing with people" (1 if most people can be trusted, 0 otherwise). Standard errors in brackets.

I. The investor's Behavior

Table 1.B. Maximum-likelihood estimates of a random effects model that estimates the level of trust and controls for unobserved individual heterogeneity.

Dependent Variable: level of trust (X/e_a)				
	CP_{LOW}			CP_{HIGH}
	(10,40)	(40,40)	(10,10)	(40,10)
Period	-0.043	0.009	-0.045	0.014
	(0.049)	(0.020)	(0.040)	(0.047)
Net earnings t – 1	-0.008*	-0.004	0.010*	0.015*
	(0.005)	(0.005)	(0.006)	(0.009)
PUN	0.012	-0.006	-0.001	0.091**
	(0.058)	(0.030)	(0.055)	(0.040)
Women	-0.218***	-0.103**	-0.178*	-0.112
	(0.050)	(0.051)	(0.105)	(0.069)
Age	0.0003	8.27e-05	-0.015	0.006
	(0.012)	(0.010)	(0.013)	(0.011)
GSS	0.150	0.084**	0.225***	-0.006
	(0.092)	(0.042)	(0.080)	(0.082)
Constant	0.417	0.121	0.760***	-0.010
	(0.278)	(0.234)	(0.293)	(0.286)
σ_u	0.107	0.062	0.225	0.169
σ_e	0.301	0.157	0.237	0.182
ρ	0.113	0.134	0.474	0.464

Notes. The set of independent variables include the period, the net earning in the previous round, a dummy variable for possibility of punishment (PUN), and the data collected in the questionnaire regarding the investor’s gender, age and the answer to the attitudinal survey question from the General Social Survey (GSS): “Generally speaking, would you say that most people can be trusted or that you cannot be careful in dealing with people?” The robust standard errors take into account matching group clustering and are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2.B. We report the outcome of pairwise comparisons between the four different distributions using the Wilcoxon signed-rank in Panel A.[1] Hypothesis testing using the χ^2-test after estimating the econometric model in Table 1 in the main text are presented in Panel B. In both panels, we report the value of the statistics for the NOPUN and PUN treatment appear in the grey and the white area respectively.

A. Wilcoxon signed-rank test

	(10,40)	(40,40)	(10,10)	(40,10)
(10,40)	-	3.51***	0.89	3.04***
(40,40)	2.51**	-	2.29**	0.06
(10,10)	0.07	2.38**	-	3.01***
(40,10)	1.14	1.87*	0.63	-

B. χ^2-test after the random-effect model

	(10,40)	(40,40)	(10,10)	(40,10)
(10,40)	-	4.47**	0.01	6.36**
(40,40)	5.64***	-	6.21**	2.03
(10,10)	1.51	3.83*	-	10.6***
(40,10)	3.08*	1.91	0.41	-

The results confirm that without punishment, behavior is primarily driven by the *endowment effect* (Result 2). Trust is not significantly different when the endowment of the investor is low (10 ECUs) or high (40 ECUs), regardless of the endowment of the allocator. However, comparing trust when investors' endowment differs becomes significant. In the case with punishment, the same comparison yields a different result: the proportion of the endowment that the investor sends in (40,10) is statistically different from the behavior in (40,40), but it is not statistically different from the level of trust if the endowment is 10 ECUs.

[1] The results are robust when considering the t-test.

Table 3.B. We can test whether the amount that investors send in the NOPUN and the PUN treatment is the same regardless of the order in which these treatments are implemented. After the estimation of our model in Table 1, we test the null hypothesis H$_0$: $\alpha_{PUNFIRST} + \alpha_{PUNxPUNFIRST} = 0$ and H$_0$: $\alpha_{PUNxPUNFIRST} = 0$. The results of the χ^2-test are summarized below:

	Null Hypothesis	χ^2_1 (p-value)
The level of trust in the PUN treatment is the same when the game is played first or second in the session.	H$_0$: $\alpha_{PUNFIRST} + \alpha_{PUNxPUNFIRST} = 0$	7.32 (0.001)
The level of trust in the NOPUN treatment is the same when the game is played first or second in the session.	H$_0$: $\alpha_{PUNxPUNFIRST} = 0$	0.81 (0.368)

In the light of these results we can conclude that the highest level of trust is achieved when PUN is the first treatment to be implemented. The level of trust when there is NOPUN is not affected by the order of the treatments.

II. Allocators' Behavior

Table 4.B. Hypothesis testing for the effects of endowment heterogeneity on the return ratio using the χ^2-test after estimating the model in Table 2 (in the main text) –value of the statistics for the NOPUN treatment (grey area) and PUN treatment (white area) respectively.

