

Social Influence on Third-Party Punishment: an Experiment[☆]

Emanuela Carbonara^a, Marco Fabbri^{a,b,c}

^a*Department of Economics, University of Bologna*

^b*Hamburg Institute of Law and Economics, Hamburg University*

^c*Rotterdam Institute of Law and Economics, Erasmus University Rotterdam*

Abstract

In this paper we study the effects of social influence on third-parties' decision to engage in costly punishment. In the context of a dictator game, we exclude payoff complementarity among punishers and elicit punishment decisions both in isolation and after having provided information regarding actual peers' punishment. We find evidence that the amount of punishment chosen by third-parties is influenced by beliefs about the amount of peers' punishment. Moreover, the larger the difference between third-parties beliefs about the level of peers' punishment and actual peers' punishment, the more likely the third-parties modify the initial punishment decision. We also find that more self-regarding third-parties are less affected by social influence. Finally, we disentangle the effect of normative social influence from that of informational social influence and we show that some subjects are responsive to the former type of social influence but not to the latter.

Keywords: Third-party Punishment, Social Influence, Peer Pressure, Experiment, Dictator Game

JEL Classification: D03, K04

[☆]We are grateful to the Alfred P. Sloan foundation for financial support. We thank Maria Bigoni, Marco Casari, Andrea Geraci, Francesco Parisi, Louis Visscher, Roberto Weber and seminar participants at Erasmus University Rotterdam for helpful suggestions and Stefano Rizzo for valuable research assistance. The usual disclaimer applies.

1. Introduction

Societies often rely on punishment for preventing and eventually responding to rule violations. Punishment is a costly activity that inflicts negative consequences to wrongdoers and that is carried on both by formal centralized institutions and by the decentralized actions of peers. When the punishment activity is inflicted directly by agents that are bearing the costs of the rule violation, we talk about second-party punishment. However, in groups composed by a large number of agents interactions are often non-repeated and the punishment activity is typically carried on by a third-party not directly affected by the consequences of the rule violation. In these cases we refer to third-party punishment.

While scholars' attention has traditionally focused on second-party and centralized third-party punishment, in recent years a growing body of contributions analyzes empirically the role of decentralized third-parties in punishing rules and norms violators (Fehr and Gächter, 2002; Fehr et al., 2003; Fowler, 2005). However, despite there has been substantial progress in identifying the determinants of decentralized third-party punishment (Bernhard et al., 2006; Lewisch et al., 2011; Lieberman and Linke, 2007; Coffman, 2011), there is only a partial understanding of what are its major determinants yet (Fehr and Fischbacher, 2004a).

In this paper we focus on decentralized punishment, examining the effects of social influence on the punishment behavior of bystanders not directly affected from the action of the wrongdoers¹. By social influence we refer to the effect of the endogenous interactions between a third-party's preferences for punishment and the preferences for punishment expressed by other bystanders (Manski, 2000). Focusing on endogenous interactions means that we

¹In order to minimize repetitions, when talking about decentralized punishment we will employ the terms "third-party" and "bystander" interchangeably throughout the paper.

investigate the influence that the punishment choices of other third-parties have on the decision of a bystander to engage in punishment, ruling out the effects produced by self-selecting into the same group (contextual interactions) or by sharing common individual characteristics (correlated effects). Focusing on preference interaction means that we study how the utility of a bystander is affected by information on other third-parties' punishment choices when this information do not modify her choice set and payoff complementarity between bystanders is excluded.

Scholars report field and experimental evidence that social influence is a major determinant of human behavior in a variety of settings characterized by important economic consequences, like teenage pregnancy (Akerlof et al., 1996), obesity (Christakis and Fowler, 2007), judicial voting patterns (Sunstein et al., 2006), investment strategies (Hirshleifer and Hong Teoh, 2003), tax evasion (Fortin et al., 2007; Galbiati and Zanella, 2012) and other criminal activities (Glaeser et al., 1996). However, empirically estimating social influence effects on decentralized third-party punishment presents two major identification problems. On one hand, it requires to rule out problems of self-selection and confounding factors like correlated effects or the possibility that third parties' material payoff is modified as a consequence of the exposure to social information. On the other hand, in most societies punishment of rule violators is carried on by a centralized system based on codified legal rules that coexists and sometimes overlap with a decentralized system based on informal norms (Akerlof, 1989; Cooter, 1998). As a consequence, it is often impossible to isolate the effects of social influence on decentralized punishment behavior analyzing field data. Therefore, we exploit the advantage that the controlled environment of a laboratory experiment offers in order to rule out self-selection problems and disentangle the effects of social influence from those of possible confounding factors.

Furthermore, Deutsch and Gerard (1955) suggest that social influence affects the behavior of an individual agent through two possible channels. On one

hand, under "informational" social influence an agent derives utility from doing what is the right action. Therefore, the agent's behavior is influenced by information received about peers' choices because of an update of her own beliefs regarding what is the correct thing to do. On the other hand, under "normative" social influence an agent derives (dis-) utility from being (dis-) liked by her peers. Therefore the agent's behavior is influenced by information regarding peers' choices because of the utility gain derived from being liked or the disutility coming from being negatively judged by them. Disentangling informational and normative social influence is important for policy purposes because, while the former has persistent effects on individual behavior, the effects of the latter is less robust and limited in time (Cason and Mui, 1998).

Therefore, in this paper we address the following questions: is social influence a major driver of third-party punishment? Does social influence play a role in bystanders' punishment decision through the channel of normative or informational influence?

Results of our experiment show that social influence is an important determinant of third-party punishment. Moreover, we find that bystanders engaging in high level of punishment are affected by social influence the most. We also find that the larger is the difference between a bystander's beliefs regarding peers' punishment and actual peers' punishment, the more informing the bystander about peers' behavior influences her punishment choices. Finally, we find that some subjects respond to normative social influence but not to informational social influence.

we proceed in this way. In the next section we provide a literature review. In section 4.3 the experimental design is presented. Section 4.4 specifies the theoretical framework and the hypothesis we test. Section 4.5 presents the experiment results and section 4.6 discusses our findings, suggests possible directions for future research and states the conclusions.

2. Literature Review

Recognizing the importance of punishment in the societal framework scholars have devoted great attention to the topic. Contributions in the early literature were mostly concerned with second-party punishment (SPP) or forms of centralized third-party punishment (TPP). The milestone of the literature in law and economics could be considered the work of Becker (1968). The author analyzes the decision of a potential criminal to violate the law in the framework of individual utility maximization, arguing that the criminal act would be carried on only if its expected benefit exceed the expected costs. Therefore, according to Becker's argument, an increase in punishment level and probability of being punished associated to criminal activities would result in the reduction of crime rate. Subsequent contributions extend Becker's original idea to the frameworks of regulation (Bose, 1995) and tax evasion (Slemrod and Yitzhaki, 2002, for a survey).

In the field of experimental economics, the seminal work by Güth et al. (1982) introduces the concept of "irrational punishment" in the context of the so called Ultimatum Game. In this game a receiver has to either accept or reject the share of an amount of money offered by a Proposer. If the offer is rejected, the receiver earns nothing but nullifies at the same time also the earnings of the Proposer. Contrary to game-theoretical prediction, the evidence shows that agents often prefer, at the price of leaving the offer of the Proposer on the table, punishing by rejecting positive, but considered unfair, offers.

Since Guth's contribution an extensive investigation of the determinants and characteristics of SPP has been conducted. Among the others, we mention the contributions of Ostrom et al. (1992), Gächter and Fehr (2000) and Fehr and Gächter (2002) that analyze costly punishment in commons and public goods setting. These studies find that the presence of a punishment mechanism substantially improves cooperation levels.

Subsequent articles further investigate the characteristic of SPP, suggesting

that it follows the law of demand (Carpenter, 2007) and it is driven more from an emotional satisfaction than from a rational need of justice (Casari and Luini, 2009), eventually leading to degeneration in riots and resources wasting (Nikiforakis, 2008).

Despite some pioneering contributions, (Axelrod, 1986; Bendor and Mookherjee, 1990; Ellickson, 1999), it is instead only starting from the last decade that scholars begin to investigate decentralized TPP. The groundbreaking articles are due to Fehr and Gächter (2002); Fehr et al. (2003); Fehr and Fischbacher (2004b). The authors show in laboratory experiments that third parties voluntarily incur costs in order to punish norm violations and that the amount of punishment increases with perceived unfairness. Subsequent works by Shinada et al. (2004) and Bernhard et al. (2006) report evidence that humans in their punishment decisions are subject to “parochial altruism”, tending to punish the rule-breakers more when he belongs to the same reference group than when he is an outsider. Okimoto and Wenzel (2011) confirm these findings showing that intra-group status influences TPP even if only symbolic and limited to the context of an anonymous laboratory experiment. Moreover Lieberman and Linke (2007) show that social categories have a significant effect on the level of TPP provided and Hoff et al. (2011) find that the legacy of cast culture in India influences norm enforcement, determining less TPP within the casts considered at the bottom of the society. More recently, Coffman (2011) shows that intermediation processes reduce both TPP and rewards and Lewisch et al. (2011) suggest that TPP suffers the free-riding problem when more than a single potential bystander are present.

In a cross-cultural study among 15 small scale societies, Henrich et al. (2006) find a significant variability on the level of TPP provided. Attempting to account for these differences, Marlowe et al. (2008) show that societies characterized by complex organizations and subject to frequent market interactions engage in higher level of TPP compared to less articulated ones. As a

consequence, the authors argue that institutions and social structures play a fundamental role in shaping our preferences for punishment. However, possibly in contrast with Marlowe’s findings, Mathew and Boyd (2011) in a following work show that third-party punishment could sustain large scale cooperation among African nomadic tribes during warfare period, suggesting that further studies are necessary in this area of research.

Contributions in applied psychology investigating the determinants of decentralized TPP have also flourished in the last decades. Kurzban et al. (2007) in a laboratory experiment find that subjects increase punishment when observers are present, arguing that TPP is influenced by the so called ”audience effect”. Subsequent works confirm that anonymity has a causal effect on TPP (Piazza and Bering, 2008), suggesting that the third party decision to sanction wrongdoer is influenced by a cost-dependent reputation effect (Nelissen, 2008) and by emotions (Nelissen and Zeelenberg, 2009). Moreover Lotz et al. (2011) suggest that differences in the level of third-parties punishment provided within a group of agents could be explained by heterogeneity in bystanders’ ”justice sensitivity”. Finally, a promising branch of research aims at explaining TPP through the investigation of the biological mechanisms governing human brain (Seymour et al., 2007; Buckholtz et al., 2008) or the behavior of other animal species (Raihani et al., 2010).

While to the best of our knowledge there are no contributions linking social influence effects and decentralized third-party punishment, nevertheless the possibility that agents’ behavior is influenced by peers has been since a long time object of interest for social scientists. Depending from the field of study and the context of the research, this behavior is called ”social influence”, ”neighborhood effect”, ”taste for conformism”, ”imitation” or ”herd behaviors”. Starting from the pioneering work of Asch (1951, 1956), contributions in experimental psychology show how individuals tend to modify and distort self-judgments under the influence of group pressure, culture influence and taste for conformism (for a survey see Bond and Smith, 1996).

Economists have been mostly interested in the implications of social influence effects for the functioning mechanisms of financial markets. Indeed, most of the contributions focus on the process of information acquisition in investment strategies (Cooper and Rege, 2008; Devenow and Welch, 1996; Scharfstein and Stein, 1990; for a survey see Hirshleifer and Hong Teoh, 2003). Also, economic scholars investigated the effects of social influence on the labor market. Studies report that peer pressure influences labor productivity (Falk and Ichino, 2006; Mas and Moretti, 2009) and that social networks characterized by an elevated percentage of unemployed individuals could generate social norms perpetuating unemployment (Akerlof, 1980; Topa, 2001). Moreover, reporting results of laboratory experiments, Falk and Fischbacher (2002) argue that social influence is a major driver of criminal behavior and Falk et al. (2010) and Krupka and Weber (2009) find that social influence plays a role in determining pro-social behavior.

