Discussion Papers in Economic Behaviour

DPEB
08/13

Carry a big stick, or no stick at all

VNIVERSITAT DE VALÈNCIA

>Vicente Calabuig
Universitat de València and ERI-CES, Spain

>Enrique Fatas
University of East Anglia

>Gonzalo Olcina
Universitat de València and ERI-CES, Spain

>Ismael Rodriguez-Lara
Universitat de València and ERI-CES, Spain

August, 2013

www.erices.es eri-ces

Carry a big stick, or no stick at all

An experimental analysis of trust and capacity of punishment

Vicente Calabuig
ERICES, Universidad de Valencia

Enrique Fatas
University of East Anglia

Gonzalo Olcina
ERICES, Universidad de Valencia

Ismael Rodriguez-Lara*
ERICES, Universidad de Valencia
LUISS Guido Carli University

Abstract

We investigate the effect of punishment in a trust game with endowment heterogeneity in which the investor may punish the allocator at a cost. Our results indicate that the effect of the punishment crucially depends on the investor's capacity of punishment, that is measured in our experiment by the proportion of the allocator's payoffs that the investor can destroy. We find that punishment fosters trust when the capacity of punishment is high (i.e., when the cost of punishing is relatively low). Otherwise, punishment fails to promote trusting behavior, crowding out intrinsic motivation to trust. Trustworthiness is higher with punishment than without punishment, except if investors have a high capacity of punishment.

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

JEL Codes: C91, D02, D03, D69.

* Corresponding author: Ismael Rodriguez-Lara, Universidad de Valencia, Dpto. Análisis Económico, Campus dels Tarongers, Avd. Naranjos s/n, 46022 Valencia. Phone:(+34)965828211. Email: Ismael.Rodriguez@uv.es

1. Introduction

Trust and trustworthiness play an essential role in strategic environments involving incomplete contracts. Examples include the hold-up problem, 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) or the principal-agent relationship. A crucial question to be addressed in all these asymmetric environments is how to increase the level of trust and trustworthiness as these factors indeed determine the level of efficiency of the interaction.[1]

Reward and punishment institutions might help to foster trust and trustworthiness. By using carrots (rewards) and sticks (punishment), principals might foster cooperation of their employees and buyers can favor the provision of high quality products by sellers. 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). Laboratory experiments do also suggest that individually costly sanctions significantly increase contribution to public goods games (Chaudhuri, 2011), even between strangers (Fehr and Gächter, 2002). However, the role of punishment on trust and trustworthiness has been studied to a much lesser extent. In fact, it is unclear whether punishment will pursue the desired outcome as incentives may crowd out intrinsic motivation to trust and reciprocate.[2]

In this paper we investigate the effect of punishment on trust and trustworthiness in a controlled laboratory experiment. We introduce a punishment-phase in the trust game (as first analyzed by Berg et al., 1995) by giving the investor the possibility of punishing the allocator's behavior (at a cost). In our experiment, the effectiveness of the punishment (i.e., the factor by which punishment reduces the allocator's payoff) is kept constant.[3] We focus instead on the capacity of punishment that is determined by the endowment heterogeneity. This feature has been disregarded by the literature and refers to the proportion of the allocator's payoff that the investor can destroy after trusting, if the allocator decides to return nothing. One example in which this capacity of punishment operates is to consider that you want to build a new house. You pay the builder but the outcome is unsatisfactory. An unlimited budget will always allow you to hire a good lawyer to sue the builder. If your resources are scarce, however, the magnitude of your resources, relative to the builder's company size, will critically

[1] A large number of studies show that tendencies to trust and reciprocate affect economic growth and yield better functioning governments (see, among others, Arrow, 1974; Knack and Keefer, 1997). There exists also evidence that the level of trust is an important determinant of economic performance (Bachmann and Zaheer, 2006; Gambetta, 1988).

[2] In psychology, Deci, Koestner and Ryan (1999) review the literature on the role of incentives and discuss the emergence of crowing-out effects. Some relevant papers in economics include Frey (1997) and Gneezy and Rustichini (2000a, 2000b). For a recent survey see Gneezy et al. (2011).

[3] Nikiforakis and Normann (2008) and Rigdon (2009) investigate how the effectiveness of punishment affects decisions in a public good game and a trust game respectively.

determine the extent to which you can punish the builder. In other words, endowment heterogeneity (rather the mere effectiveness of the punishment) becomes the main determinant of the damage you can inflict to the builder.

In our within-subject design, investors and allocators receive a low or high endowment, depending on the round and play the trust game with and without the punishment-phase. The level of trust is defined as the proportion of the endowment that the investor decides to transfer. We measure the level of trustworthiness by the return ratio, which refers to the proportion of the received amount that the allocator returns. As we manipulate the endowment of both the investor and the allocator, we systematically explore how trust and trustworthiness are affected by the *capacity of punishment* when the possibility of punishment is at stake. A high capacity of punishment (i.e., having a big stick) implies that the investor can destroy a high proportion of the allocator's payoffs if nothing is received back. The opposite occurs when the capacity of punishment is low. In that vein, what crucially determines the capacity of punishment is the relation between the investor and the allocator's initial endowments.

Our experimental data suggest that the capacity of punishment is a key element to explain the effect of punishment on trust and trustworthiness. A high capacity of punishment significantly increases trust because it prevents any crowding-out effect when we introduce the possibility of punishment in the game. Although we find that a high capacity of punishment is beneficial for the level of trust, our results also suggest that it fails to increase the relative level of trustworthiness (i.e., the return ratio). In particular, our data provide evidence that allocators are willing to return a higher share of the generated surplus to the investors if there is punishment with only one exception: the distribution in which the investor has a high capacity of punishment. This occurs even when allocators received a relatively larger share of the investor's endowment.

The rest of the paper is organized as follows. Next we discuss the related literature and explain what makes our model divert from previous research that introduce the punishment-phase in the trust game. In Section 2, we describe our trust game with punishment and provide a formal definition of the capacity of punishment. We discuss the experimental design in Section 3. The results are summarized in Section 4. Section 5 presents final remarks and relates our findings with previous research in other environments, where the effects of incentives on individual behavior are analyzed.

Related literature

We will not attempt to review the large and growing experimental literature on the trust game. Camerer (2003), Cooper and Kagel (2009) and Eckel and Wilson (2011) provide excellent surveys that the reader may wish to consult. Instead, we focus our attention on some recent papers that are particularly relevant to our study as these articles incorporate a punishment-phase in the bilateral trust game of Berg et al. (1995). In

particular, we discuss previous modifications in the trust game in which investors are given the possibility of punishing allocators.[4]

The seminal article of Fehr and Rockenbach (2003) focuses on the effect of punishment on the levels of trustworthiness. They find evidence for a crowding-out effect since allocators are less likely to reciprocate after facing threats from investors. More precisely, 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.[5] Because choosing or not to fine might have a *signaling* meaning, Houser et al. (2008) aim to disentangle allocators' intentions and incentives by comparing threats imposed by investors and by nature. Again, trustworthiness is higher when allocators are not threatened with sanctions. The allocators' behavior depends also on how large the requested amount is relative to the sanction and this result does not vary with investors' intentions.

Although these findings suggest that punishment might have a detrimental effect on trustworthiness, its effect on trust is not analyzed in deep. In addition, the imposition of the fine is supposed to affect only the allocator's payoff, being costless for investors. In Rigdon (2009) the possibility of punishment is exogenously given by the experimenter, but costly for the investor. Investors may send up to half of their endowment to the allocator, and the rest of their endowment can be used to punish the allocator if she does not fulfill the requested amount. Rigdon (2009) analyzes how the *effectiveness* of punishment affects the subjects' decisions and find that punishment increases the level of trust only if it is very effective (or relatively cheap).[6] Our results are in line with this view, but in our setting the effectiveness of the punishment is kept constant and investors' capacity of punishment is crucially affected by the difference between the investor and the allocator's level of endowments. This issue is disregarded from her analysis.

A final noteworthy aspect of our game detailed below relies on the absence of a desired payback. In our view, this is important to understand the role of incentives when contracts are incomplete. When the original trust game is modified to allow investors to choose a desired-payback, other psychological considerations such as guilt aversion might play a role (see Charness and Dufwenberg 2006; Battigalli and Dufwenberg, 2007). By eliminating the desired-payback, we introduce uncertainty about the investor's expectations and let punishment occur regardless of whether agents are aware of the desired return of the transaction. Note that while in previous research investors were allowed to punish only when the return did not match the requested payback, in

[4] The effect of third-party punishment is studied in Bohnet et al. (2001) or Charness et al. (2008), among others.

[5] A larger desired payback 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.

[6] In her low-punishment treatment, 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. Rigdon (2009) finds higher levels of trust only in the high-punishment treatment.

our game allocators can be punished regardless of the amount they return to investors, what perfectly fits into the idea of an incomplete contract.

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 $e_b \geq 0$. Any amount X in $[0, e_a]$ that the investor sends to the allocator is tripled by the experimenter, 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:λ. 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) $CP(e,\lambda;X,Y) = \frac{\lambda \cdot \pi^a}{\pi^b} = \frac{\lambda \cdot (e_a - X + Y)}{(e_b + 3X - Y)}$

where $CP(e,\lambda;X,Y) \geq 1$ if the investor can destroy the allocator's payoffs completely (i.e., $P' \geq P^*$).

Since investors might differ in their endowment and in order to make meaningful comparisons on their behavior, we refer hereafter to the level of trust (x) 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). If we rewrite equation (6) in terms of these variables we obtain our expression for what we call the *capacity of punishment* (CP).

Definition. *The* capacity of punishment *refers to the 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.*

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

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 devote for making $\pi^b=0$. 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.