	(10,40)	(40,40)	(10,10)	(40,10)
(10,40)	-	0.55	0.07	0.02
(40,40)	0.09	-	0.19	0.67
(10,10)	0.13	0.00	-	0.17
(40,10)	3.28*	2.57*	2.15	-

We find that the return ratio does not change within distributions in the NOPUN treatment (p-values > 0.46) so the allocator's behavior is roughly the same in that regard. In the PUN, the return ratio in (40,10) is significantly different (and actually smaller) than the return ratio in the distributions where the investor have a low capacity of punishment (10,40) and (40,40).

Table 6.B. We can test whether the return ratio in the NOPUN or the PUN treatment is the same regardless of the order in which treatments are implemented. After the estimation of our model in Table 2 (in the main text), we test the null hypothesis H_0: $\alpha_{PUNFIRST} + \alpha_{PUNxPUNFIRST} = 0$ and H_0: $\alpha_{PUNxPUNFIRST} = 0$. The results of the χ^2-test are summarized below:

	Null Hypothesis	χ^2_1 (p-value)
The return ratio in PUN is the same when the game is played first or second in the session.	H_0: $\alpha_{PUNFIRST} + \alpha_{PUNxPUNFIRST} = 0$	1.80 (0.180)
The return ratio in NOPUN is the same when the game is played first or second in the session.	H_0: $\alpha_{PUNxPUNFIRST} = 0$	1.25 (0.264)

In the light of these results we can conclude for any given treatment PUN or NOPUN, the return ratio is not affected by the order of the treatments (e.g., the return ratio with PUN is roughly the same when PUN is introduced after NOPUN and when PUN is the first treatment to be implemented).

III. Efficiency, final payoffs and punishment behavior

Table 6.B. Final (average) payoffs of investors and allocators in each distribution with and without punishment. We report the p-values for the Wilcoxon-test.

	Distribution			
	CP$_{\text{LOW}}$ **(10,40)**	**(40,40)**	**(10,10)**	CP$_{\text{HIGH}}$ **(40,10)**
Investor's payoffs				
■ NOPUN	9.000	36.958	9.270	39.375
□ PUN	8.375	36.146	10.021	35.687
t-test (t)	0.418	0.652	0.187	0.062
Wilcoxon test (Z)	0.609	0.756	0.567	0.042
Allocator's payoffs				
■ NOPUN	46.333	55.167	15.604	19.583
□ PUN	40.875	43.687	10.625	21.750
t-test	0.018	0.033	0.001	0.568
Wilcoxon test (Z)	0.009	0.011	0.000	0.350

Our data suggest that investors do not send a higher proportion of the endowment to allocators except if the capacity of punishment is high (Result 1 in the main text). We have also found that the return ratio is higher with punishment, except when the capacity of punishment is high (Result 5 in the main text). Our results in Table 1D are consistent with these findings. If we look at the allocator's payoffs, for example, we can see that investors are better off in the absence of punishment, except if the capacity of punishment is high. This result can be explained because investors are not more willing to transfer money with punishment in (10,10), (10,40) and (40,40), but allocators are less likely to reciprocate in these distributions (i.e., the return ratio is higher with punishment). Besides, the punishment destroys part of their endowment so that allocators would prefer the situation without punishment. The investor's problem is a little bit different. If they do not have a high capacity of punishment, they do not send more money to allocators, but they receive more money back. This would be beneficial for them by increasing their payoffs. However, allocators use the punishment and end up with a payoff that is similar to the one without punishment.

Table 7.B. Punishing behavior: Spearman correlation coefficients

	Decision to punish (Yes/No)	Amount punish (P)	Relative punish ($P/\overline{\pi}^a$)
Return ratio (y = Y/X)	-0.270***	-0.298***	-0.397***
Reciprocity ($Y - X$)	-0.249***	-0.289***	-0.402***
Endowments ($e_a^H - e_b^H$)	0.181	0.143	0.040
Number of obs.	117	117	117

As indicated by the first column, the willingness to punish is affected by the return ratio. In particular, the larger the return ratio ($y = Y/X$), the less likely investors are to punish. Along these lines, investors devote less resources to punish (in terms of the amount of punishment inflicted and the proportion of the interim payoffs they used to punish) the larger the return ratio is. These findings are consistent if we focus instead on the difference between Y and X, which can also be used to measure reciprocity.[2] In line with our findings in the manuscript, the endowment heterogeneity does not seem to affect punishing behavior.

[2] When we investigate punishing behavior and relate it to the level of trustworthiness (Y) the results are also clear-cut. Investors punish less frequently and devote a smaller proportion of their interim payoffs to punish the higher the level of trustworthiness is (r = -0.29, p-value=0.001 and r = -0.48, p-value=0.000, respectively).