Finally, we signal a series of recent policy interventions that exploit social influence effects in order to achieve welfare-improving results (for a discussion of further similar policies see also Thaler and Sunstein, 2008). The first framework of intervention is related to tax compliance. We already mentioned in the previous section that for an individual the likelihood to engage in tax evasion is affected by her peers' rate of tax evasion. However, social influence properly combined with framing effect might also help in increasing tax compliance. Indeed, Coleman (1996) reports the result of a field experiment conducted in Minnesota with the objective to find a costless strategy to increase tax compliance. In the experiment, a letter is sent to taxpayers by the tax authority few days before the annual tax file deadline. The letter could contain different information according to different treatments: a reminder to the civic obligation to pay taxes, information regarding the procedure to follow, a simple reminder of the deadline or information regarding the number of taxpayers that have already complied with tax payment (a percentage above 90%, that a survey analysis reveals people typically under-

estimated). This last treatment turns out to be the only treatment registering a statistically significant increase in the number of compliant tax payers.

Another example of behavioral public policy based on social influence is an interventions that encourages socializing nondrinking. Perkins et al. (2010) report results from a field experiment run in Montana, where surveys revealed that college students systematically overestimate the fraction of peers consuming alcoholic beverages. In the experiment, a random sample of college students were exposed to the (true) information that the overwhelming majority of the people in the state and of the students on campus consumes moderate quantities of alcohol. Results show that those exposed to this information decrease the consumption of alcohol compared to peers in the control group.

The last example we report is described in Cialdini (1993) and concerns a field experiment run in the Petrified National Forest, Arizona. Apparently, visitors tend to take home petrified fossils as a souvenirs, a behavior that over the years created serious concerns about the preservation of the park. Signs along the park trails ask people not to take samples away. However, Cialdini found that when the request on the sign is framed as an injunctive norm ("Please don't remove fossils from the park in order to preserve the Petrified Forest") people are significantly more compliant than when the request conveys information about other visitors' unlawful behavior ("Many past visitors have removed the petrified wood from the park, changing the state of the Petrified Forest"). This finding confirm that social influence, either for good or bad, represents a major driver of human behavior.

3. Experimental Design

The Game. we conduct a variant of Fehr and Fischbacher (2004) dictator game with TPP. Following Cox et al. (2007) and Swope et al. (2008), in the game a dictator has the possibility to take from a passive receiver some or all of the experimental monetary units (tokens) of the initial endowment

provided by the experimenter. The game has 3 possible roles: receiver (Participant A), dictator (Participant B) and Third-party (Participant C).

The game has 2 periods. Each period of the game is divided in two stages. At the beginning of each period, each participant is endowed with 30 tokens by the experimenter. In the first stage of each period, Participant B has the possibility to take from 0 up to 30 tokens (in multiples of 5) from A. Participants A cannot undertake any action during the game.

In the second stage of each period, Participant C has the opportunity to impose a costly punishment to B. Specifically, C could use up to 20 units of her initial endowment to reduce B's payoff. For each token used by C, the payoff of Participant B is reduced by 4 tokens. Participants C specify how many tokens they use in order to reduce B's payoff for each possible action chosen by B (strategy method). The tokens C uses for punishment in one period and the consequent reduction of B's payoff have no effect on the payoff of player A. Agents have full information regarding the rules of the game.

Before the game starts, participants' beliefs about the average punishment choices of the peers are elicited. To do so, we use an incentivized coordination game similar to Krupka and Weber (2013). We refer to this part of the experiment as the "Beliefs elicitation game". We present to participants an hypothetical situation identical to the game described above. We ask each participant to indicate, for each of the 7 possible actions of B, the number of tokens $[0; 20]$ that in their opinion C would use to punish B. We explain that, once each participant present in the laboratory has provided her answers, the computer selects one of the seven possible actions of B. For the selected action, a participant earns 40 tokens if the number she indicated is equal, bigger or smaller by one unit to the average number indicated by all the participants in the experimental session. Therefore, in this part of the experiment participants have incentives to reveal their true beliefs regarding peers' choices of punishment.

Treatments. We propose 2 effect treatments (INFORMATIONAL and NORMATIVE) and a control treatment (CONTROL). The elicitation of beliefs and the first period of the game are identical in all treatments. Specifically, the amount of tokens B decides to take from A in the first stage is not immediately observed by C. Instead in stage 2 of the first period C’s decisions of punishment are elicited employing the “strategy method”: for each possible action of B, C states his decision of punishment. Participants are informed that only the punishment decision corresponding to the actual choice made by B determines payoffs. The punishment tokens used by C in correspondence of the other possible choices of B do not have payoff consequences. First period earnings and choices of peers are not revealed to participants at the end of the first period.

At the beginning of the second period, participants’ endowments are restored at the initial level. Earnings of the first period are independent from those of the second. The first stage of the second period is identical to the first stage of the previous one: B is endowed with the same amount of tokens and may take part or all of A’s endowment.

Also in the second stage of the second period, C has to indicate the level of punishment inflicted for each possible action of B. The difference between treatments consists in the kind and amount of information disclosed to participants C before the punishment choices. In the INFORMATIONAL treatment, each participant C receives information about the average number of tokens used to punish B in the first period by the participants C taking part to the experimental session.

In the NORMATIVE treatment, each participant C receives the same information of INFORMATIONAL. However, she is additionally informed that her punishment decisions of the second period will be revealed to 5 peers randomly selected among the experiment participants. After observing these choices, the 5 peers vote for sending an emoticon that will appear to the screen of the participant C. The 5 peers could vote for a smiling emoticon or

a sad emoticon. If the majority vote for a smiling emoticon, on player C's monitor will appear a smiling emoticon. A sad emoticon will appear on the screen otherwise. Participants are informed that the emoticon received has no effect on earnings and that it disappears after one minute.

In the CONTROL treatment, no relevant information about participants' punishment choices is disclosed in the second period. However, we have to rule out the possibility that a change in punishment behavior between periods is driven by factors other than the exposure to social influence. One possible confounding factor is subjects' experience that increases between periods. Another possible confounding factor is that processing new information imposes a cognitive effort to subjects. Hence these factors, not the social content of the information received, could be responsible for a modification of the punishment choice. In fact, in the INFORMATIONAL and NORMATIVE treatments subjects have to process some sort of information, and there is evidence that individuals exposed to a cognitive load tend to modify their behavior (for discussion on this point see Cason and Mui, 1998).

In order to rule out these confounding factors and isolate social influence effects, we then expose the CONTROL group to some social irrelevant information. Specifically, we ask at the beginning of the session to each participant her day of birth $\in [1; 31]$ and we take the average. Since we do not ask nor we report them neither the year nor the month of birth, reporting this measure do not convey any relevant social information. However, in this way participants in CONTROL are affected by the same cognitive burden of participants in TREATED and the only difference lies in the exposure to relevant social information .

Endowments, Choice Sets and Payoffs.

Participants's initial endowment in each period and treatment is 30 tokens. The initial endowment is restored at the beginning of each period. Earnings

of the first period are independent from those of the second one. In each period, first B decides how many tokens to take from A. B could take from 0 up to 30 tokens in multiples of 5 from A. Then, C decides how many tokens to use for punishing B. C could use from 0 up to 20 tokens of his initial endowment. For participants C, the cost of reducing the payoff of B of 4 tokens is 1 token. Only punishment of integer tokens is allowed. Participants are informed that eventual negative earnings would be deducted from the participation fee. C has to take 7 punishment decisions. In fact, C does not observe the actual choice of player B. Instead, C reports his punishment choice for each of the 7 possible B's actions.

In all treatments, the per period payoffs are calculated as:

- $\Pi_A = 30 - t$
- $\Pi_B = 30 + t - 4 \cdot p$
- $\Pi_C = 30 - p$

where:

- t = tokens taken by B from A
- p = punishment tokens used by player C

Results and earnings of the beliefs elicitation game and of the first period of the dictator game are not revealed to subjects until the end of the experiment. In order to calculate individual earnings, participants are randomly divided in groups of 3. Each group is composed of one participant A, one B and one participant C. The final payment for each group follows this procedure:

- In the dictator game, after the second period is concluded, one of the two periods is randomly selected. This period will be called the "payment period".

- For each participant, earnings relative to the payment period are added to earning collected in the beliefs elicitation game. Earnings from the non-selected period are not paid out.
- A 5 euro participation fee is added to total payments.

Given the experimental design, we are able to isolate the effect of *endogenous interactions* in the form of *preferences interactions* on TPP. This is possible because we randomly select and assign participants to roles, we exclude payoff complementarity among third-parties and we explicitly present them a choice set that remains unchanged throughout the experiment. we also feel confident that our design rules out the possibility that individuals are influenced by "epistemic norms" (Hetcher, 2004). Epistemic norms emerge when agents experience scarcity of information and so just "follow the crowd" in taking a decision. To be precise, epistemic norms are not even norms but conventions motivated by simple self-interest (Kahan, 1997). Independently from how these norms are named, we believe they are not playing a role in the experiment, since every player has perfect information about possible actions in his choice set, and payoff complementarity and strategic actions are ruled out by design.

The Procedure. The experiment was programmed using the software z-Tree (Fischbacher, 2007). Every session was conducted at the Bologna Laboratory for Experimental Social Sciences at the University of Bologna, Italy, between November and December 2013. Participants were for the vast majority graduate and undergraduate students of the University of Bologna, plus some private citizens, and were recruited through the online system ORSEE (Greiner, 2004). In each session participants were split into 5 groups of 3 subjects². Overall, 9 sessions were run, 3 for each treatment, that results in

²In one session of the INFORMATIONAL treatment there were only 4 groups, for a total of 12 participants.

a total of 132 participants (56% female).

In each session, before each of the three part of the experiment (elicitation of beliefs, first period punishment and second period punishment), a printed copy of the instructions was distributed and read aloud by the experimenter³. Participants had additional images and tables summarizing the instructions on their computer screen. Information regarding payoff functions and rules of the game was common knowledge. Participants had the possibility to ask questions before the experiment started.

At the end of each session participants completed a brief socio-demographic questionnaire⁴. Each participant took part in one session only. Peers' identities were maintained unknown even after the end of the experiment. In order to guarantee anonymity, participants were individually and privately paid after the experiment finished. No communication among participants was allowed.

The part of the session concerning beliefs elicitation and treatments lasted around 20 minutes. However, due to the impossibility of learning throughout periods and the limited number of decisions each participant had to take, we were concerned about the possibility that instructions may were not fully understood. In order to minimize this possibility, we adopt special care in writing detailed instructions and providing multiple examples, and we also asked subjects to correctly answer control questions before proceeding with each part of the experiment. As a result, each experimental session lasted in total about 45 minutes. Tokens where converted into euro at a rate of 5 tokens for 1 euro. Subjects earned on average 11 euros for experimental session.

³Original instructions are in Italian and are available upon request. A copy of the instructions for the NORMATIVE treatment translated in English is included in Appendix B.

⁴In one session of the INFORMATIONAL treatment subjects' socio-demographic characteristic were not recorded due to a technical problem.

4. Hypothesis

Following the customary assumptions, the predictions of the game outcomes are straightforward. Agents' utility is an increasing function of individual wealth and agents are individual payoff maximizer. Hence, in any treatment, no punishment should be observed, since the payoff-maximizing strategy for third-parties is to punish nothing and keep the initial endowment. Anticipating the absence of punishment, dictators should take all the tokens from receivers.

However past dictator game experiments have shown two behavioral regularities. On one hand, even in games where the dictator faces no threat of punishment, positive amounts of tokens are transferred (in our setting: are left) to the receiver. On the other hand, third-parties engage in costly punishment for dictator's levels of transfer (in our setting: for dictator's levels of taking) perceived as unfair. In this study we are interested in verifying how, given the action of a dictator, the punishment choices of other third-parties affects the utility that a bystander derives from punishing the dictator.

Consider the choice of a third-party *we* to use p tokens of her initial endowment in order to punish a dictator that takes z tokens from a passive receiver. Third-parties' individual utility is an increasing function of the final monetary earnings x . Moreover, given a dictator's action, third-parties have some inherent preferences p_z^k for the amount of tokens she wants to use for punishment. $p_{i,z}^k$ could be interpreted as reflecting the individual sense of justice of the third-party *we*. If a third-party chooses to punish the dictator a quantity different from her inherent preference, she has to bear a cost s that increases when the absolute difference between p^k and the p increases. Furthermore, third-parties have some beliefs $E(\bar{p})$ regarding the average amount of tokens that the other bystanders will use for punishing dictators. A third-party incurs a cost c for punishing a quantity of tokens different from $E(\bar{p})$, and this cost becomes larger when the absolute difference between individual punishment and the average punishment of the peers increases. c

incorporates both the costs imposed by the other bystanders observing the third-party that deviates from the average punishment and the disutility the third-party experiences in not conforming with the peers' behavior independently from the fact that her action is observed.