Some papers that investigate the efficiency of the punishment highlight the importance of its effectiveness to achieve the desired outcomes (e.g., Nikiforakis and Normann 2008, Rigdon 2009). It is clear from equation (7) that an increase in the effectiveness of punishment (λ) allows the investor to destroy a higher proportion of the allocator's payoffs, making the punishment more credible. In addition, 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 capacity of punishment decreases (increases) in the level of trust (the level of trustworthiness). The investor will then 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:*

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

This lemma allows us to rank different capacities of punishment depending on the level of endowments. To illustrate how the capacity of punishment depends on $e=(e_a,e_b)$ we rely on the worst possible scenario for the investor in which she trusts sending x but receives nothing back form the allocator (y = 0). Figure 1 depicts the investors' capacity of punishment for each possible value of x in $[0,1]$. We consider three different distributions of endowment satisfying $e_a^-<e_a^0=e_b^0<e_a^+$.

Graph here?

For any level of trust, it is always the case that the investor can destroy a higher proportion of the allocator's endowment, the higher (smaller) the value of e_a (e_b). In our experiment, we vary e_a and e_b so that investors are given different capacities of punishment. Our paper is an attempt to investigate how this capacity affects the subjects' decision in the game.

In our context, the capacity of punishment is also affected in our context by the level of trust (e.g., and trustworthiness (e.g., Importantly, these decisions can interact with the level of endowments in our context.

4.4 A comment on the endogenous capacity of punishment.

In all the previous behavioral models, the credibility of the punishment (i.e., the size of the stick) is exogenously determined. In our context, the capacity of punishment depends on the level of endowments but it is also affected by endogenous decisions. This, in our view, might affect the level of trust and trustworthiness.

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.[7] This reduction in the level of trust should be more severe when the capacity of punishment is low. By the same token, the allocator may want to return less, so as to affect the investor's capacity of punishment. The effect in this case goes in the opposite direction as allocators will return less when the investor's capacity of punishment is high. For low capacities of punishment, the effect on trustworthiness should be smaller.

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), all of them 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*), kept constant through the session. Each subject went through a sequence of eight one-shot games with two different treatments: No Punishment (NOPUN) and Punishment (PUN). Each treatment consisted of four rounds

[7] In Rigdon (2009), investors can only send up to half of their endowment to allocators. The remaining amount can be kept by investors or devoted to punish. Thus, the capacity of the punishment does not depend so crucially on the level of trust.

in which subjects with different roles were matched in pairs. To control for order effects, subjects played either NOPUN or PUN first in half of the sessions.

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*). They interacted with each other using a perfect-stranger protocol within treatments, as they never made more than one decision with the same pair in the same treatment, and a partners protocol across treatments (NOPUN and PUN). Subjects from different matching groups never interacted with each other throughout the session.

Subjects received at the beginning of each treatment (PUN or NOPUN) a printed copy of the experiment instructions.[8] 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$\in\{a,b\}$. Subjects then played the four different distribution of endowments $(e_a, e_b) \in \{(10,10),(10,40),(40,10),(40,40)\}$ in each of the two treatments.[9] Subjects made decisions four times per treatment, one per distribution, in a very similar way. After knowing (e_a, e_b), 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. Subjects were informed about their payoffs in the round and were then re-matched.

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. Sending one point to the allocator cost 1 ECU to the investor and decreased allocator's payoff in λ = 3 ECUs.[10] 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.

At the end of the session one of the two treatments (PUN or NOPUN) was randomly selected to pay subjects. We paid a whole treatment rather than a round to avoid extreme variance in subjects' payoffs -e.g., if round 1 was selected to pay, we would had some subjects who played the distribution (10,10) and subjects who played

[8] A translated version of the instructions is available in Appendix A. This contains further details about the experimental design (including some screenshots).
[9] We control for the order in which distributions were played so that not all subjects went through the same sequence of decisions.
[10] Note that this corresponds to the high-punishment treatment in Rigdon (2009).

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.[11]

Subjects received on average 15 Euros.[12] 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

The central question our experiment addresses is whether the possibility of punishment affects the levels of trust and trustworthiness. By the same token, we want to analyze whether the capacity of punishment have some predictive power.

In our experiment, the effectiveness of punishment is kept constant ($\lambda = 3$) therefore the capacity of punishment crucially depends upon the level of endowments as follows:

$CP(40,10, \lambda; x,y) > CP(10,10, \lambda; x,y) = CP(40,40, \lambda; x,y) > CP(10,40, \lambda; x,y)$, for any given pair (x ,y) such that $x \geq 0$ and $y \geq 0$ and a fixed value of λ.

We refer hereafter to the distribution (40,10) as the one in which the investor has a high capacity of punishment, and (10,40) as the one in which the capacity of punishment is low.

Different models on behavior posit different predictions. The self-interest model, for example, assumes that subjects are exclusively motivated by their material payoff so that the levels of trust and trustworthiness will not depend at all on the capacity of punishment. More precisely, the level of trust and trustworthiness will be zero in all distributions and punishment will not occur. This outcome is clearly inefficient and contrasts with the observed behavior in many laboratory experiments. Even in the absence of punishment investors usually trust on allocators by sending part of their initial endowment and allocators then reciprocate this behavior being trustworthy (see Camerer, 2003; Cooper and Kagel, 2009; Eckel and Wilson, 2011 for a review of the results). In the light of this evidence, we see that behavioral models on social preferences and intrinsic motivation are much more promising. In what follows, we present the prediction of these models.

4.1 Models of social preferences.

[11] One might argue that paying one treatment instead of a round can be problematic because as long as there are differences in endowment between investors and allocators in a particular round, but not within treatments. We note, however, that subjects were not aware that all of them played the four different distributions (see the instructions).

[12] Exchange rate: 10 ECUs = 1 Euro.

Models of social preferences (e.g. Rabin 1993, Fehr and Schmidt 1999, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006) 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. Importantly, the presence of a significant fraction of punishers in the population might induce changes on the behavior of selfish individuals (Fehr and Gachter 2000, Chaudhuri 2011). As the punishment is more effective (see Nikiforakis and Normann 2008, Rigdon 2009) the impact of punishers on aggregate behavior is greater; i.e., the levels of cooperation in the population increase.

If the models of social preferences apply in our setting, the prediction would be that the levels of trust and trustworthiness increase monotonically with the capacity of punishment.[13]

Hypothesis 1 (*Social Preferences*). a) The level of trust and trustworthiness is higher with punishment than without punishment. b) Once we allow for punishment, the levels of trust and trustworthiness increase monotonically with the capacity of punishment.

4.2 Models of intrinsic motivation: the hidden cost of punishment

Fehr and Rockenbach (2003) show experimentally that sanctioning allocators may crowd out voluntary reciprocation. This crowding out effect of extrinsic motivation (punishment) on intrinsic motivation has been reported in other field and laboratory experiments in different settings (see, among others, Deci et al. 1999, Frey 1997, Gneezy and Rustichini 2000b, Gneezy et al. 2011).

In the trust game, intrinsic motivation is typically related to the investor's willingness to trust and the allocator's willingness to return. While a direct *price-effect* may alter subjects' extrinsic motivation, making investors (allocators) more willing to trust (reciprocate), incentives may also have an indirect and opposite psychological effect on this intrinsic motivation, *crowding out* the incentivized behavior. In particular, the investor's intrinsic motivation to trust might be reduced when punishment is allowed because trusting losses its *image* or *reputation value* (see Benabou and Tirole 2003, Gneezy et al. 2011). The investor gives money away, not because she is nice and trusts, but because she can punish the allocator if he does not pay back. As a result, the investor's intrinsic motivation would be decreasing on his capacity of punishment. But this crowding out effect on intrinsic motivation also operates for the allocator and his willingness to reciprocate. Explicit incentives might signal distrust in a principal-agent relationship (Kosfeld and Falk 2006; Ellingsen and Johanhesson 2008). In our trust

[13] In a dynamic theoretical model, Olcina and Calabuig (2013a, 2013b) show that the levels of trust and trustworthiness increase with the capacity of punishment, although this capacity is exogenous in their setting (e.g., the maximum proportion of the allocators' payoff that investors can destroy is determined by the legal system of the country).

game allocators can perceive as an unfriendly the possibility of punishment. This negative perception will be increasing in the capacity of punishment of the investor and can potentially have detrimental effects on behavior. The possibility of punishment might produce a shift from positive reciprocity to negative reciprocity. It becomes unclear for the allocator whether the investor is trusting him because she is being nice or because she can punish him if case of an unsatisfactory outcome.

Obviously extrinsic motivation to give and to pay back will be increasing on the capacity of punishment both for the investor and the allocator. As it is suggested in Gneezy et al. (2011), this price or incentive effect will dominate if incentives are high enough (i.e., if the stick is sufficiently big). Otherwise, punishment may crowd-out intrinsic incentives to give and pay back.

Hypothesis 2 (*Intrinsic and extrinsic motivation*). a) Whereas punishment will affect extrinsic motivation, increasing the level of trust and trustworthiness, there may exist a crowding-out effect that dumper trust and trustworthiness, b) The incentive effect dominates the crowding-out effect as the capacity of punishment increases.

Thus, models on intrinsic and extrinsic motivation will not necessarily predict that punishment fosters trust and trustworthiness. In particular, the relationship between the capacity of punishment and the levels of trust and trustworthiness might be non-monotonic, although the punishment should have a positive effect if the capacity of punishment is sufficiently high.

4.3. A comment on the order effect

Our within-subject design outlined in Section 3 makes subjects play with and without punishment in each of he sessions. Since we control for the order in which treatments are implemented, we can analyze how subjects react when incentives are introduced and removed. One important insight from the literature on intrinsic and extrinsic motivation described in Section 4.2 is that an increase in extrinsic motivation 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 idea is that once the incentives are removed the intrinsic motivation will remain in the new and low levels, so that the behavior may be affected not only by current incentives but also by the incentives offered previously (see Gneezy and Rustichini 2000b). In other words, "*Once incentives are removed, people may pursue the desired outcome less eagerly.*" (page 192, Gneezy et al. 2011). This translates into considering that removing punishment will affect the investor's behavior if the capacity of punishment is high, so that the level of trust can decrease specially in this case.

Hypothesis 3 (*Lasting effect on Intrinsic motivation*)**.** When punishment is removed, the level of trust and trustworthiness will decrease, especially if the investors' capacity of punishment is high.

Notice that a model of social preferences will also predict a decrease of the levels of trust when punishment is removed but for all capacities of punishment.

5. Results

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

5.1. The Investor's Behavior

5.1.1. The effects of punishment and endowment heterogeneity on the level of trust

We begin by examining the effect of punishment on trust. We present an overview of our data in Figure 2, where we plot the distribution of trust in both treatments by considering each possible distribution separately.[14] 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 2 around here]

We can see in Figure 1 that the distribution of trust is roughly the same with and without punishment, unless the investor holds a high capacity of punishment -the average level of trust nearly doubles that case (11.1% versus 19.9%), while keeping roughly the same in the rest of distributions. 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.[15]

Result 1. (Effect of punishment on the level of trust) *Punishment fosters the level of trust only if the investor's capacity of punishment is high. Otherwise, the level of trust is not affected.*

Another insight that can be gleaned from Figure 2 concerns the investor's behavior across distributions within each treatment. If there is no punishment, the level of trust is

[14] In what follows, distributions are ordered according to the capacity of punishment. We assume in this section that investors consider the worst possible scenario in which they trust by sending x but think that will receive nothing back from the allocator (y=0). Given the level of trust that is observed in our data, we can then say that CP(40,40) < CP(10,10).

[15] Additional results supporting this evidence are presented in our econometric analysis of Section 5.1.3 and Table 1.B in the appendix, where we consider a random effect specification to control for unobserved individual heterogeneity.

smaller if the investor's endowment is high (40 ECUs) compared with the case in which her endowment is low (10 ECUs). This result is in line with previous findings suggesting that the proportion of the endowment sent by investors decreases as the size of her endowment increases (e.g., Johansson-Stenman et al., 2005; Johnson and Mislin, 2011). One interesting finding, however, is that the *endowment effect* vanishes with a high capacity of punishment. If there is punishment, the level of trust in the (40,10) distribution is roughly the same (19.95%) as the level of trust in distributions where the investor's endowment is low (21.81% and 25.83%). Still, the level of trust in the (40,40) distribution is significantly smaller than in the rest of the distributions.[16]

Result 2. (Endowment effect and capacity of punishment) *If there is no punishment, the level of trust decreases as the level of the endowment increases (endowment effect). If there is punishment, a high capacity of punishment compensates for this endowment effect.*

5.1.2. Order effect: Trust and crowding-out

Models of social preferences predict that punishment will foster trust, regardless of the capacity of punishment and the order in which treatments are implemented. Result 1 provides some evidence against these models and highlights that intrinsic and extrinsic motivation might play an important role in determining the effect of punishment on the level of trust. In our experimental design, subjects played either with or without punishment in the first part of the session. Figure 3 depicts the percentage changes in the level of trust for each distribution, both when the punishment is introduced in the second part of the experiment (solid line) and when it is removed (dotted line). For the sake of completeness, we present the effect of the punishment in the pooled data (bars).

[Figure 3 around here]

According to models of social preferences, introducing the possibility of punishment should be beneficial for trust. Figure 3, however, suggests that punishment has a detrimental effect on the level of trust (*crowding-out effect*), except if the capacity of punishment is high –this is the only case in which punishment does not significantly change the proportion sent by investors ($t = 0.616$, p-value = 0.544). Although the extinction of the institution should be detrimental for the level of trust, we find that the effect is only significant in distribution (40,10), in which trust decreases by 70%, going from 26.9% to 7.7% ($t = 3.255$, p-value = 0.003). In the rest of the distributions, changes on trusting behavior are around 25% but not significant (p-values > 0.208), so that we find evidence in favor of our Hypothesis 3.[17]

[16] Table 2.B. in the appendix reports the p-values of pairwise comparisons between the four different distributions using the Wilcoxon signed-rank test. The results presented in the appendix are robust when considering the t-test.

[17] Note that the solid and the dotted line and in the negative 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

Result 3. (Order effect and trust)

a) *The introduction of punishment crowd-outs the level of trust by decreasing the relative amount sent, unless there exists a high capacity of punishment.*
b) *The extinction of punishment is detrimental for the level of trust when the capacity of punishment is high. [and trust is high]*

In a nutshell, Results 1, 2 and 3 provide evidence against models of social preferences (Hypothesis 1), and seem to support models on intrinsic and extrinsic motivation (Hypothesis 2 and 3). The econometric analysis presented in Section 5.1.3 is in line with this view and suggest that men, and those who think that majority of people can be trusted are more likely to trust.

5.1.3. Regression analysis and behavioral determinants of trust

To disentangle the effects of punishment and endowment heterogeneity on the level of trust, Table 1 presents the maximum-likelihood estimates of a random effects model that controls for unobserved individual heterogeneity. In our specification, the dependent variable is the level of trust -i.e., the proportion of the endowment that the investor sends to the allocator. The set of independent variables include the period in which the decision is made (PERIOD), the investor's earnings in the previous round (PREVEARN) 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).

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. 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

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, where we also argue that the endowment effect and the importance of the high capacity of punishment to compensate it (Result 2) occur regardless of the order in which treatments are implemented.

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 = 0.0736 and 0.0316, 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 = 0.6519). These findings support models on intrinsic and extrinsic motivation (Hypothesis 2) rather than models of social preferences (Hypothesis 1). In line with Hypothesis 3, the χ^2-tests suggest that removing the punishment is detrimental for the level of trust if the capacity of punishment is high (p-value = 0.001), but not otherwise (p-values > 0.179).

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 = 0.0347). These results confirm Results 1 and 3 and provide further evidence against models of social preferences.[18]

Finally, we find that previous 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. 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.[19]

Result 4. (Demographics and trust) *Women trust less than men in all possible distributions. Similarly, those who think that majority of people can be trusted trust more.*

5.2. The Allocator's Behavior: Trustworthiness

5.2.1. The effects of punishment on the level of trustworthiness

We examine in this section how the allocator reciprocates by analyzing the return ratio (y), which is smaller than 1 (larger than 1) when the allocator gives back to the

[18] 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 2 regarding the existence of an endowment effect that is compensated by the capacity of punishment when investors are allowed to punish (see Table 2.B in the Appendix).

[19] Note that Table 1 suggests a significant effect of gender and trust in the baseline distribution. Additional regressions suggest something similar happens in all distributions, with no gender differences in the probability of giving (e.g., see Table 1.B and 3.B in the appendix). We do not discuss these results in detail because we focus on the effect of punishment on the level of trust and trustworthiness. Further results on the determinants of trust and trustworthiness are analyzed in a companion paper.

investor less than (more than) what he has received. Figure 4 plots the distribution of the return ratio grouping the cases in which it is less than 1, equal to 1 and greater than 1. We report the average value of the return ratio and the standard deviation (in brackets) in the table below the figure. The table includes the results of the t-test and the Wilcoxon signed-rank that compare allocators' behavior in each distribution with and without punishment.[20]

[Figure 4 around here]

Figure 4 provides some interesting insights into the allocator's behavior. The proportion of allocators who transfer less than (more than) what they received from their paired investor decreases (increases) with the possibility of punishment, except when there exists a high capacity of punishment. In that case, distributions are roughly the same in both treatments. The descriptive statistics highlight that the average return ratio is always higher with punishment, with the smallest difference in the distribution (40,10) in which the return ratio is similar in both treatments (0.66 versus 0.71).

Result 5. (Effect of punishment on the level of trustworthiness) *Punishment increases the return ratio, except if the investor has a high capacity of punishment.*

This is a striking result. Allocators were expected to reciprocate more, especially if the capacity of punishment is high. Surprisingly, this is not the case. Punishment fosters trustworthiness except in the (40,10) distribution. Moreover, we find that the return ratio decreases with the capacity of punishment when there is punishment, going from 1.266 in the distribution (10,40) to 0.711 in the distribution (40,10).

One might argue that allocators are inequity averse (e.g., Fehr and Schmidt 1999). They might want to equalize payoffs being less likely to reciprocate when their endowment is lower than the investor's one. If this were the case, the same pattern should be observed in the absence of punishment. However, allocators return roughly 72% of what they receive without punishment and no significant differences exists when comparing behavior across distributions.[21]

Result 6. (Allocator's behavior across distributions) *If there is no punishment, the return ratio does not change within distributions. If there is punishment, the return ratio decreases as the capacity of punishment increases.*

Hypothesis 1 is then rejected for the case of trustworthiness. Models on intrinsic and extrinsic motivation suggest that the fear of punishment triggers extrinsic motivation to

[20] In our analysis, we drop all observations in which allocators receive no money from investors. This feature complicates the within-subject analysis (i.e., allocators who receive nothing cannot return any amount to investors). For that reason, we use a conservative t-test and Wilcoxon signed-ranks test: when testing the effect of the punishment in each distribution, we only consider those subjects who received a positive transfer in both treatments.