Therefore, in their punishment decisions a third-party maximizes individual utility taking into account the cost of using tokens for punishing a dictator and so reducing her monetary payoff, the cost for deviating from her inherent preference for punishment and the cost of not conforming to the peers' average punishment:

$$\begin{aligned} \max_{p_{i,z,t}} \quad & U_{i,z,t} = x_{i,z,t} - (s(E_i(\bar{p}_{z,t}) - p_{i,z,t})^2 + c(p_{i,z}^k - p_{i,z,t})^2) \\ \text{s. t.} \quad & y_w e = x_{i,z,t} + p_{i,z,t} \end{aligned} \quad (1)$$

Where y is third-party's initial endowment. Assuming an interior solution exists, equation (1) generates the following first order conditions:

$$p_{i,z,t}^* = \frac{sE_i(\bar{p}_{z,t}) + cp_{i,z}^k - 1}{s + c} \quad (2)$$

Therefore, according to our model of social influence the optimal punishment choice of a third-party is an increasing function of the expected punishment chosen by her peers. Furthermore, the higher is for a third-party the cost s of not conforming to other bystanders' average punishment relative to the cost c of deviating from inherent preferences, the higher will be the tendency to conform to the peers' average punishment. Allowing for concavity of agents' utility function, the intuition of the previous results will still work.

In order to test our predictions, as a first step we verify if there is a positive association between a third-party's beliefs regarding peers' average punishment and her first period punishment. As a second step, we then investigate how participants modify their punishment choices between first and second period. Assume that third-parties in TREATED revise beliefs about average peers' punishment substituting their initial priors with the actual punish-

ment level revealed them after the first period, hence $E_i(\bar{p}_{z,t}) = \bar{p}_{z,t-1}$. The punishment variation across periods is given by:

$$(p_{i,z,2}^* - p_{i,z,1}^*) = \frac{s(\bar{p}_{z,1} - E_i(\bar{p}_{z,1}))}{s + c} \quad (3)$$

For the moment, focus only on the distinction between participants in CONTROL and the other participants grouped together, that we call TREATED. Third-parties in CONTROL are not exposed to social relevant information between period 1 and 2. Instead, bystanders in TREATED are exposed to information that may induce them to update their initial beliefs regarding peers' average punishment and so influence their second period punishment decision. Therefore, if social influence has an effect on third-party punishment decision, we expect participants in TREATED to be more likely to modify their punishment decisions in the passage between the first and second period punishment compared to participants in CONTROL.

Thus, according to our model revealing to a bystander her peers' average punishment may trigger a change in her second punishment decision as a consequence of a beliefs updating process. Specifically, for a bystander the likelihood to change punishment decision in the second period increases when the absolute difference between her beliefs regarding peers' average punishment and the actual average punishment of the first period is large. Therefore, we test the following hypothesis:

1. Zero Social Influence hypothesis: *In the first period, punishment decisions of bystanders are not influenced by their beliefs regarding peers' average punishment. Moreover, bystanders in TREATED are as likely as bystanders in CONTROL to modify their initial punishment decisions.*

Second, we want to identify who are the bystanders more responsive to social influence. Third-parties deciding to use tokens for punishing a dictator are

reducing their final monetary payments. Hence, every time we observe a bystander punishing a positive amount, according to our model we infer that $sE(\bar{p}) + cp^k - 1$ is positive. This could mean that the bystander has inherent preferences for punishing a positive amount ($p^k > 0$) and at the same time she attaches a positive weight to this component of the utility function ($c > 0$). However, it is also possible that the bystander attaches a positive weight to the social component of the utility function ($s > 0$) and she expects peers to punish on average a positive amount of tokens ($E(\bar{p}_{z,t}) > 0$)⁵. If this last possibility is true, the higher is a bystander's punishment in the first period the more she attaches weight to the social component of the utility function and so the more likely she is to modify second period punishment decision. Now consider the difference between first and second period punishment of a bystander. Inherent preferences for punishment are stable, so they do not play a role in the decision to eventually modify punishment choice. Instead, according to the prediction of our model, the larger is s for a bystander, the more she responds to the social information regarding peers' punishment. Therefore, holding constant $E(\bar{p}_{z,t}) - \bar{p}_{z,t-1}$, we expect that the more a bystander punished in the first period, the more she is likely to revise her punishment decisions in the second period.

Moreover, we also consider the difference between a bystander's beliefs regarding peers' average punishment and her first period punishment. In the first period, a bystander could punish an amount different from her beliefs regarding peers' average punishment because she only cares about her monetary payoff or because her inherent preference for punishment differs from the expected average punishment and the cost s of non conforming to peers' average punishment is small compared to the cost c of non following inherent preferences. In both cases, the choice of the bystander reveals that in her punishment decisions she is little influenced by peers' behavior. As a

⁵Of course, it is possible that what it is observed is a combination of this two possibilities.

consequence, we expect the more a bystander punishes in the first period a quantity different from her beliefs about peers' average punishment, the less she will be responsive to social influence.

2. Differential Social Influence hypothesis: *Third-parties that engage in high punishment in the first period are the most responsive to social influence. Conversely, the higher the absolute difference between a bystanders' beliefs regarding average peers' punishment and first period individual punishment, the lower the bystander likelihood to modify punishment choices in the second period.*

Finally, we investigate the psychological mechanisms triggering social influence. In our experiment, we give bystanders in NORMATIVE and INFORMATIONAL the same information about peers' punishment. However, in the INFORMATIONAL treatment, the second period choices of the third-party are not observable ex post by other participants. As a consequence, in the INFORMATIONAL treatment a third-party has no incentives to conform to peers' punishment choices if her only goal is being liked by them. Hence, a bystander would eventually modify his punishment strategy only if informational social influence is at work.

On the other hand, in the NORMATIVE treatment a bystander is aware that her punishment decisions of the second period will be observed by peers and that they will express a judgement regarding those choices. As a consequence, for bystanders in NORMATIVE the cost s of not conforming to the average peers' punishment has been modified in the passage between first and second treatment since we added a normative social influence component. Hence, if some bystanders are responsive to normative but not to informational influence, the NORMATIVE treatment will show social influence effects different from those resulting from the INFORMATIONAL treatment. The difference

in the way bystanders modify their punishment decisions between NORMATIVE and INFORMATIONAL treatments isolates the effect of normative social influence on third-party punishment.

3. Equivalence of Normative and Informational Influence hypothesis: *Social influence effects on third-party punishment are the same for subjects exposed to informational and normative influence.*

5. Results

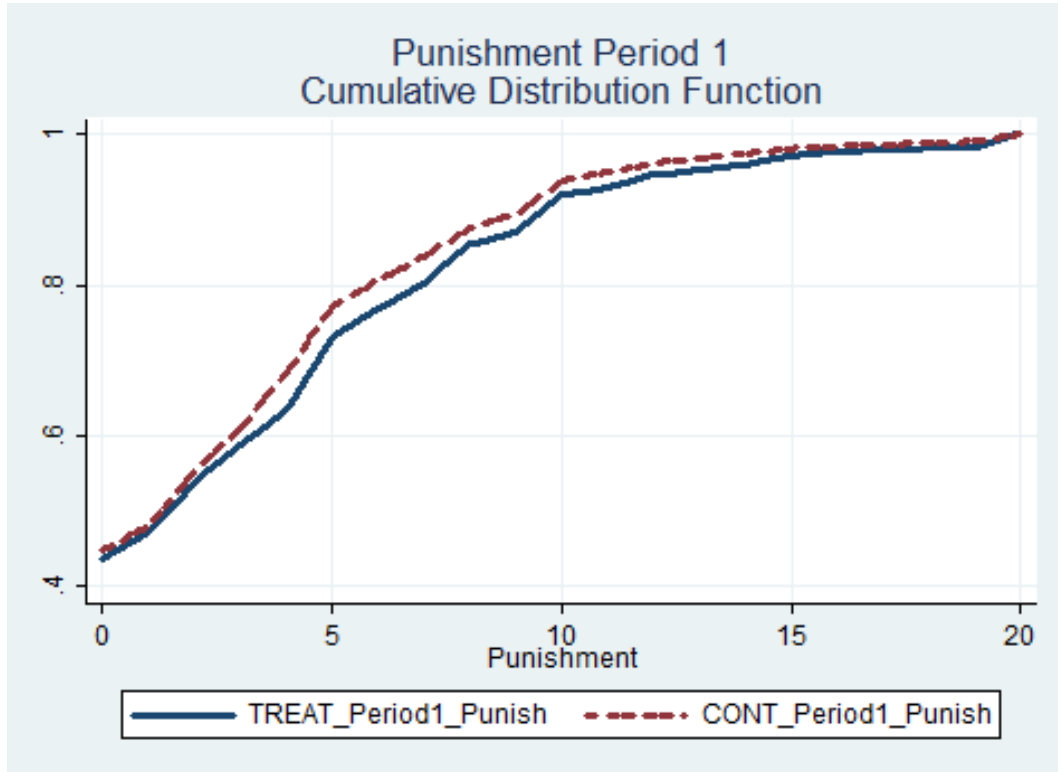


Figure 1: Punishment Period 1 Cumulative Distribution Function

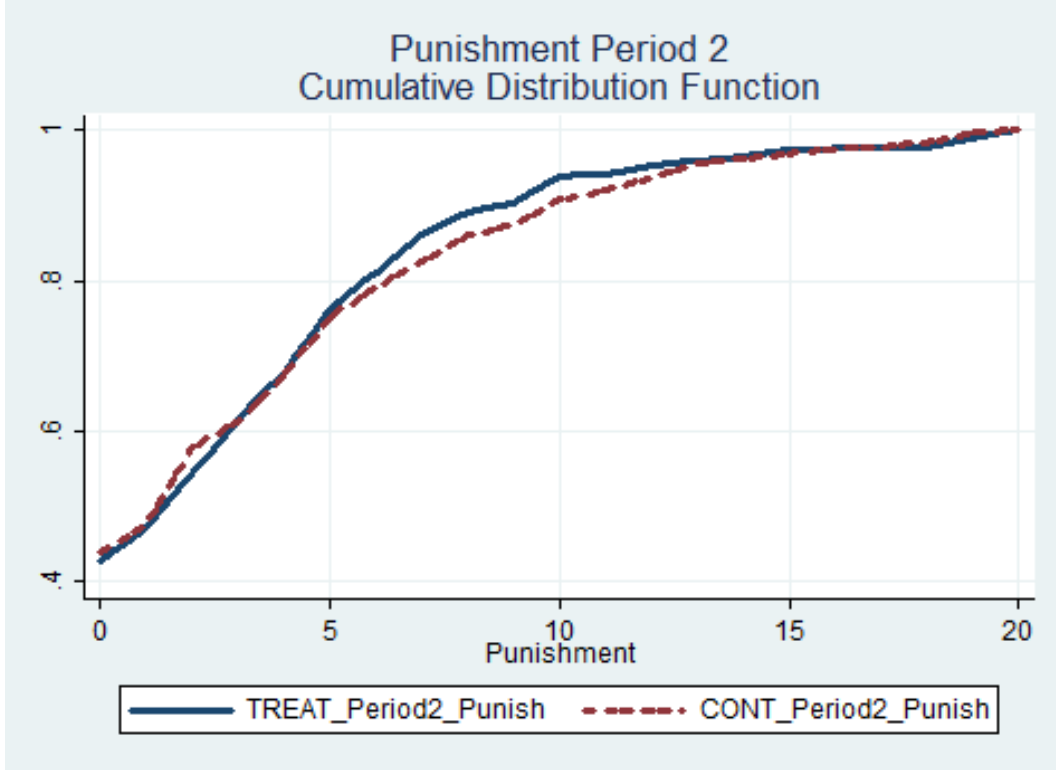


Figure 2: Punishment Period 2 Cumulative Distribution Function

Table 1 reports summary statistic relative to our data⁶. Dictators leave approximatively 36% of receivers' endowment. This finding is consistent with results from other comparable experiments where dictators have to take tokens from the endowment of a passive receiver (List, 2007; Krupka and Weber, 2013)⁷.