[21] This corresponds to giving back on average 24% of the generated surplus, what is consistent with the observed behavior in the meta-study of Johnson and Mislin (2011). Another result that is consistent with the literature (e.g., Camerer 2003; Cooper and Kagel 2009) is the fact that investors do not retrieve on average what they sent (i.e., the return ratio is smaller than one in all the distributions).

return, but the allocator's intrinsic motivation critically depends on the capacity of punishment. Although investors carrying a big stick send more, allocators might be less willing to reciprocate because the investor's behavior might be associated to her high capacity of punishment and not her intrinsic motivation.[22] According to Hypothesis 2 any crowding-out effect in the intrinsic motivation to return would be compensated by the price effect as the capacity of punishment increases. However, we do not find evidence for that hypothesis in the pooled data analysis. Next, we show that Hypothesis 3 that predicts a decrease in trustworthiness when punishment is removed -especially in the distribution (40,10)- is also rejected.

5.2.2. Order effect: Trustworthiness and endogenous capacity of punishment

To analyze how introducing and removing the possibility of punishment affects the allocator's behavior Figure 5 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 appears in the grey bars. As discussed in Section 5.2.1, this effect is decreasing in the capacity of punishment.

[Figure 5 around here]

Introducing punishment has a positive effect on trustworthiness that vanishes as the capacity of punishment increases. When punishment is removed, the return ratio goes down by roughly 40%, except if the capacity of punishment is high. In that case, the extinction of the punishment is beneficial for the return ratio, although the effect is not significant.

Result 7. (Order effects and trustworthiness)

a) *The introduction of punishment fosters the return ratio, unless there exists a high capacity of punishment.*
b) *The extinction of punishment is detrimental for the return ratio, unless there exists a high capacity of punishment.*

These findings provide evidence against Hypothesis 3. The introduction of punishment might decrease (increase) the intrinsic (extrinsic) motivation to reciprocate, but punishment should foster trustworthiness when the capacity of punishment is sufficiently high. One plausible explanation of why punishment backfires in that scenario is to consider that punishment is being perceived as unfair by allocators. This argument is frequently used to explain behavior in public good games (e.g., Denant-Boemont et al. 2007, Nikiforakis 2008) and can explain the findings in Fehr and

[22] The Pearson's correlation coefficient supports this idea. The return ratio is negatively correlated with the amount received when there is punishment ($r = -0.192$, p-value=0.032) but the correlation is not significant when there is not ($r = -0.076$, p-value=0.437). We interpret that when allocators receive a higher transfer with punishment (as it actually happens when the capacity of punishment is high), they might be less willing to reciprocate by increasing the return ratio, since they perceive the larger offer as being associated to the possibility of being sanctioned, and not to the investor's intrinsic motivation.

Rochemback (2003), where the punishment crowds-out reciprocity if investors threaten with the imposition of a fine. In Fehr and Rochemback (2003) the fine is costless for investor whereas in our experiment the punishment is costly but depends on the capacity of punishment (i.e., the highest the capacity of punishment, it is less costly for the investor to destroy the allocator's payoff completely). In our case, the capacity of punishment is indeed endogenous. As we argue in the concluding remarks, this distinctive feature of our experiment can reconcile with our data. Next we present our econometric analysis that supports Results 5, 6 and 7, and highlights that neither the allocator's gender nor the GSS question has a predictive power on the return ratio.

5.2.3. Regression analysis and behavioral determinants of trustworthiness

Table 2 provides the results of a panel random-effect regression where the dependent variable is the return ratio (y).[23] The set of independent variables includes the amount of ECUs received from the investor, the period in which the decision is made (PERIOD), the allocator's earnings in the previous round (PREVEARN), the levels of endowment (i.e., the dummies e_a^H and e_b^H defined above), the existence of the institution (PUN), the order in which the punishment is implemented (PUNFIRST) and the data collected in the questionnaire regarding the allocators' gender, age, and the answer to the GSS question.

[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 punishment is not allowed. The effect of introducing punishment on the return ratio is positive and significant. In line with our Result 7, the χ^2-test suggests that the positive effect of punishment on the return ratio takes place in every distribution except in the one with a high capacity of punishment. When punishment is removed, the return ratio decreases in all the distributions except the one in which the capacity of punishment is high. The χ^2-tests also support Result 6. The levels of endowment (e_a^H and e_b^H) do not determine the return ratio in the absence of punishment. With punishment, the allocator's behavior in (40,10) is statistically different from the behavior in the rest of distributions.[24]

[23] Our results are invariant if we use instead the model of Ashraf et al. (2006) and Chaudhuri and Gangadharan (2007) where the dependent variable is the proportion of the generated surplus that allocators return (Y/3X).

[24] The results of the χ^2-tests for the effect of introducing and eliminating punishment are presented in Table 5.B in the appendix. Table 6.B reports the p-values of the χ^2-tests for pairwise comparisons across distributions. In Table 7.B we show that the order in which treatments are implemented is not important to explain the allocator's behavior.

Table 2 suggests that the amount that allocators receive from investors has a non-significant effect on the return ratio.[25] Similarly, the period, the GSS question and the allocator's gender do not have any predictive value when estimating the return ratio. Our result for gender confirms previous research (e.g., Rigdon, 2009).

Result 8. (Demographics and trustworthiness)

Return behavior is gender invariant. Similarly, the GSS question that measures attitudinal trust does not have effect on the return ratio.

5.3. Efficiency of punishment and punishing behavior

We have seen that investors send a higher proportion of their endowment to allocators when their capacity of punishment is high. If investors refrained from punishing, this would be beneficial for the level of efficiency through an increase in the total payoffs. Arguably, the use of the punishment will yield an efficiency loss by destroying both subjects' payoffs.

In Figure 6 we investigate how the possibility of punishment affects the level of efficiency in each of the distributions. The vertical axis represents how large is the sum of the final payoffs with respect to the initial endowments. We compute the percentage of generated surplus in the trust game using the measure EG = $(\overline{\pi}^a+\overline{\pi}^b)/(e_a+ e_b)$ - 1. In the first column we plot the proportion of the generated surplus in the treatment without punishment as a natural benchmark. The second column plots the relationship between the interim payoffs and the initial endowments (i.e., the level of efficiency *before* investors employ the punishment). Finally, the third column shows the ratio between final payoffs and initial endowment once punishment has taken place. Errors bar reflect one standard error.

[Figure 6 around here]

As already suggested by Result 1, there are no interim efficiency gains associated to punishment, except if the investor's capacity of punishment is high (p-value = 0.017).[26]

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

[25] 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 = 0.3728, p-value=0.0001; PUN: r_s = 0.3852, p-value=0.0000). Arguably, "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).

[26] The reported test statistics 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.

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). Overall, these results suggest that investors decide to punish and generate efficiency losses, as already found in public good experiments (Chaudhuri, 2011).[27]

Result 9. (Efficiency of the punishment) *Investors punish allocators in all the distributions and destroy the gains of efficiency. Only if the capacity of punishment is high, the sum of the final payoffs 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.*

Although we consider that studying punishing behavior is beyond the scope of this paper, we investigate partial correlations between the investor's punishing behavior and decisions in the trust game for the sake of completeness.[28]

We know that trusting reduces the investor's capacity of punishment in our setup. As a result, we find that investors are less likely to punish the higher the level of trust is (r = -0.251, p-value= 0.007). Strikingly enough, once investors decide to punish (what occurs roughly 42% of the times) they devote more resources to punish the more they trust (r = 0.697, p-value= 0.000). The results for trustworthiness 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).[29]

6. Discussion and concluding remarks

In this paper, we have studied the extent to which allowing for investors' punishment affect the levels of trust and trustworthiness in the trust game (Berg et al., 1995). Using a within-subjects design, we manipulated both the investor and allocator's endowment and find that the capacity of punishment (measured by the relation of between the level of endowments) crucially determines the outcome.

Punishment fosters trust when the investor's capacity of punishment is high, but fails otherwise. Trustworthiness is higher with punishment than without punishment, except if investors have a high capacity of punishment.

[27] Interestingly, punishment decreases investors' payoffs only in the (40, 10) distribution, while allocators' earnings are lower with punishment in all distributions except in that distribution (see Table 8.B in the appendix).

[28] When considering partial correlations, we isolate the effect of the level of endowments. t

[29] We consider a regression analysis in the appendix. We show tha

These findings are clearly incompatible with models of social preferences, and can be explained by models of intrinsic and extrinsic motivation only partially. A complete explanation requires a precise understanding of the relation between incentives and motivation in our particular setting, where the extrinsic motivation, the intrinsic motivation and the *"credibility"* effect are linked in a peculiar and strategic way

Models on intrinsic and extrinsic motivation highlight that when punishment is allowed, a price effect is generated. If the threat is credible, then the extrinsic motivation to give and to pay back increases (monotonically) with the capacity of punishment. Thus, the incentivized behavior becomes more attractive (Prendergast 1999). Arguably, the introduction of punishment may have an opposite psychological effect on intrinsic motivation, causing a *crowding-out* effect on the incentivized behavior (see Deci et al., 1999; Frey, 1997; Gneezy et al., 2011).

In the trust game, the intrinsic motivation of the investor declines because trusting losses its self-image value or reputation value. As for the allocator, he perceives the investor's decision as not as truly *genuine*, but probably interested or based in an implicitly threat of punishment if she does not pay back enough.

In our setting, there is an additional effect on behavior due to the fact that trust and trustworthiness determine, along with the endowments and the effectiveness of the punishment, how credible the punishment is. When the capacity of punishment is low, investors might want to keep money in their pocket to make the punishment more credible. This *credibility* effect will be weaker and probably negligible when the capacity of punishment is high.[30] The natural consequence is that any positive *price-effect* of punishment may be severely mitigated if investors choose to keep a significant part of their endowment to make their threat credible. The effect works in the opposite direction for the allocator. He is aware that a high return ratio will reinforce the investor's capacity of punishment. Thus, the allocator might want to reduce the level of trustworthiness (weakening the capacity of punishment) when the capacity of punishment is high. The effect will be very low when the capacity of punishment is low. As a result, the *credibility effect* may decrease trust when the capacity of punishment is low, and may dumper trustworthiness for high capacity of punishment.

This idea is very consistent with our results. Punishment fosters trust when the *stick* is big because the price effect dominates the crowding effect (as models on intrinsic and extrinsic motivation predict), but also because there is no need to make punishment credible when the capacity of punishment is high.[31] Trustworthiness significantly increases with punishment, except if the investor's capacity of punishment is high. We argue that investors carrying a big stick send more, what reduces the allocator's intrinsic motivation. The fear of punishment, on the contrary, triggers some extrinsic motivation

[30] In other words, by sending too much, the investor loses the opportunity of punishing an allocator that sends back nothing. 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.

to return back but incentives to reciprocate are affected by the capacity of punishment. If it is high, allocators might decide that punishment is unfair and react accordingly by not increasing trustworthiness. Allocators do not need to weaken a low capacity of punishment and are more likely to reciprocate when the capacity of punishment is low.

Whereas models of social preferences cannot explain our results for trust and trustworthiness, models of intrinsic and extrinsic motivation do it only partially. In particular, we find that it is important to account for the credibility effect when explaining allocator's behavior. We find that our study extends the "pay-enough-or-do-not-pay-at-all" argument in Gneezy and Rustichini (2000a) and Gneezy et al. (2011) in at least two different directions. First, we focus the debate about intrinsic and extrinsic motivation in the punishment setting rather than in the reward one. Second, we consider an scenario in which parties contribute to the credibility of the punishment, so that the capacity of the institution is not exogenously given.

References

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

Ashraf, N., Bohnet, I., Piankov, N., 2006. Decomposing trust and trustworthiness. Experim. Econ. 9, 193–208.

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

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Amer. Econ. Rev., Pap. Proc. 97, 170-176.

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

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

Bohnet, I., Frey, B., Huck, S., 2001. More order with less law: On contract enforcement, trust, and crowding. Amer. Pol. Science Rev., 95, 131–44.

Boyd, R., Gintis, H. Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proceed. Nat. Acad. Sciences of the 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. J. Econ. Behav. Organiz. 68, 466-476.

Camerer, C. F., 2003. Behavioral game theory: experiments in strategic interaction. Russell Sage Foundation, New York, Princeton University Press.

Charness, G., Cobo-Reyes, R., Jiménez, N., 2008. An investment game with third-party intervention. J.Econ. Behav. Organiz. 68, 18–28.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica. 74, 1579-1601.

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

Chaudhuri, A., Gangadharan, L., 2007. An experimental analysis of trust and trustworthiness. South. Econ. J. 73, 959–985.

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.

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

Denant-Boemont, L., Masclet, D., Noussair, C., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Econ. Theory. 33, 145-167.

Dufwenberg, M. and 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. and Johannesson, M., 2008. Pride and Prejudice: The Human Side of Incentive Theory, American Economic Review, 98(3), 990-1008.

Falk, A. and Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior, 54(2), 293-315.

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

Fehr, E. and 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. Exper. Econ.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. J. Econ.Behav. Organiz. 69, 226–240.

Glaeser, E., Laibson, D., Scheinkman, J., Soutter, C., 2000. Measuring trust. Quart. J. Econ. 115, 811-846.

Gneezy, U., Rustichini., A., 2000a. Pay enough or don't pay at all. Quart. J. Econ. 115, 791–810.

Gneezy, U., Rustichini, A., 2000b. A fine is a price. J. Legal Stud. 29, 1–18.

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

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.

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

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

Johnson, N. D., Mislin, A., 2011. Trust games: A meta-analysis. J. Econ. Psyc. 32, 865–889.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? A cross-country investigation. Quart. J. Econ. 112, 1251-1288.

Kosfeld, M. and Falk,A., 2006. The Hidden Cost of Control, American Economic Review, 96(5), 1611-1630.

Nikiforakis, N., 2008. Punishment and counter-punishment in public goods game: Can we still govern ourselves? J. Public Econ. 92, 91-112.

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

Olcina G. and Calabuig, V., 2013a. Coordinated Punishment and the Evolution of Cooperation, Journal of Public Economic Theory, forthcoming.

Olcina G. and Calabuig, V., 2013b, Cultural Transmission and the Evolution of Trust and Reciprocity in the Labor Market, Documentos de Trabajo, 11, Fundación BBVA.

Prendergast, C., 1999. The provision of incentives in firms. J. Econ. Lit. 37, 7–63.

Rabin, M., 1993. Incorporating Fairness into Game Theory and Economics. The American Economic Review, 83 (5), 1281-1302

Rigdon, M., 2009. Trust and reciprocity in incentive contracting. J. Econ. Behav. Organiz. 70, 93–105.

Figure 1. Capacity of punishment for different level of endowments

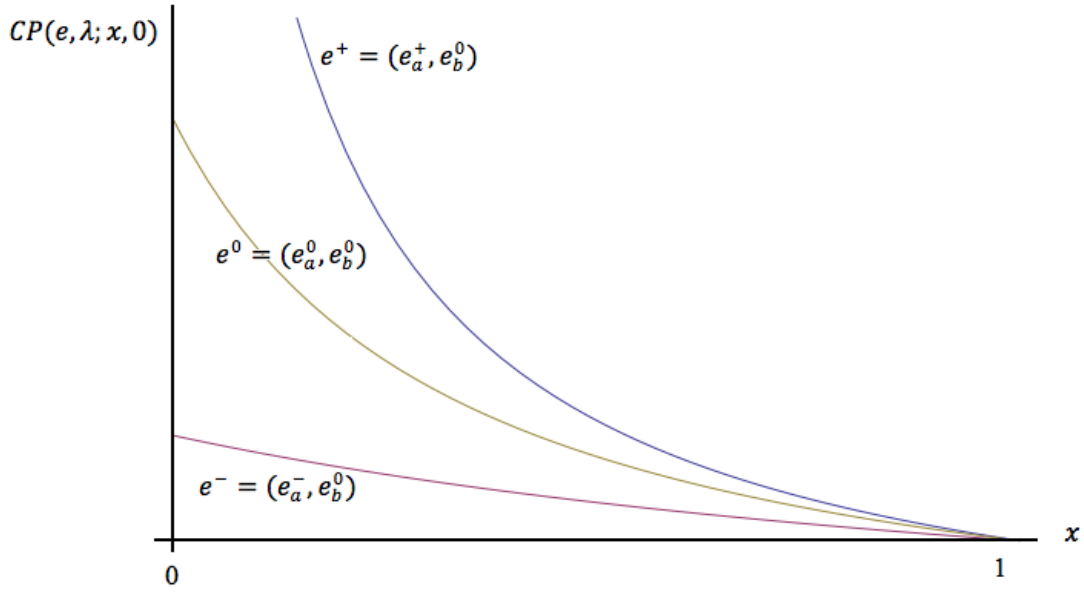
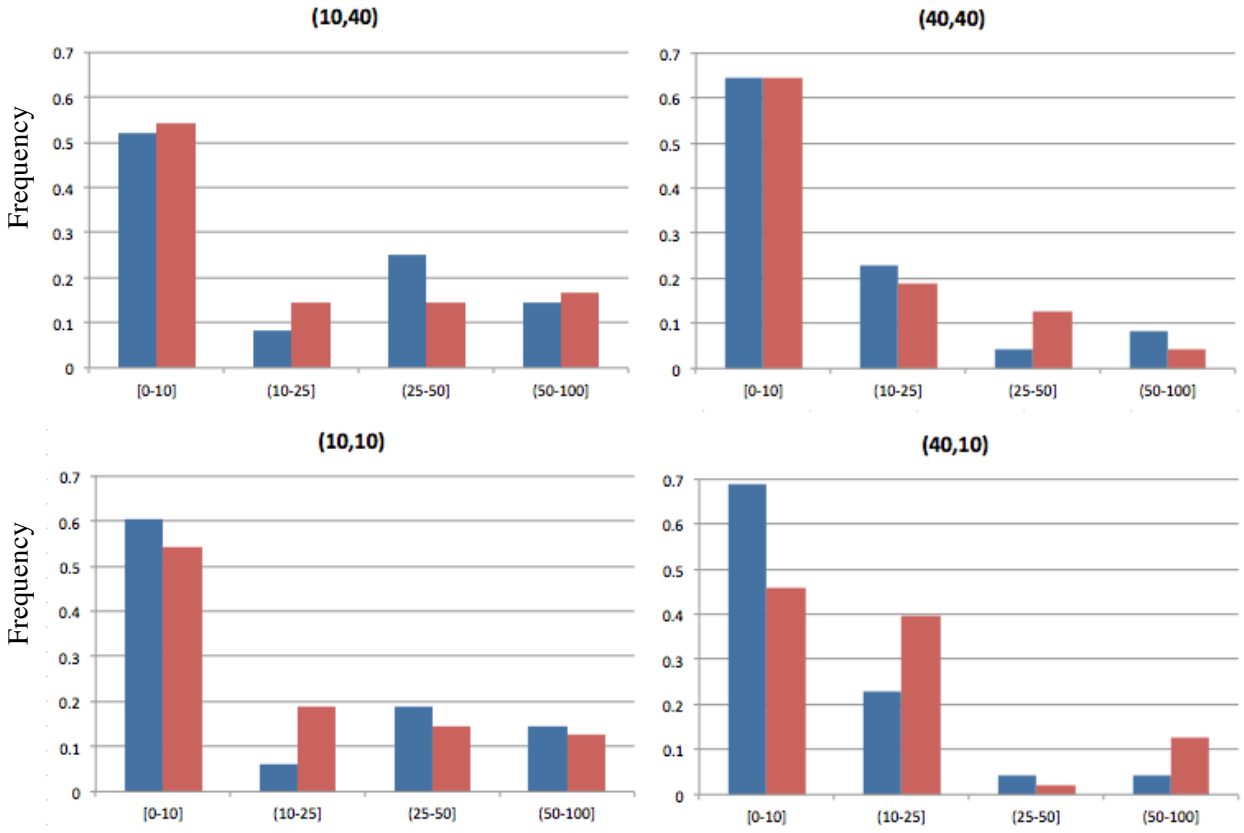


Figure 2. The effect of punishment on the level of trust: relative transfer in each distribution



Treatment	Distribution (10,40)	(40,40)	(10,10)	(40,10)
NOPUN	0.266	0.151	0.243	0.112
	(0.32)	(0.26)	(0.34)	(0.22)
PUN	0.258	0.141	0.219	0.199
	(0.34)	(0.21)	(0.29)	(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 3. The effect of introducing and eliminating punishment on the investor's behavior.