⁶Additional summary statistics where we consider separately the 7 possible punishment choices in each period, are reported in Tables A.7 and A.8 in the appendix.

⁷In the classical dictator game without punishment a dictator has the possibility to give part of his endowment to a passive receiver. In a meta-study, Engel (2011) found that on average dictators give roughly 25% of their endowment to the receiver. However, in our design dictators has to *take* money from receivers' endowment instead of *giving* them. This difference and the possibility of being punished that characterizes our design are

On average bystanders punish approximately 3.5 tokens, decreasing punishment amount in the second period. When the dictator takes all the money from the receiver, third parties spend approximately 6 tokens in punishment. Average punishment then progressively decline, reaching virtually 0, when levels of dictators' taking decrease. Also this result is consistent with previous findings on third-party punishment in dictator games (Fehr and Fischbacher, 2004b). However, if we consider the 3 treatments separately, we see that in CONTROL punishment slightly increases in the second period, while in both NORMATIVE and INFORMATIONAL it decreases. Considering third-parties' beliefs regarding peers' punishment behavior, beliefs are on average are higher than actual punishment.

we proceed considering, for each bystander in a single punishment period, the average of her 7 punishment choices corresponding to different levels of dictator taking. we compare the cumulative distribution of this measure in TREATED and CONTROL. Figures 1 and 2 report the cumulative distribution functions of punishment respectively in period 1 and 2. In the first period, the cumulative punishment choice distribution in CONTROL exceeds the distribution in TREATED for any possible punishment level. However, a Kolmogorov-Smirnov test cannot reject the null hypothesis that the distributions are equivalent. In the second period instead, the cumulative punishment choice distribution of TREATMENT exceeds the distribution of CONTROL for some punishment levels greater than 5. However, also in the second period a Kolmogorov-Smirnov test cannot reject the equivalence of the two distributions, nor the samples means are statistically different (t-test two sided p-value 66%). Now we test our hypothesis.

likely to explain the slightly more fair allocation we registered compared to the standard dictator game (on this point, see also Krupka and Weber, 2013).

5.1. Zero Social Influence Hypothesis

We start by investigating if bystanders punish in the first period according to their beliefs regarding the average punishment they expect peers' will use. As a first step, we test if there is a significant difference between the punishment used by a bystander and her beliefs about peers' average punishment. We conduct a t-test comparing the two averages under the null hypothesis that they are the same. Third-parties punish on average 3.6 tokens in the first period while their beliefs about peers' average punishment is 5.1 tokens.

Results of the t-test reject our hypothesis and indicate that bystanders in the first period punish significantly less than what they think peers on average will do (t-test two-tails, p-value < 1%).

We want to verify if this result is driven by those third-parties that during the experiment punish always 0 (we name them "selfish"). Excluding selfish punishers from the sample, bystanders' beliefs about peers' punishment remain higher than the punishment they provide (6.0 versus 5.4 tokens), however the difference is not statistically significant (p-value 12%). This results suggest that selfish subjects are responsible for the aforementioned gap.

we continue the analysis regressing the quantity of punishment tokens a bystander uses in the first period with her beliefs regarding peers' average punishment and a set of socio-demographic characteristics. Results are reported in Table 2⁸. The variable *Beliefs_Punish* indicates bystanders' beliefs about peers average punishment. The coefficient is positive and significant at the 1% level in any model specifications. According to the model estimation, bystanders spend additional 0.4 token for every unit of increase in expected average peers' punishment. Hence, data suggest that third-parties are influenced in their first period punishment decisions from beliefs about peers' punishment. This finding goes against the Zero Social Influence hypothesis.

⁸Table C.9 in Appendix C reports a description of each variable employed.

Table 1: Summary Statistics

Treatment	male	age	dictatorTake	Beliefs	PunishPer1	PunishPer2
Control						
(Mean)	.33	24.68	18.38	5.08	3.33	3.54
(Median)	0	25	20	4.29	2.71	2.71
(SD)	.48	2.76	11.19	3.71	3.42	3.82
Normative						
	.58	26.18	18.28	4.80	3.30	3.03
	1	25	20	4.57	3	2.29
	.50	5.38	11.40	3.40	3.34	3.37
Informational						
	.37	25.15	17.32	5.41	4.15	3.81
	0	25	16.25	4.86	4.29	3.86
	.49	3.72	10.85	3.67	3.71	3.90
Total						
	.44	25.37	18.01	5.09	3.58	3.45
	0	25	20	4.71	3.43	3.07
	.50	4.17	11.08	3.57	3.48	3.69

we proceed in the analysis verifying how third-parties in CONTROL and in TREATED modify punishment choices between first and second period. As a first step, we sort bystanders into 2 main categories: those who never change punishment decisions across periods and those who change at least once. In the TREATED group 53 subjects (61%) change at least one punishment decision between periods, while in the CONTROL treatment 24 subjects (53%) do so. This difference is not statistically significant. If we repeat the same test excluding selfish bystanders, it turns out that in TREATED 87% and in CONTROL 80% of third-parties change punishment decisions at least

Table 2: Determinants First Period Punishment

	(1)	(2)
Beliefs	0.405*** (0.05)	0.432*** (0.06)
male	-0.515 (0.58)	-0.407 (0.62)
age	-0.078 (0.06)	-0.047 (0.07)
degree	-0.842** (0.41)	-1.435*** (0.45)
worker	0.759 (0.74)	0.860 (0.80)
social	0.378 (0.73)	0.399 (0.85)
arts	0.234 (1.06)	0.021 (1.28)
field_other	0.409 (0.63)	-0.131 (0.75)
risk	0.138 (0.10)	0.037 (0.11)
logic	-0.305 (0.40)	0.057 (0.47)
impulsivity	-0.636** (0.27)	-0.687** (0.32)
Instruction	0.000 (0.00)	-0.000 (0.00)
DictatorTake	-0.007 (0.01)	0.026 (0.02)
_cons	6.246*** (1.90)	7.889*** (2.04)
N	924	630
R^2	0.281	0.275
BIC	5203.7	3602.6

Notes: OLS regression: dep. var. *Strat_Punish*,
SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

once. However, also in this case the difference is not statistically significant. We also verify how many times on average each punisher change decision across periods. In TREATED bystanders change decision 2.5 times, while in CONTROL they change 2.1 times. This difference is not statistically significant and it remains roughly unchanged even if we exclude selfish bystanders. Therefore, these results do not provide evidence against the Zero Social Influence hypothesis. The result seems to be driven by the high percentage of participants (53%) in the CONTROL group that modifies punishment choices at least once, even if they did not receive any relevant social information.

As a second step, we test if there is a difference in the likelihood that participants in CONTROL and TREATMENT change punishment decisions. We create the dummy variable *DummyP1p0* that takes the value 1 when punishment in the second period differs from punish in the first one and 0 otherwise.

We implement a logistic regression to estimate the likelihood of changing punishment choice across periods. Results of the model are presented in Table 3. The dummy *TREATED* equal to 1 for participants in NORMATIVE and INFORMATIONAL. The coefficient of the dummy is positive and statistically significant in any of the model specifications⁹. Therefore, we conclude that the results of the logistic regression do not support the Zero Social Influence hypothesis and indicate that participants in TREATED modify punishment decisions across periods more often than those in CONTROL.

As a third step, we investigated how third-parties modify their punishment choices. In CONTROL bystanders reduce punishment in the second period 48 times (15%), increase 49 times (16%) and do not change 218 times (69%). In TREATED, bystanders reduce punishment 140 times (23%), increase 93

⁹Model 2 differs from Model 1 because excludes selfish participants. Model 3 add the control variables *Strat_Punish*, indicating punishment exerted in the first period, and *Abs0Belifs*, reporting the absolute difference between a bystander’s punishment in the first period and her beliefs regarding peers’ average punishment. Model 4 excludes selfish participants from the sample.

(15%) and do not change 376 times (62%). This choices result for CONTROL in an average increase in punishment from period 1 to period 2 of 0.21 tokens (from 3.3 to 3.5) and in an average decrease in TREATED of 0.30 tokens (from 3.7 to 3.4). The mean punishment difference across periods is not statistically different in CONTROL and TREATED (p-value 0.16, t-test two-sided).

However, we could expect that bystanders have no reason to punish a dictator when she does not take any amount of money from the receiver. Hence, when the dictator chooses to take 0 from the receiver, we expect little or no punishment both in CONTROL and TREATED. In fact, if we exclude the situations in which the dictator takes 0 from the receiver, the average difference between bystanders' punishment in the two periods is weakly statistically significant higher in CONTROL versus TREATED (p-value 0.09). Furthermore, if we consider only situations in which the dicator takes half or more of receiver's initial endowment, this difference between CONTROL and TREATMENT becomes significant at the 5% level.

Hence, results of this third set of tests suggest that, at least for situations where dictators subtract positive amounts of tokens from receivers, bystanders exposed to social influence significantly reduce the amount of punishment provided compared to bystanders in CONTROL. These results provide evidence against the Zero Social Influence hypothesis.

Fourth, we test the hypothesis that a large absolute difference between a bystander's beliefs regarding peers' average punishment and actual peers' average punishment increases the likelihood to modify the initial bystander punishment choice. Third-parties receive information regarding actual peers' punishment in the NORMATIVE and INFORMATIONAL treatments only, so we restrict the analysis to these treatments. we test this hypothesis using a logistic model. Results are reported in Table 3. From models 7 and 8, we can see that the coefficient of the variable *Abs_BeliefAvgPunish* is positive as expected, however only weakly significant. The estimations suggest that

an increase of one unit in the difference *Abs_BeliefAvgPunish* increases on average the probability of modifying second period punishment by 3.5%¹⁰. This result provides evidence against the Zero Social Influence hypothesis. Finally, we report a series of statistical tests that account for the direction of punishment deviation between periods. First, we consider the variable *P1p0* that reports the difference between punishment in period 2 and period 1. Table 5.1 provide summary statistics of this variable, grouping subjects according to the difference between individual beliefs about average punishment and actual average punishment observed. When beliefs match exactly the average punishment observed ($P1p0_BAP = 0$), subjects confirm punishment choices in the second period 76% of times. Instead, when beliefs are larger or smaller than the actual average punishment observed ($P1p0_BAP > \text{and} < 0$ respectively), subjects confirm first period choice respectively 51% and 73% of times. Considering how subjects modify their decisions, we see that those observing actual average punishment smaller than beliefs reduce on average punishment in the second period of 0.30 tokens. Instead, subjects observing actual average punishment equal to beliefs reduce punishment of 0.14 tokens between the two periods, and those observing average punishment smaller than beliefs increase punishment in the second period of 0.27 tokens.

Additionally, we want to test if subjects that receive information regarding session average punishment modify their second period punishment decisions in order to conform to peers' average punishment. We consider each participant's punishment decision in the first and in the second period (

¹⁰we also consider the possibility that a large absolute difference between a bystander punishment in the first period and the average peers' punishment increases the likelihood to change punishment decision in the second period. We create the variable *Abs_Signalp0*, reporting the absolute difference between individual punishment in the first period and average peers' punishment. Results of the logistic estimations are reported in model 7 and 8 of Table 3. The coefficient of *Abs_Signalp0* is not statistically different from 0 in any model specification. As a consequence, we conclude that *Abs_Signalp0* has no impact on subjects' likelihood to modify punishment decision.

Table 3: Probability modify punishment across periods

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
male	-0.219*** (0.08)	-0.258*** (0.09)	-0.168** (0.07)	-0.215** (0.09)	-0.098 (0.12)	-0.019 (0.14)	-0.089 (0.12)	-0.017 (0.14)
risk	0.030* (0.02)	0.014 (0.02)	0.019 (0.02)	0.008 (0.02)	-0.015 (0.02)	-0.040 (0.03)	-0.020 (0.02)	-0.043* (0.03)
logic	-0.077 (0.06)	-0.033 (0.07)	-0.061 (0.06)	-0.031 (0.07)	0.135 (0.09)	0.242** (0.10)	0.149 (0.09)	0.246** (0.10)
TREATED	0.155** (0.06)	0.131* (0.07)	0.124** (0.06)	0.122* (0.07)				
Strat_Punish			0.040*** (0.01)	0.021** (0.01)	0.044*** (0.01)	0.025** (0.01)	0.047*** (0.01)	0.029** (0.01)
Abs_p0Belifs			-0.010 (0.01)	-0.003 (0.01)	-0.026** (0.01)	-0.019* (0.01)	-0.032** (0.01)	-0.026 (0.02)
Abs_Signalp0							0.002 (0.01)	0.006 (0.01)
Abs_BelAvgPun							0.024 (0.02)	0.023 (0.03)
Beliefs			0.015 (0.01)	0.009 (0.01)	0.025** (0.01)	0.021* (0.01)	0.015 (0.01)	0.009 (0.01)
Other contr	Y	Y	Y	Y	Y	Y	Y	Y
N	924	630	924	630	609	420	609	420
pseudo R^2	0.092	0.059	0.207	0.095	0.206	0.101	0.226	0.123

Notes: Logistic regression: dep. var. *DummyP1p0*, marginal effect at means, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01. Other Controls include: degree worker social arts field_other DictatorTake impulsivity age.