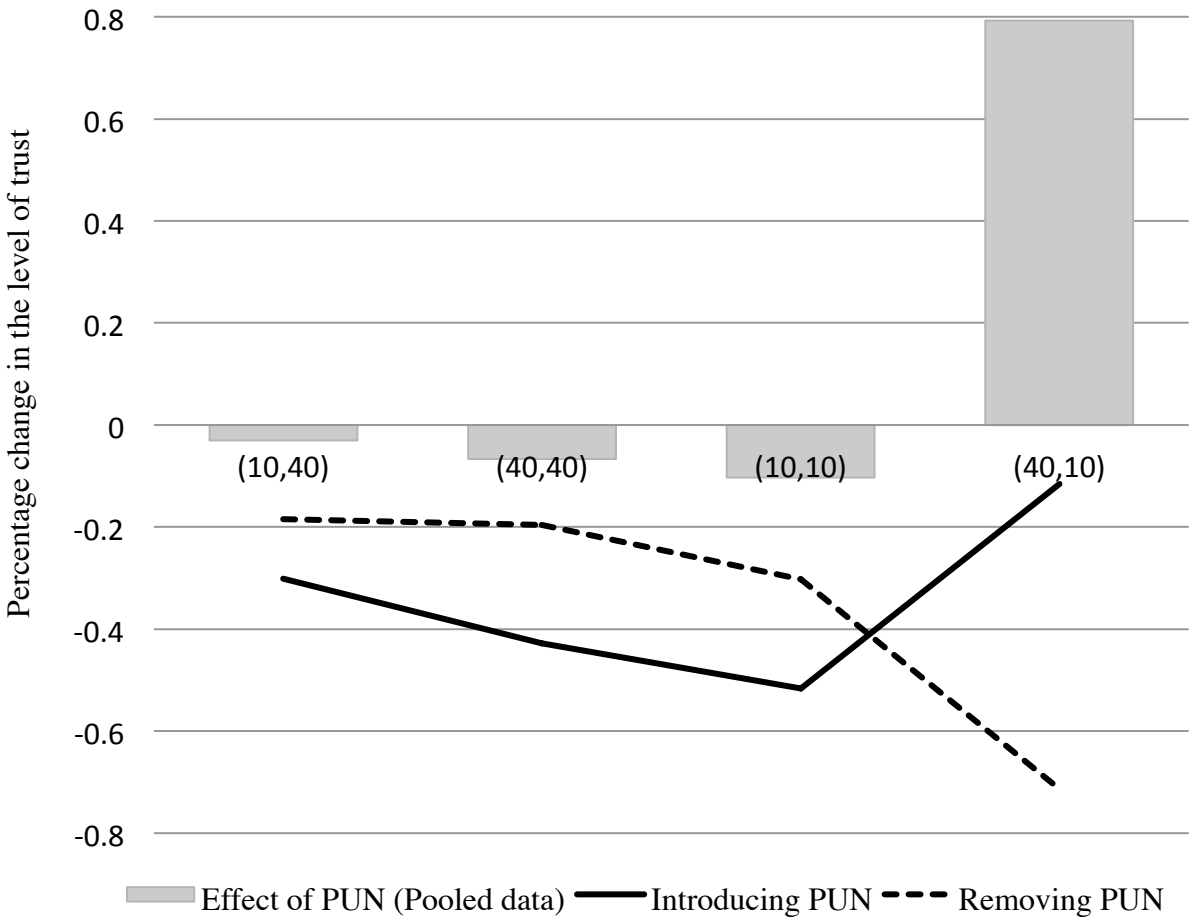
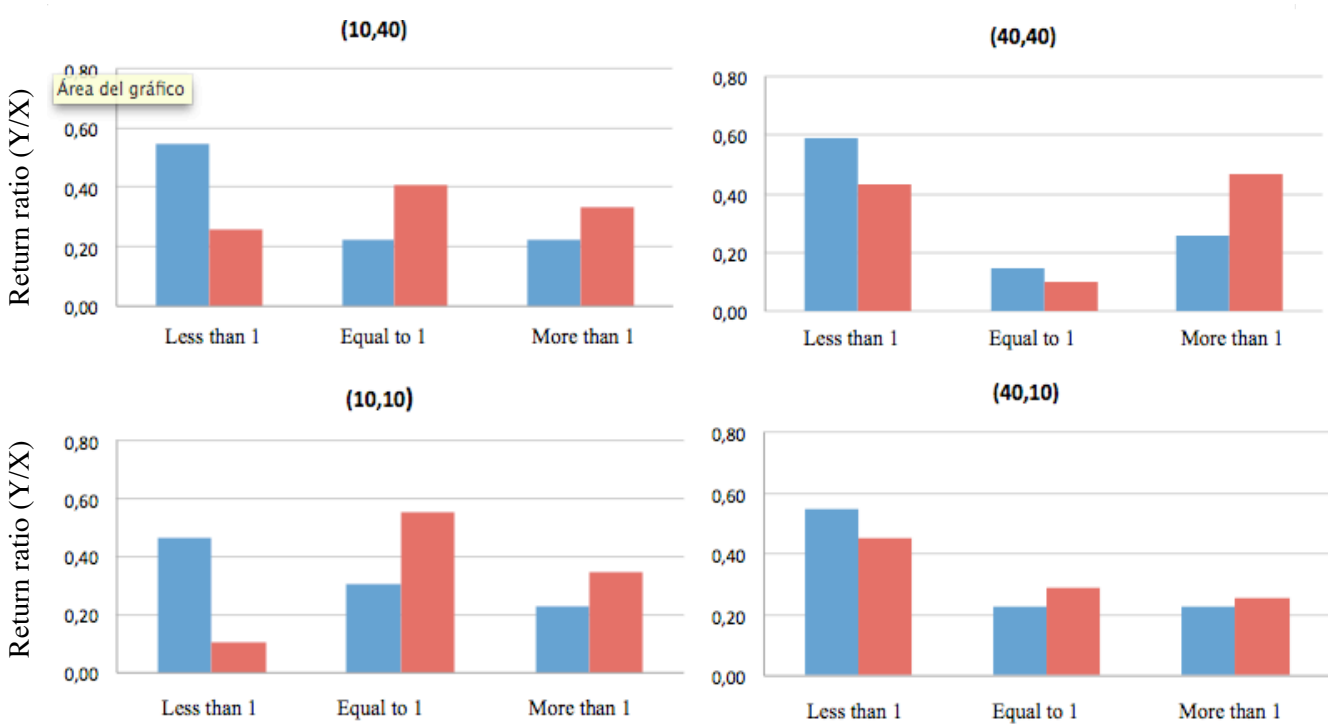


Figure 4. The return ratio in each treatment for each distribution



Treatment	Distribution (10,40)	(40,40)	(10,10)	(40,10)
NOPUN	0.659	0.713	0.857	0.660
	(0.69)	(0.72)	(0.87)	(0.58)
PUN	1.266	1.078	1.241	0.711
	(0.62)	(0.94)	(0.62)	(1.02)
t-test (t)	0.001	0.048	0.027	0.449
Wilcoxon test (Z)	0.001	0.071	0.090	0.413

Figure 5. The effect of introducing and eliminating punishment on the allocator's behavior.

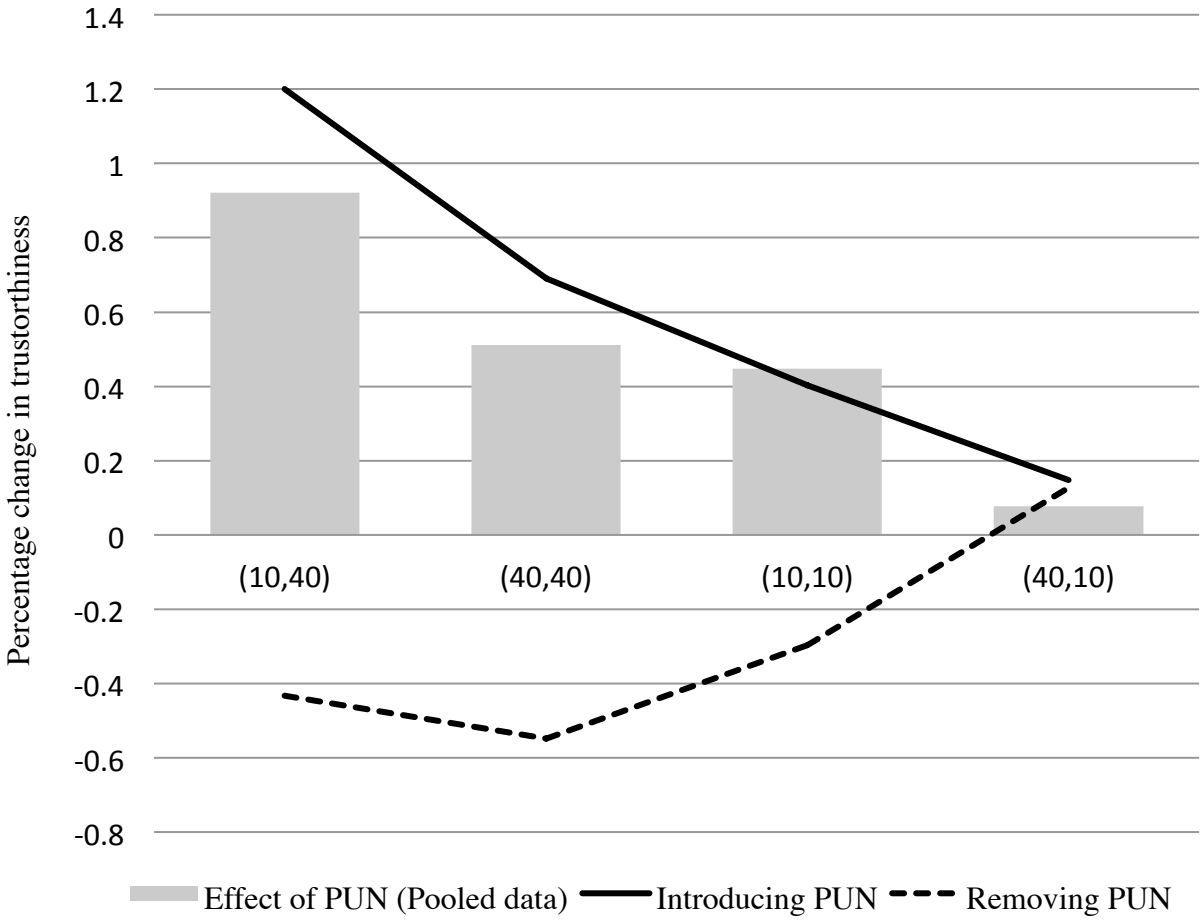


Figure 5. Efficiency gains in each distribution comparing the sum of the final payoffs and the sum of the initial endowments

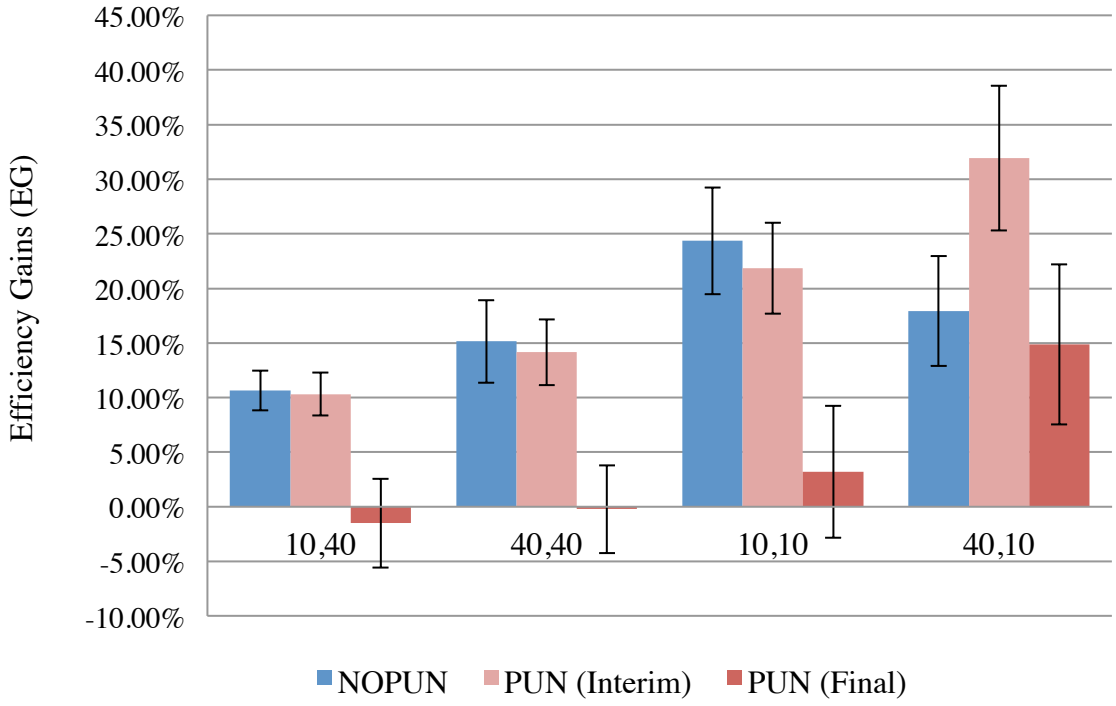


Table 2. Maximum-likelihood estimates of a random effects model that controls for unobserved individual heterogeneity.

INDEPENDENT VARIABLES	(1)	(2)	(3)
PERIOD	-0.024**	-0.024**	-0.024**
	(0.0120)	(0.0120)	(0.0120)
PREVEARN	0.001	0.001	0.001
	(0.0011)	(0.0011)	(0.0011)
PUN	-0.121***	-0.0331	-0.121***
	(0.0452)	(0.0480)	(0.0451)
PUNFIRST	-0.063		-0.023
	(0.0667)		(0.0637)
PUN x PUNFIRST	0.175***		0.175***
	(0.0542)		(0.0538)
e_a^H	-0.134***	-0.135***	-0.134***
	(0.0424)	(0.0418)	(0.0441)
e_b^H	-0.007	-0.007	-0.004
	(0.0357)	(0.0357)	(0.0349)
$e_a^H e_b^H$	0.061	0.060	0.052
	(0.0456)	(0.0457)	(0.0451)
PUN e_a^H	0.134***	0.134***	0.135***
	(0.0402)	(0.0399)	(0.0404)
PUN e_b^H	0.047	0.047	0.047
	(0.0475)	(0.0473)	(0.0473)
PUN $e_a^H e_b^H$	-0.146*	-0.146*	-0.147*
	(0.0802)	(0.0799)	(0.0800)
WOMEN	-0.173***	-0.178***	
	(0.0530)	(0.0521)	
AGE	-0.004	-0.005	
	(0.00773)	(0.0068)	
GSS	0.092**	0.099***	
	(0.0413)	(0.0366)	
Constant	0.477***	0.479***	0.292***
	(0.175)	(0.149)	(0.0582)
σ_u	0.1632	0.1599	0.1632
σ_e	0.2019	0.2073	0.2019
ρ	0.3959	0.3730	0.3952

standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2. Return ratio: Maximum-likelihood estimates of a random effects model that controls for unobserved individual heterogeneity.

Independent Variables	(1)	(2)	(3)
X (ECUs received)	-0.00169	-0.00653	-0.00352
	(0.00904)	(0.00870)	(0.00902)
PERIOD	0.0376	0.0430	0.0429
	(0.0737)	(0.0733)	(0.0730)
PREVEARN	0.00156	0.000618	0.000883
	(0.00283)	(0.00279)	(0.00279)
PUN	0.626**	0.527**	0.569**
	(0.261)	(0.238)	(0.255)
PUNFIRST	-0.244		-0.215
	(0.213)		(0.219)
PUN x PUNFIRST	-0.108		-0.0737
	(0.252)		(0.247)
e_a^H	-0.104	-0.0867	-0.0910
	(0.268)	(0.267)	(0.266)
e_b^H	-0.00400	-0.0864	-0.0594
	(0.264)	(0.255)	(0.260)
$e_a^H e_b^H$	0.266	0.313	0.252
	(0.359)	(0.355)	(0.356)
PUN e_a^H	-0.255	-0.261	-0.253
	(0.344)	(0.341)	(0.340)
PUN e_b^H	0.133	0.175	0.187
	(0.337)	(0.330)	(0.331)
PUN $e_a^H e_b^H$	-0.0113	-0.0578	-0.0315
	(0.484)	(0.479)	(0.476)
WOMEN	0.0609	0.0726	
	(0.155)	(0.165)	
AGE	0.0445*	0.0370	
	(0.0233)	(0.0244)	
GSS	-0.0300	-0.0943	
	(0.207)	(0.216)	
Constant	-0.479	-0.341	0.559*
	(0.608)	(0.627)	(0.313)
σ_u	0.2684	0.3464	0.2667
σ_e	0.7118	0.7031	0.6891
ρ	0.1393	0.1953	0.1303

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Carry a big stick, or no stick at all

An experimental analysis of trust and endogenous punishment

Vicente Calabuig
ERICES, Universidad de Valencia

Enrique Fatas
University of East Anglia

Gonzalo Olcina
ERICES, Universidad de Valencia

Ismael Rodriguez-Lara
ERICES, Universidad de Valencia
LUISS Guido Carli University

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 supplementary econometric analyses of our data, which supports the findings discussed in our manuscript.

APPENDIX A:

INSTRUCTIONS [1]

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.

[1] 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:

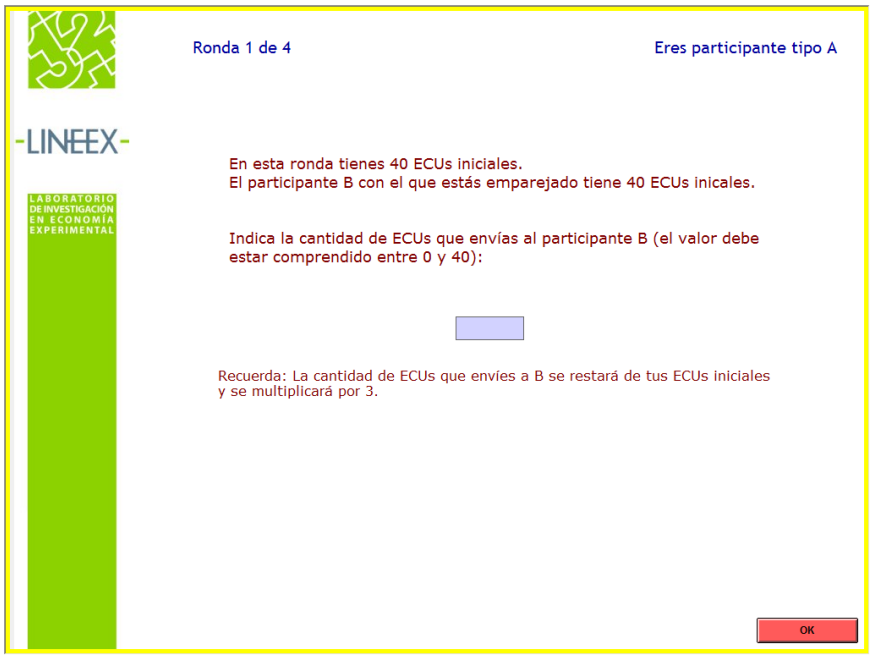
Final payoff player A = Provisional payoff player A – points sent by A

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.