$P1$ and $P2$ respectively) and the information regarding session average first period punishment that the TREATED group receive between first and second period (*AvgPunish*). If a subject wants to conform her punishment behavior to the punishment choices of the peers, she will choose a second period punishment $P2$ that reduces the gap between first period punishment $P1$ and the revealed peers' average punishment *AvgPunish*. Therefore, we create the variable *Diff_Deviation*, that compares for each punishment choice the absolute value of the difference between participants' first pe-

riod average punishment and individual punishment in the second and in the first period: $Diff_Deviation = |P2 - AvgPunish| - |P1 - AvgPunish|$. In CONTROL, where subjects are not exposed to information regarding peers' punishment behavior, $Diff_Deviation$ increases of 0.20 unit. Conversely, $Diff_Deviation$ is on average reduced of 0.25 unit for subjects belonging to the TREATED group. A t-test test comparing the mean values of $Diff_Deviation$ between groups under the alternative hypothesis that this measure is smaller in TREATED than in CONTROL reports weakly statistical evidence against the null hypothesis (t-test one side, p-value 0.06).

Table 4: Punishment difference across periods

Treatment	P1p0	P1p0_BAP>0	P1p0_BAP=0	P1p0_BAP<0
Informational	294	142	27	125
(mean)	-.33	-1.04	-.11	.42
(sd)	3.28	3.96	.42	2.52
Normative	315	170	60	85
	-.28	-.49	-.15	.06
	2.58	3.09	1.63	1.90
Total	609	312	87	210
	-.30	-.74	-.14	.27
	2.94	3.52	1.37	2.29

Notes: Variable $P1p0$ indicate the difference between punishment in first and second period. $P1p0_BAP >, <, = 0$ indicate respectively $P1p0$ when individual beliefs regarding average punishment are $>, <, =$ actual average punishment.

Finally, we test the hypothesis that subjects conform to peers' average punishment by means of a logistic regression. First, we create the dummy variable *ConvergeDummy* that takes value 1 when $Diff_Deviation$ is positive, that is when a subject chooses a second period punishment that reduces the gap between her first period punishment and peers' average punishment. Then we implement a logistic regression to test if the probability of reducing

this gap is the same for participants in CONTROL and TREATED groups. Results are reported in Table 5. As we can see from model (1), the coefficient of the dummy TREATED is positive and statistical significant at the 1% level. This result suggests that, if compared to participants in the CONTROL group, subjects exposed to information regarding peers' average punishment are roughly 14% more likely to modify their second period punishment choice in order to conform to peers' punishment behavior.

- Conclusion relative to the Zero Social Influence hypothesis: *Bystanders' punishment choices in the first period are positively associated with their own beliefs about average peers' punishment. However, we find that on average bystanders punish less than the expected average peers' punishment. This result seems to be driven by those bystanders that always decide not to punish dictators. We also found evidence that subjects in TREATED are more likely to change punishment decision across periods. Moreover, in CONTROL the amount of punishment in the two periods remains constant, while in TREATED it decreases. The mean punishment difference across periods is statistically higher in control than in treated if we consider situations where dictators take positive amounts from receivers' endowment. Additionally, a large absolute difference between a bystander's beliefs regarding peers' punishment and actual peers' average punishment increases her likelihood to modify punishment decisions across periods. Finally, if compared to participants in CONTROL, subjects in TREATED choose more often a second period punishment that reduces the gap between individual first period punishment and peers' average punishment. These results provide evidence against the Zero Social Influence hypothesis. Hence, we conclude that the Zero Social Influence hypothesis is not supported by the results of the experiment.*

Table 5: Probability Second Period Individual Punishment Converges Toward Average Participants' First Period Punishment

	(1)	(2)	(3)	(4)	(5)
TREATED	0.144*** (0.05)				
NORMATIVE		0.021 (0.08)	-0.018 (0.08)	0.201* (0.11)	0.226** (0.10)
male	-0.089* (0.05)	-0.082 (0.09)	-0.089 (0.08)	-0.195 (0.12)	-0.224** (0.11)
Strat_Punish	0.047*** (0.01)	0.046*** (0.01)	0.049*** (0.01)	0.020*** (0.01)	0.041*** (0.01)
Beliefs	0.000 (0.01)	0.004 (0.01)	0.002 (0.01)		0.010 (0.01)
Punish	-0.033*** (0.01)	-0.023* (0.01)	-0.025* (0.01)		-0.030** (0.01)
Abs_BeliefAvgPunish	0.003 (0.01)	0.014 (0.02)	0.032 (0.02)		-0.010 (0.01)
Abs_p0Belifs			-0.045*** (0.02)		-0.012 (0.01)
NormAbs_BeliefAvgPunish			-0.025 (0.02)		
NorAbs_p0Belifs			0.044** (0.02)		
Other contr	Y	Y	Y	Y	Y
<i>N</i>	924	609	609	197	197
pseudo <i>R</i> ²	0.195	0.163	0.185	0.198	0.310

Notes: Logistic regression: dep. var. *ConvergeDummy*, marginal effect at means, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01. Other Controls include: age risk impulsivity logic student.

5.2. Differential Social Influence hypothesis

First, we want to verify if bystanders that engage in less punishment in the first period are also less responsive to social influence. Third-parties receive relevant social information in the NORMATIVE and INFORMATIONAL treatments only, so we restrict the analysis to these treatments. For each of the 7 dictators' decisions, we characterize bystanders that punish in the

first period above the median as "high punishers"¹¹. For each transfer level considered, the percentage of third-parties modifying punishment decision across periods among high punishers is almost double than among the other punishers. If we exclude selfish bystanders we still have similar results.

we test the hypothesis implementing a logistic model. We estimate the probability of modifying punishment decision including the independent variable *Strat_Punish* that reports the level of punishment provided in the first period. Results are reported in Table 3. In any model specification, the coefficient associated with *Strat_Punish* is positive and significant at the 1% level. The coefficient of *Strat_Punish* suggests that, holding constant at their means the other controls, a bystander spending 1 additional token in first period punishment is 3% to 5% more likely to revise her punishment choice in the second period.

Hence, we conclude that this first set of results supports the Differential Social Influence hypothesis.

Second, we want to verify if bystanders that choose to punish in the first period a quantity different from their beliefs regarding average peers' punishment are less responsive to social influence compared to the other bystanders. we implement a logistic regression estimating the probability that a bystander modifies punishment decisions across periods. As independent variable, we introduce *Abs_p0Belifs*, the absolute difference between a bystander's beliefs regarding average peers' punishment and her individual punishment in the first period. We report results in Table 3.

The coefficient of *Abs_p0Belifs* is negative and statistically different from 0 in model specifications 7 and 8, in which we include all the controls¹². The estimations suggest that increasing of one unit the absolute difference

¹¹Results are substantially the same if we choose the average punishment as a criterion for classification.

¹²In model 8 we exclude from the analysis selfish bystanders and the coefficient becomes only weakly significant.

Abs_p0Belifs decreases for a bystander the probability of modifying punishment decision across periods of approximately 3%.

Therefore, we conclude that also this second set of results supports the Differential Social Influence hypothesis

- Conclusion relative to the Differential Social Influence hypothesis: *There is evidence that the more a bystander punishes in the first period, the more she is responsive to the social information received. We also find evidence that the larger is the absolute difference between a bystander's beliefs regarding peers' average punishment and her first period punishment, the less likely is that she modifies punishment decisions across periods.*
Therefore, we conclude that the results of our experiment support the Differential Social Influence hypothesis.

5.3. Equivalence of Normative and Informational Influence hypothesis

We conclude this section reporting results on the difference between bystanders exposed to both normative and informational social influence and those exposed only to the latter. From the summary statistics reported in Table 1, we could see that in the first period third-parties in INFORMATIONAL punish on average 4.1 tokens versus 3.3 of those in NORMATIVE. This difference is not statistically significant (p-value 26%). In both treatments, on average bystanders reduce punishment between first and second period: NORMATIVE of 0.28 tokens and INFORMATIONAL of 0.33. Also this difference is not statistically significant.

We test the hypothesis that third-parties in NORMATIVE are more likely to revise their second period punishment decisions. We create the dummy variable *NORMATIVE* equal to 1 for third-parties in the normative treatment and we implement a logistic regression. The dependent variable is *DummyP1p0* equal to 1 when punishment is modified across periods. Results are reported in Table 5.3.

Table 6: Probability Modify Punishment Across Periods: Treated Groups

	(1)	(2)	(3)	(4)
NORMATIVE	0.064 (0.11)	0.075 (0.09)	0.034 (0.17)	0.154 (0.15)
Strat_Punish			0.054*** (0.01)	0.039** (0.02)
Abs_p0Belifs			-0.085*** (0.02)	-0.091*** (0.03)
NorStratPunish			0.016 (0.02)	0.014 (0.02)
NorAbs_p0Belifs			0.103*** (0.03)	0.116*** (0.04)
Abs_Signalp0			-0.003 (0.02)	0.001 (0.02)
NorAbs_Signalp0			0.007 (0.02)	-0.003 (0.02)
Abs_BelAvgPun			0.094*** (0.02)	0.094*** (0.03)
NormAbs_BelAvgPun			-0.104*** (0.03)	-0.117*** (0.04)
DummySignalp0			0.064 (0.10)	0.152 (0.11)
NorDummySignalp0			0.095 (0.14)	-0.038 (0.17)
Other contr	Y	Y	Y	Y
<i>N</i>	609	420	609	420
pseudo <i>R</i> ²	0.092	0.092	0.316	0.221

Notes: Logistic regression: dep. var. *DummyP1p0*, marginal effect at means, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01. Other Controls include: male age degree worker social arts field.other risk logic impulsivity Instruction DictatorTake.

From the coefficient of *NORMATIVE* in model 1 to 4 we could see that, on average, there is no statistical difference between treatments in the likelihood of modifying punishment decision. Moreover, we test if participants in *NORMATIVE* and *INFORMATIONAL* choose a second period punishment

that reduces the gap between first period individual punishment and peers' punishment with the same probability. We implement a logistic regression and we report the results in Table 5. From the coefficient of the dummy variable NORMATIVE in models (2) and (3) we could see that there is no statistical difference between the two effect treatments.

However, when instead we disentangle the effect of individual determinants of the probability to modify punishment decision across periods it is possible to find differences between treatments. First, consider the tests we did for the Differential Social Influence hypothesis. Models 7 and 8 of Table 3 suggest that increasing one unit the absolute difference across punishment in the first period and individual beliefs regarding the average punishment in the session (the variable *Abs_p0Belifs*) decreases of approximately 3% the likelihood to modify punishment between period. However, the result is only weakly significant. Nevertheless, the estimation could be affected by the fact that in the models of table 3 we constrained the slope of *Abs_p0Belifs* to be the same for NORMATIVE and INFORMATIONAL. Therefore, in model 3 and 4 of Table 5.3 we introduce the interaction term *NorAbs_p0Belifs*, that isolate the effects of the absolute difference between punishment in the first period and beliefs about peers' average punishment for bystanders in the NORMATIVE treatment. The coefficient is positive and significant at 1% level for both model specification, and the coefficient of *Abs_p0Belifs* becomes negative and significant at 1% level. Interpreting the coefficients, we can see that for third-parties in NORMATIVE *Abs_p0Belifs* has no effect on the probability of modifying punishment across periods. Instead, for bystanders in INFORMATIONAL an increase of one unit in *Abs_p0Belifs* diminishes of roughly 8% the probability of modifying punishment across periods.

We can find a similar difference between treatments for the probability that subjects choose a second period punishment that decreases the gap between individual first period punishment and peers' average punishment. From model (3) in Table 5 we can see that the unconstrained coefficient of

$Abs_{p0}Belifs$ is negative and statistically significant at the 1% level. However, the coefficient of the variable $NorAbs_{p0}Belifs$ is positive and statistically significant. These results confirm that $Abs_{p0}Belifs$ has a statistically significant effect only for subjects in the INFORMATIONAL treatment. Therefore, the results contrast the Differential Social Influence hypothesis for bystanders in the NORMATIVE treatment, while the hypothesis finds support for subjects in the INFORMATIONAL treatment.

As a possible explanation for this difference across treatments, we conjecture that for bystanders in NORMATIVE there are additional incentives to revise their punishment decisions compared to bystanders in INFORMATIONAL. In fact, in NORMATIVE bystanders are told that their punishment choices of the second period will be revealed to other participants and that these peers will express their judgements. Therefore, it seems that the threat of revealing individual choices to other participants triggers the decision to modify first period punishment. In order to further investigate this conjecture, in models (4) and (5) of Table 5 we restrict the analysis to those punishment choices within the TREATED group where $Abs_{p0}Belifs$ is larger than the average. In fact, participants choosing a first period punishment largely different from their beliefs about average punishment show that they are not interested in conforming to peers' behavior and so they should be unaffected by informational social influence. On the other hand, subjects in NORMATIVE before stating the second punishment decision are told that peers will observe and judge their individual choices. Therefore, this additional exposure to normative social influence might induce some of the subjects unresponsive to informational social influence to conform with the peers' punishment behavior. Indeed, in models (4) and (5) the coefficient of the dummy *NORMATIVE* is positive and statistically significant. These estimations suggest that, within the subsample of subjects showing lack of interest for conformity with peers' punishment behavior, participants exposed to normative social influence are roughly 20% more likely to choose a second period punishment that reduces

the gap between first period individual punishment and peers' average punishment.

Furthermore, consider the absolute difference between a bystander's beliefs regarding peers' average punishment and the actual peers' average punishment. Models 7 and 8 of Table 3 indicate that increasing of one unit the coefficient of the variable *Abs_BeliefAvgPunish* increases for a bystander the probability to modify punishment decision across periods by 3%. However, this result comes from models where we constrained the coefficient of *Abs_BeliefAvgPunish* to be the same in NORMATIVE and INFORMATIONAL.

We verify if the coefficient is the same in both treatments estimating the effect of *Abs_BeliefAvgPunish* for the two groups separately. We do so interacting the variable *Abs_BeliefAvgPunish* with the dummy NORMATIVE and so creating the variable *NormAbs_BeliefAvgPunish*. From results of models 3 and 4 in Table 5.3, we can see that the coefficient of *NormAbs_BeliefAvgPunish* is negative and statistical significant at 1% level. On the other hand, the coefficient of *Abs_BeliefAvgPunish* in the unconstrained model becomes positive and significant at the 1% level, while it was only weakly statistically significant in the constrained model. Specifically, for subjects in the INFORMATIONAL treatment an increase of one unit in *Abs_BeliefAvgPunish* raises by approximately 9% the probability of modifying punishment across periods.

In order to further investigate this result, we check how bystanders in the two treatments modify their punishment choices across periods conditional to the sign of the difference between beliefs regarding peers' average punishment and actual peers' average punishment. Table 5.1 reports these summary statistics. Bystanders in both treatments reduce punishment in the second period when actual average punishment is lower than expected. However, bystanders in INFORMATIONAL on average reduce punishment of more than 1 tokens, while those in NORMATIVE of less than 0.5. Conversely, when

actual average peers' punishment exceeds a bystander's expectations, in INFORMATIONAL third-parties increase punishment of 0.4 tokens on average, while bystanders NORMATIVE do not modify punishment decisions.

It is possible that the lower variability registered in NORMATIVE derives from the fact that individual choices are observable by peers. We conjecture that in NORMATIVE bystanders refrain from modifying punishment decisions, in particular from reducing punishment, because of the disutility of being eventually judged and targeted with the "sad" emoticon by peers.

Finally, we also test if the slope of the variable *Strat_Punish* differs NORMATIVE and INFORMATIONAL. We created the variables *NorStrat_Punish*, isolating the effect of *Strat_Punish* for bystanders in the NORMATIVE treatment. As expected, results for these unconstrained models reported in Table 5.3 show that there is no statistical difference between treatments in the data.

- Conclusion relative to Equivalence of Normative and Informational Influence hypothesis: *We find mixed evidence regarding our hypothesis. On one hand, at an aggregate level the likelihood to modify punishment choices is the same in NORMATIVE and INFORMATIONAL. However, disentangling the determinants that push bystanders to modify punishment choices across periods, we find differences between the two treatments. Therefore, we conclude that the empirical evidence is mixed and the hypothesis is not fully supported by the data.*

6. Conclusions

Human organizations need mechanisms to enforce rules and regulations upon which are founded. On one hand, societies developed a centralized apparatus of enforcement for this purpose. However this centralized systems coexist with a decentralized practice of punishment carried on by members of the

societies itself. Understanding the nature and characteristics of decentralized punishment might help legal scholars and policymakers to design effective policies in a variety of situations. Therefore, which are the major drivers of decentralized third-party punishment is an important question for social scientists.

In this paper we examine through a laboratory experiment the effect of one of these drivers, social influence, on the punishment decisions of third parties. Scholars in psychology, law and economics underline the relevance of third party punishment for the cohesion of human societies (Fehr and Fischbacher, 2004b; Marlowe et al., 2008) and the importance of social influence in various fields of application (Bernheim, 1994; Turner, 1991; Kahan, 1997; Becker, 1991). However, to the best of our knowledge this paper is the first work that investigates empirically social influence effects within the framework of third party punishment.

In a modified dictator game, we elicit the punishment choices of third parties before and after having exposed them to information regarding the punishment behavior of their peers. We compare those choices with decisions made by bystanders not exposed to social relevant information. The main finding of this article is that social influence is a major driver of bystanders' decision to engage in third-party punishment. Results of the experiment show that third-parties receiving information about peers' punishment revise their punishment choices more often and on average reduce punishment across periods compared to bystanders exposed to social irrelevant information. This last effect seems to be driven by the fact that bystanders' beliefs regarding peers' average punishment are higher than the actual punishment peers exert. Indeed, consistently with the model predictions, the empirical analysis shows that the larger is the absolute difference between a bystander's beliefs about peers' average punishment and peers' actual punishment, the more likely is the bystander to revise his initial decisions.

We also disentangled the effect of two possible channels of social influence.

Results suggest that some third-parties are only responsive to the discomfort for disagreeing with the majority, that is at the base of normative social influence and their punishment choices are not influenced by the "need to be right" on which is based informational social influence. Distinguishing between these two channels of social influence is of primary importance for social analysts, since previous studies document that informational social influence causes a permanent change in behavior (see for example Newcomb et al., 1967). On the other hand, normative social influence is more ephemeral and leads to modifications of behavior that are subject to specific circumstances¹³ (Deutsch and Gerard, 1955; Cohen and Golden, 1972; Burnkrant and Cousineau, 1975).

These findings have two major implications. On one hand, they remark the importance in our societies of citizens' perception about peers' behavior. This is especially important in situations where beliefs of the general population systematically overestimate the frequency of socially undesirable behaviors, like frequently happens for perceived crime, benefit frauds or percentage of non-voters¹⁴. In these situations, policymakers might often achieve welfare-improving results by means of ad-hoc communication strategies that could overperform alternative and often more costly policies (see for example Casal and Mittone, 2014, where the authors discuss an application of social stigma to tax evasion).

On the other hand, even when population beliefs are not biased, the possibility of resorting to social influence as a subsidiary tool for achieving compliance has been advanced by scholars in an array of situations of economic

¹³Nevertheless, scholars proposed models of endogenous preferences, arguing that even individuals initially adopting compliant behaviors by means of normative social influence may endogenously modify their preferences (Akerlof, 1989; Klick and Parisi, 2008).

¹⁴For example, the Royal Statistical Society reports that 58% of the UK population estimates that crime is rising, while data show how crime rate in the country is 19% lower than the previous year and 53% lower than 1995. For discussion of other examples and additional details see <http://www.kcl.ac.uk/newsevents/news/newsrecords/2013/07-July/Perceptions-are-not-reality-the-top-10-we-get-wrong.aspx>.

importance (Ela, 2008; Posner, 2000; Cooter, 1998; Zasu, 2007). As a society, we invest considerable amount of resources with the objective of shaping individual beliefs and direct them toward social desirable outcomes. Policymakers might want to encourage, by means of a social influence approach, third party interventions in situations where a centralized sanctioning authority might lack the ability or the resources to result effective. This is the case for example of the recent campaign aiming at prevention of social offenses "*Bringing in the Bystander*" promoted in the UK by the National Sexual Violence Resource Center. The campaign aims at reducing social offenses employing a marketing campaign that explicitly encourages third parties intervention¹⁵.

We agree that third-party punishment plays a key role in cementing human societies together (Mathew and Boyd, 2011). In this article we argue for the first time about the possibility for policymakers to take advantage of social influence effects in promoting third-party punishment, reporting evidence from a laboratory experiment that social influence significantly affects bystanders' interventions. Given the importance and wide possibilities of application in the societal framework, we hope that future researches further investigate the connection between social influence and third party punishment, in particular verifying the robustness of our findings in a field setting and the persistence of the effects in a longer term horizon.

¹⁵"Using a bystander intervention approach combined with a research component, this program assumes that everyone has a role to play in prevention [...] The Know Your Power campaign is the social marketing component of *Bringing in the Bystander*".

Appendix A.

First Period Punishment

Treatment	StratP0	StratP5	StratP10	StratP15	StratP20	StratP25	StratP30
Control							
(mean)	1.18	2.07	2.40	2.96	4.13	4.82	5.79
(median)	0	0	2	3	4	5	5
(SD)	3.31	4.09	2.74	2.97	3.94	4.68	5.72
Normative							
	.36	1.73	2.58	3.58	4.29	4.87	5.73
	0	1	2	2	4	5	5
	1.05	2.86	3.13	4.36	4.44	5.03	6.01
Informational							
	1.38	1.90	3.36	3.74	5.14	6.17	7.33
	0	1	3	4	5	6.5	8.5
	3.60	2.16	3.83	3.63	5.00	5.48	6.15
Total	.96	1.90	2.77	3.42	4.51	5.27	6.26
	0	1	2	3	5	5	6
	2.88	3.14	3.25	3.69	4.46	5.06	5.96

Table A.7: Average First Period Punishment by Levels of dictator Taking. StratP0 = dictator take 0 tokens from receiver

Second Period Punishment

Treatment	Punish0	Punish5	Punish10	Punish15	Punish20	Punish25	Punish30
Control							
(mean)	1.02	2.02	2.73	3.36	4.42	5.02	6.22
(median)	0	1	2	2	4	4	6
(SD)	3.18	3.45	3.49	3.58	4.38	5.24	6.26
Normative							
	.62	1.69	2.47	3.18	3.87	4.2	5.18
	0	1	2	3	3	4	4
	1.99	2.86	3.27	3.63	4.31	4.83	6.02
Informational							
	1.5	2.26	2.95	3.60	4.60	5.48	6.31
	0	1	3	4	5	6	8
	4.07	3.92	3.66	3.87	4.24	4.83	5.54
Total							
	1.04	1.98	2.71	3.37	4.29	4.89	5.89
	0	1	2	3	4.5	5	6
	3.16	3.41	3.45	3.67	4.29	4.96	5.93

Table A.8: Average Second Period Punishment by Levels of dictator Taking. Punish0 = dictator take 0 tokens from receiver

Appendix B. Instructions

Welcome! This is a study on individual decision-making. Participants' answers are completely anonymous. It will not be possible to data analysts to link individual answers to the participants that provided them. You earned euro 5 for showing up on time today. Additionally, you can collect other earnings. The amount of these earnings depends on your choices and from the choices other participants will make during the study. During the study you will earn "tokens". Each 10 tokens earned, euro 1 will be paid out to you. In the unlikely case you will collect negative earnings, those losses will be subtracted from your participation fee. If you have questions at any time, please raise your hand and wait for a researcher that will answer your questions privately. Please switch off and remove from the table any electronic device, do not talk or communicate with other participants during the study. The study is composed of more parts. Earnings obtained in each part of the study are independent from those obtained in the other parts. Your final earnings are composed by:

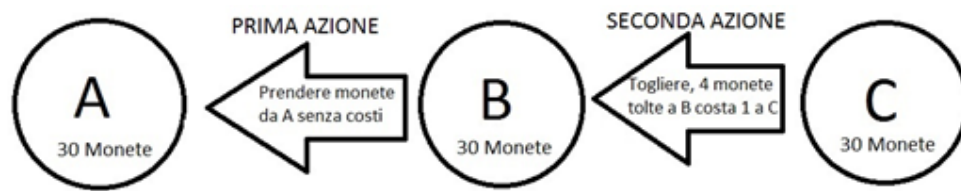
- Euro 5 of the participation fee
- Earnings collected in the first part of the study
- Earnings collected in one part after the first one. At the end of the study the computer will randomly select the part after the first one your earnings will be paid out to you

Final earnings will be paid privately and cash at the end of the study

First Part Instructions: description of the situation (Instructions on this part are the same in the 3 treatments)

Consider a situation with 3 people. Each person is randomly assigned to a role: one "Person A", one "Person B" and one "Person C". A, B and C could make decisions and earn tokens.