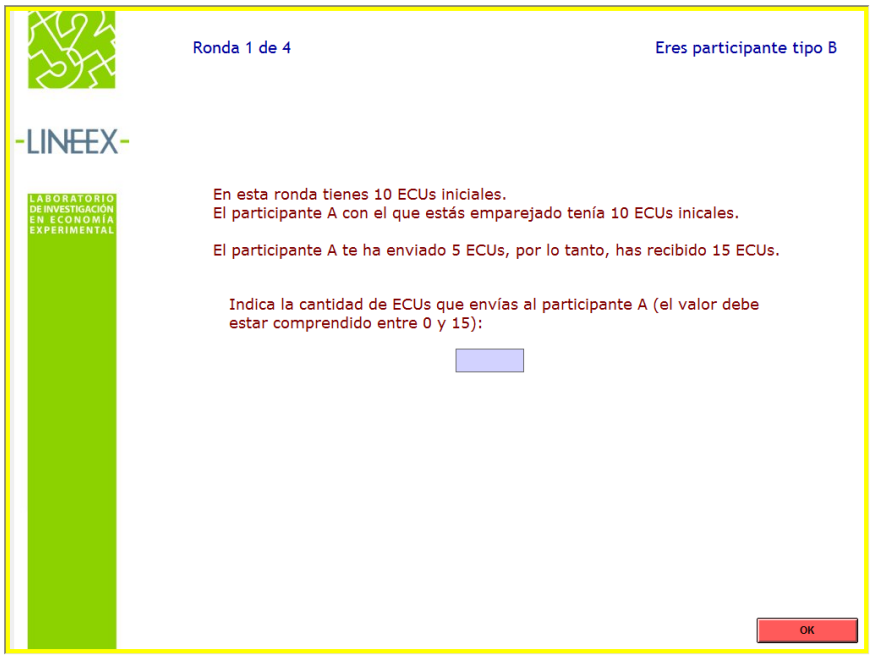
SCREENSHOTS

I. The investor's Behavior: Trust



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



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.

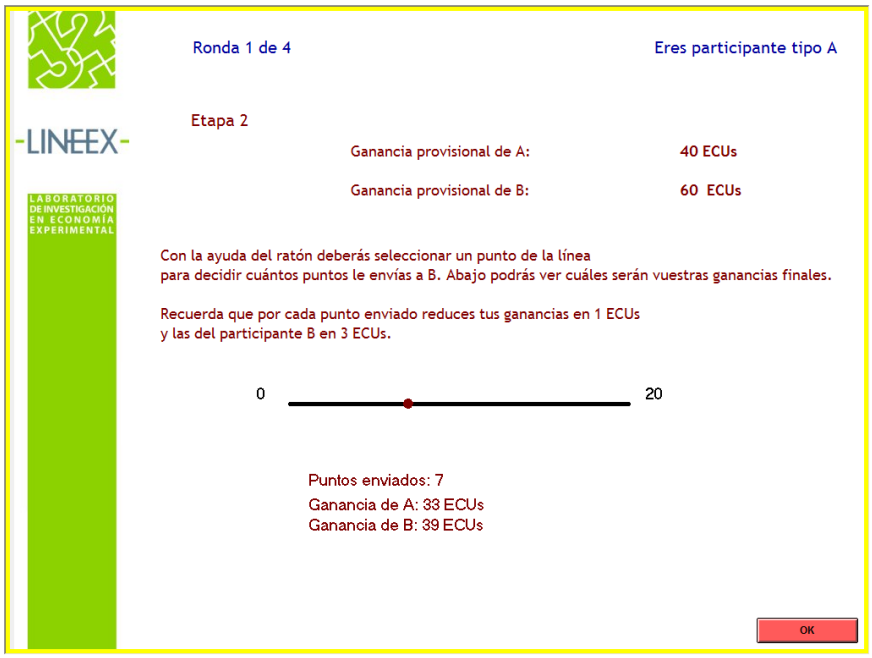
We decided to inform subjects about their earnings at the end of each round because in the punishment treatment, this information must be available for investors to decide whether to punish or not. With our design, we wanted to avoid that subjects received more information (feedback) in the treatment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



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 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 Table 4.B) Results on efficiency are reported in Section II (Table 5.B).

I. The investor's Behavior

Table 1.B. Maximum-likelihood estimates of a random effects model that controls for unobserved individual heterogeneity. The dependent variable is the relative transfer send by investors in each of the distributions (X/e_a). The set of independent variables include the period, 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 (in brackets) take into account matching group clustering.

Independent Variables	(10,10)	(10,40)	(40,10)	(40,40)
PERIOD	-0.0389	-0.0472	-0.0112	-0.0273
	(0.0248)	(0.0315)	(0.0402)	(0.0286)
PUN	-0.0250	-0.00833	0.0875**	-0.00990
	(0.0481)	(0.0548)	(0.0416)	(0.0445)
WOMEN	-0.122	-0.202***	-0.135***	-0.176***
	(0.0754)	(0.0533)	(0.0490)	(0.0310)
AGE	-0.0144	-0.0167*	0.00990	-0.000991
	(0.0118)	(0.00871)	(0.00759)	(0.00719)
GSS	0.153**	0.210***	-0.0361	0.0800**
	(0.0600)	(0.0776)	(0.0564)	(0.0320)
Constant	0.698**	0.827***	0.00129	0.324*
	(0.286)	(0.174)	(0.139)	(0.168)
Wald-test	16.15***	39.27***	9.76*	223.36***

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In line with our Result 1 in the main text, we find that the possibility of punishment (PUN) does only have a significant positive effect when there exists a high capacity of punishment; i.e., in the distribution (40,10).

Table 2.B. We report the outcome of pairwise comparisons between the four different distributions using the Wilcoxon signed-rank in Panel A of Table 2.B. Hypothesis testing using the χ^2-test after estimating the econometric model of section 5.1.3 are presented in Panel B of in Table 2.B. The p-values for the NOPUN and PUN treatment appear in the grey and the white area respectively.

A. Wilcoxon signed-rank test

	(10,10)	(10,40)	(40,10)	(40,40)
(10,10)	-	0.3725	0.0027	0.0219
(10,40)	0.9474	-	0.0024	0.0004
(40,10)	0.5300	0.2556	-	0.9492
(40,40)	0.0173	0.0122	0.0609	-

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

	(10,10)	(10,40)	(40,10)	(40,40)
(10,10)	-	0.8401	0.0016	0.0227
(10,40)	0.3548	-	0.0170	0.0602
(40,10)	0.6523	0.1689	-	0.1745
(40,40)	0.0715	0.0064	0.0765	-

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.

Table 3.B. Logit model that estimates the probability that the investor decides to transfer part of her endowment to the allocator, Prob(x) > 0). The set of independent

variables coincides with the one reported in Table 1B above. The robust standard errors (in brackets) take into account matching group clustering.

Independent Variables	(10,10)	(10,40)	(40,10)	(40,40)
PERIOD	-0.247	-0.0721	-0.365**	-0.276*
	(0.182)	(0.114)	(0.147)	(0.158)
PUN	0.174	-0.225	0.519	0.173
	(0.283)	(0.196)	(0.360)	(0.265)
WOMEN	-0.00991	-0.270	-0.0986	-0.276
	(0.426)	(0.296)	(0.268)	(0.321)
AGE	-0.0575	-0.0617	0.0407	0.0321
	(0.0626)	(0.0419)	(0.0401)	(0.0408)
GSS	0.497	0.697***	-0.206	0.0325
	(0.383)	(0.248)	(0.416)	(0.203)
Constant	1.914	1.959**	-0.0106	0.299
	(1.557)	(0.857)	(0.964)	(0.890)
Pseudo R^2	0.047	0.036	0.098	0.066
Log-pseudolikelihood	-62.44	-62.11	-59.54	-60.58

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Our regression suggests that the probability of trusting does not change with the possibility of punishment in any of the distributions.

Table 4.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 2, 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_1^2 (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: $\alpha_{PUNFIRST} + \alpha_{PUNxPUNFIRST} = 0$	6.05 (0.011)
The level of trust in the NOPUN treatment is the same when the game is played first or second in the session.	H_0: $\alpha_{PUNxPUNFIRST} = 0$	0.88 (0.347)

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. Efficiency and final payoffs

Table 5.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			
	(10,10)	**(10,40)**	**(40,10)**	**(40,40)**
Investor's payoffs				
NOPUN	9.270	9.000	39.375	36.958
PUN	10.021	8.375	35.687	36.146
t-test (t)	0.187	0.418	0.062	0.652
Wilcoxon test (Z)	0.567	0.609	0.042	0.756
Allocator's payoffs				
NOPUN	15.604	46.333	19.583	55.167
PUN	10.625	40.875	21.750	43.687
t-test	0.001	0.018	0.568	0.033
Wilcoxon test (Z)	0.000	0.009	0.350	0.011

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.