- Person A receives 30 tokens and does not make decisions
- Person B receives 30 tokens. Moreover, B could take some or all A's tokens and add them to his own earnings without incurring costs. Precisely, B could take 0, 5, 10, 15, 20, 25 or 30 tokens from A.
- Person C receives 30 tokens, observes B's action and could eliminate some of B's tokens, incurring a cost. For each 4 tokens eliminated from B's earnings, A has to pay 1 token. Person C could use up to 20 tokens to reduce B's earnings. C's decision does not affect A's earnings



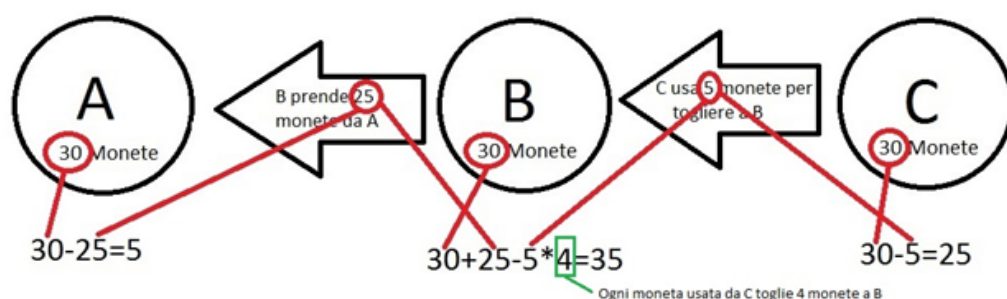
Therefore, A, B and C earnings are:

- Person A: $(30 \text{ initial tokens}) - (\text{tokens taken by B})$
- Person B: $(30 \text{ initial tokens}) + (\text{tokens taken from A}) - (4 * \text{tokens used by C})$
- Person C: $(30 \text{ initial tokens}) - (\text{tokens used for reducing B's earnings})$

Example 1) (please look at your computer screen): B takes 25 tokens from A. After observing B's choice, C decides to use 5 tokens to reduce B's earnings. Therefore participants' final earnings are:

- Person A = 5 tokens (tokens left by B)

- Person B = 35 tokens (30 initial tokens + 25 tokens taken from A – $5 \cdot 4 = 20$ tokens coming from the 5 tokens used by C to reduce B's earnings)
- Person C = 25 tokens (30 initial tokens – 5 tokens used to reduce B earnings)

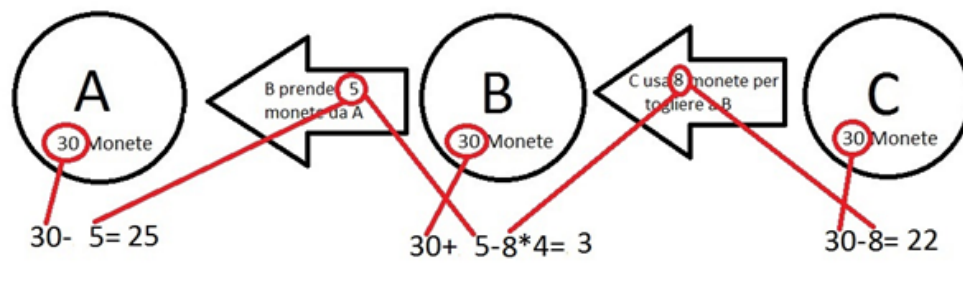


Example 2) (please look the computer screen): B takes 5 tokens from A. After observing B's choice, C uses 8 tokens to reduce B's earnings. Therefore participants' final earnings are:

- Person A = 25 tokens (left by B)
- Person B = 3 tokens (30 initial tokens + 5 tokens taken from A – $8 \cdot 4 = 32$ tokens coming from the 8 tokens used by C to reduce B's earnings)
- Person C = 22 tokens (30 initial tokens – 8 tokens used to reduce B's earnings)

Your actions and earnings

Person C observes how many tokens B takes from A. You and the other participants in the laboratory have to indicate the number of tokens, an



integer between 0 and 20, that C in your opinion will use. When everyone have answered, we calculate the average of the individual amounts indicated by you and the other participants. If the number you indicated is equal, bigger or smaller by one unit respect the average, you receive 40 tokens that will be added to your final earnings (if you indicate 0, you will receive the forty tokens if the average is 0, 1 or 2; if you indicate 20, you will receive the 40 tokens if the average is 20, 19 or 18). Instead, you do not earn tokens in this part of the study if the number you indicate is bigger or smaller by more than one unit with respect to the average.

Example 1) (please look at the computer screen): Consider the action of B “take 20 tokens from A and collect 50 tokens, leaving 10 tokens to A”. You indicate that C uses 11 tokens. You receive 40 tokens if on average all the participants to the study indicated “11”, “10” or “12” tokens. If the average is different from these numbers, you will not earn tokens for this part of the study

Example 2) (please look at the computer screen): Consider the action of B “take 0 tokens from A and collect 30 tokens, leaving 30 tokens to A”. You indicate that C uses 3 tokens. You receive 40 tokens if on average all the participants to the study indicated “3”, “2” or “4” tokens. If the average is different from these numbers, you will not earn tokens for this part of the study.

Tempo rimanente per prendere una decisione 113

La tabella sotto elenca le possibili scelte della Persona B. Quante monete userà la Persona C? Guadagni 40 monete se la tua risposta è uguale oppure maggiore/minore di una moneta a quella media fornita dai partecipanti.

Scegli Persona B	Quante monete userà la Persona C?
Prendere monete 30 dalla Persona A (La Persona B viene pagata monete (30 - 4*monete usate da C), la Persona A monete 0, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 25 dalla Persona A (La Persona B viene pagata monete (25 - 4*monete usate da C), la Persona A monete 5, la persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 20 dalla Persona A (La Persona B viene pagata monete (20 - 4*monete usate da C), la Persona A monete 10, la Persona C monete (30-monete usate))	<input type="text" value="11"/>
Prendere monete 15 dalla Persona A (La Persona B viene pagata monete (15 - 4*monete usate da C), la Persona A monete 15, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 10 dalla Persona A (La Persona B viene pagata monete (10 - 4*monete usate da C), la Persona A monete 20, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 5 dalla Persona A (La Persona B viene pagata monete (5 - 4*monete usate da C), la Persona A monete 25, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 0 dalla Persona A (La Persona B viene pagata monete (0 - 4*monete usate da C), la Persona A monete 30, la Persona C monete (30-monete usate))	<input type="text"/>

[Continua](#)

**Guadagni le 40 monete se in media i partecipanti allo studio avranno indicato "11" oppure "10" oppure "12".
Guadagni 0 se la media e' diversa da questi valori.**

Tempo rimanente per prendere una decisione 293

La tabella sotto elenca le possibili scelte della Persona B. Quante monete userà la Persona C? Guadagni 40 monete se la tua risposta è uguale oppure maggiore/minore di una moneta a quella media fornita dai partecipanti.

Scegli Persona B	Quante monete userà la Persona C?
Prendere monete 30 dalla Persona A (La Persona B viene pagata monete (30 - 4*monete usate da C), la Persona A monete 0, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 25 dalla Persona A (La Persona B viene pagata monete (25 - 4*monete usate da C), la Persona A monete 5, la persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 20 dalla Persona A (La Persona B viene pagata monete (20 - 4*monete usate da C), la Persona A monete 10, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 15 dalla Persona A (La Persona B viene pagata monete (15 - 4*monete usate da C), la Persona A monete 15, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 10 dalla Persona A (La Persona B viene pagata monete (10 - 4*monete usate da C), la Persona A monete 20, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 5 dalla Persona A (La Persona B viene pagata monete (5 - 4*monete usate da C), la Persona A monete 25, la Persona C monete (30-monete usate))	<input type="text"/>
Prendere monete 0 dalla Persona A (La Persona B viene pagata monete (0 - 4*monete usate da C), la Persona A monete 30, la Persona C monete (30-monete usate))	<input type="text" value="3"/>

[Continua](#)

Guadagni le 40 monete se in media i partecipanti allo studio avranno indicato "3", "2" oppure "4". Guadagni 0 se la media e' diversa da questi valori

You are required to indicate how many tokens Person C uses for each possible action of B (B takes 30 tokens from A; B takes 25 tokens. . . ; B takes 0 tokens from A). At the end of the study, the computer will randomly select one of the 7 actions of Person B. Relatively to this action, we will verify if you earned the 40 tokens. Your decisions and those of the other participants relative to other possible actions of Person B will be discarded and will not affect your final earnings.

Tempo rimanente per prendere una decisione: 00:01

La tabella sotto elenca le possibili scelte della Persona B. Quante monete uscirà la Persona A? Quante monete uscirà la Persona C? (Seleziona 00 monete se la tua risposta è uguale oppure maggiore/minore di una moneta a quella scelta dalla persona).

Quante monete uscirà la Persona B?	Quante monete uscirà la Persona A?	Quante monete uscirà la Persona C?
La Persona B viene pagata moneta (30) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 30 dalla Persona A	
La Persona B viene pagata moneta (25) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 25 dalla Persona A	
La Persona B viene pagata moneta (20) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 20 dalla Persona A	
La Persona B viene pagata moneta (15) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 15 dalla Persona A	
La Persona B viene pagata moneta (10) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 10 dalla Persona A	
La Persona B viene pagata moneta (5) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 5 dalla Persona A	
La Persona B viene pagata moneta (0) - 40 monete uscirà da C; la Persona A moneta 1; la Persona C moneta (30 monete uscirà)	Prendere moneta 0 dalla Persona A	

Durante questa prima parte indichi quante monete uscirà C. Devi indicare 7 risposte, una per ogni possibile scelta di B.

Alla fine dello studio il computer seleziona in maniera casuale una delle 7 azioni di B.

Solo la tua risposta data in corrispondenza di questa azione di B modifica i tuoi guadagni. Tutte le altre risposte vengono scartate.

Before starting this first part of the study, we ask you to answer some control questions. Answers to these control questions will not affect your final earnings.

(Participants answer control questions on their computers. The Ztree file containing the control questions is available upon request to the authors).

Instruction second part: description of the situation (instructions on this part are the same in all treatments)

Consider the same situation described in the first part, where 3 people are present, A, B and C, that can make decisions and earn tokens. Exactly as in the first part:

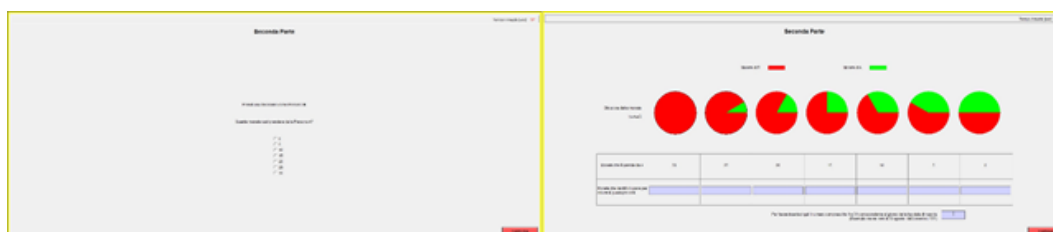
- A receives 30 tokens and do not makes decisions

- B receives 30 tokens and could take some or all the tokens of A
- C receives 30 tokens, observes the action of B and could reduce earnings of B paying a cost (every 4 tokens of reduction of B's earnings C has to pay 1 token)

Your actions and earnings

In this second part you and the other participants have to make decisions first as “Person B” then as a “Person C”. Therefore, you have to indicate:

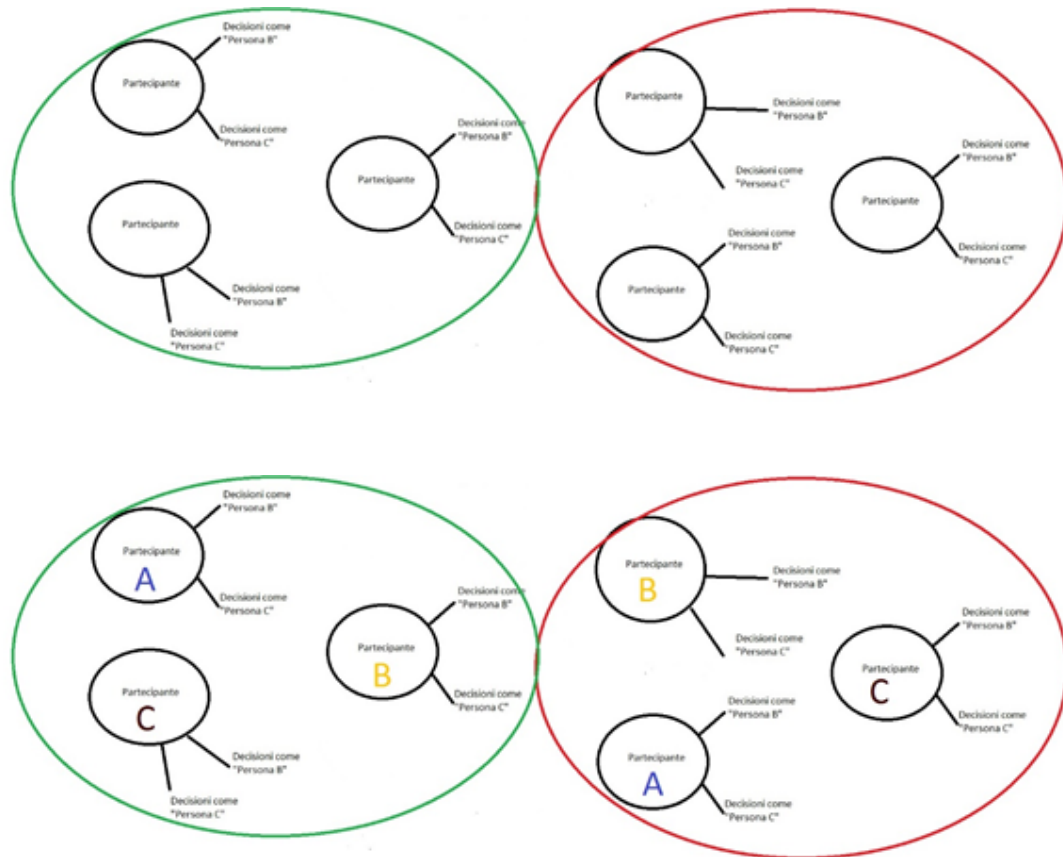
- First, as “Person B”, how many tokens you take from A
- After, as a “Person C”, for any possible action of B how many tokens you use for reducing B's earnings



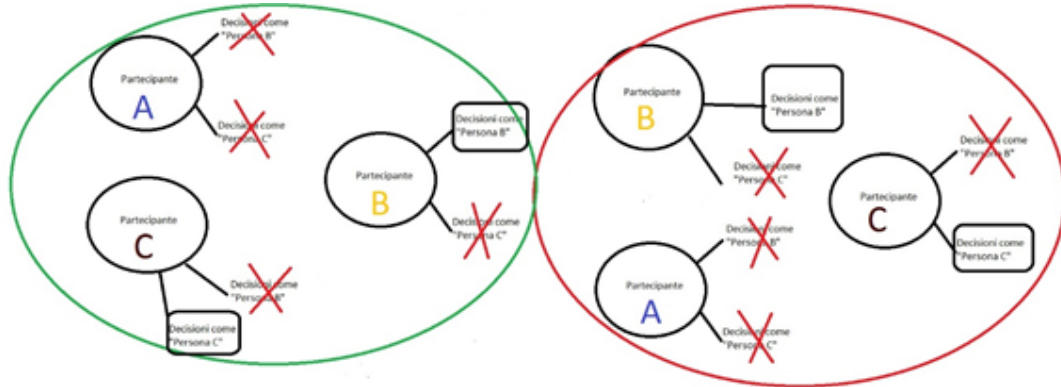
Why do you have to make decisions both as “Person C” and as a “Person B”? In calculating final earnings, each participant is associated to a unique role: either Person A or Person B or Person C. However, you and the other participants will not know which role has been assigned to you until the end of the study today. Indeed, you and the other participants will be randomly divided in groups of 3

Within the group, each one of the 3 participants is assigned either to role A, B or C

Assignment to groups and assignment of roles is completely random and each participant has 1 possibility over 3 of being assigned a specific role.



Therefore, if you are assigned the role “Person A”, your final earnings are determined by the tokens left you by the Person B that is in your same group. Other decisions you make as a Person B or C will be discarded and have no influence on your final earnings nor on the earnings of the other participants. Similarly, participants assigned to the role “Person B” determine their final earnings and those of the other group components only by the decisions make as Person B. Decisions made as Person C have no effects on final earnings. Finally, also Participants assigned to role “Person C” only influence final earnings only by decisions make as C.



During this second part of the study we will also ask to indicate the day of the month in which you were born (E.g. if you were born January 25th 1983 you should report “25”).

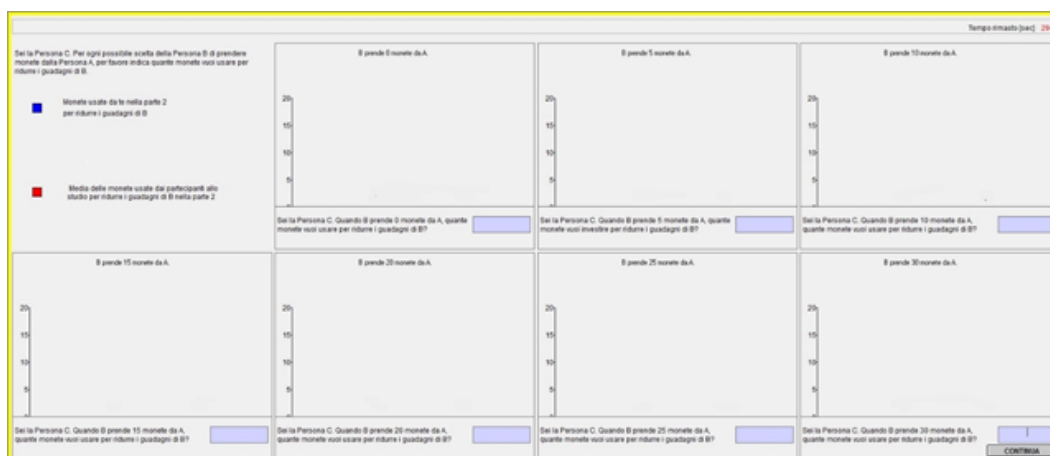
Earnings of A, B and C in this second part are determined exactly as in the first part:

- Person A: $(30 \text{ initial tokens}) - (\text{tokens taken by B})$
- Person B: $(30 \text{ initial tokens}) + (\text{tokens taken from A}) - (4 * \text{tokens used by C})$
- Person C: $(30 \text{ initial tokens}) - (\text{tokens used for reducing B's earnings})$

Before starting this first part of the study, we ask you to answer some control questions. Answers to these control questions will not affect your final earnings.

Instruction third part (Normative Treatment; instructions for Control and Informational are available upon request)



Now it starts the third and last part of this study. After the end of this part, we will ask you to fill in a brief questionnaire and then we will proceed with payments. Consider exactly the same situation of the second part of the study, same roles of A, B and C, same possible decisions that B and C have to make and same initial endowments and possible earnings. As in the second part, you have to make decisions first as a Person B then as a Person C. Additionally, in this third part before making your decisions you will receive information regarding the other participants. You will receive information on decisions make as Person C by the participants at today study. You will know how many tokens on average participants used in the second part of the study to reduce B's earnings. You will receive this information for any of the 7 possible B's choices.



Furthermore, before the end of the study, individual decisions as “Person C” that you are going to make in this third part will be revealed to 5 participants randomly selected. Similarly, you will received information regarding the individual choices made as Person C by 5 other participants. Each participant will be randomly assigned to an ID number. The ID number assigned is independent from the number of PC you are sit on. After you saw the individual choices of the other 5 participants, you and the other

Monete che B prende da A	Media monete riduzione guadagni B parte 2	Monete usate Partecipante 1 parte 3	Monete usate Partecipante 2 parte 3	Monete usate Partecipante 3 parte 3	Monete usate Partecipante 4 parte 3	Monete usate Partecipante 5 parte 3
0	0	0	0	0	0	0
5	0	0	0	0	0	0
10	0	0	0	0	0	0
15	0	0	0	0	0	0
20	0	0	0	0	0	0
25	0	0	0	0	0	0
30	0	0	0	0	0	0

participants will be able to vote for sending a smiling or sad emoticon

Partecipante 15	<input type="checkbox"/> Sorridente	<input type="checkbox"/> Triste
Partecipante 16	<input type="checkbox"/> Sorridente	<input type="checkbox"/> Triste
Partecipante 17	<input type="checkbox"/> Sorridente	<input type="checkbox"/> Triste

You receive a smiling emoticon if the majority of the five participants that saw your choices vote for “smiling”. Otherwise you will receive a sad emoticon. The emoticon will remain on your screen for one minute, then disappears automatically. After this minute has passed, you will know your final earnings.

If you have questions, please raise your hand and we will answer to you privately. Otherwise push “Continue” button and start with the third part.

Appendix C. Description of the Variables Used in the Regressions

Table C.9: Variables

Variable	Description
degree	1 if subject completed 8th grade ("scuola media"), 2 if subject completed high school, 3 if subject has a bachelor degree or equivalent, 4 if subject has a master degree or equivalent, 5 if subject has a PhD or equivalent
worker	binomial variable, 1 if worker
male	binomial variable, 1 if male
age	subject's age
social	binomial variable, 1 if subject is a student in social sciences and medicine
arts	binomial variable, 1 if subject is a student in arts or humanities
field_other	binomial variable, 1 if subject not in social or arts
DictatorTake	total amount of tokes a subject when choosing as a dictator takes to the receiver in the 2 periods
risk	$\in [1, 10]$, 1 if to question "In general, do you consider yourself ready to take risks?" the answer is "Not at all", 10 if the answer is "Totally ready to take risks"
logic	$\in [0, 2]$, 1 point for each correct answer. See figures C.3 and C.4 below for the 2 questions.

Variable	Description
impulsivity	$\in [0, 3]$, 1 point for each correct answer. See figures C.5, C.6 and C.7 below for the 3 questions.
NORMATIVE	binomial variable, 1 for subjects in normative treatment
TREATED	binomial variable, 1 for subject either in normative or in informational treatments
Strat_Punish	punishment first period
Beliefs_Punish	beliefs about peers' average punishment first period
Abs_p0Belifs	absolute value (Strat_Punish - Beliefs_Punish)
Abs Signalp0	absolute vale (Strat_Punish - Peers' average punishment period 1)
Abs BelAvgPun	absolute value (Beliefs_Punish - Peers' average punishment period 1)
NorStratPunish	$NORMATIVE * Strat_Punish$
NorAbs p0Belifs	$NORMATIVE * Abs_p0Belifs$
NormAbs_BeliefAvgPunish	$NORMATIVE * Abs_BeliefAvgPunish$
Diff_Deviation	$abs (Strat_Punish - Peers' average punishment period 1) - abs (Punish - Peers' average punishment period 1)$
ConvergeDummy	dummy, = 1 if Diff_Deviation > 0
Instructions	total time employed by subjects to correctly answer control questions

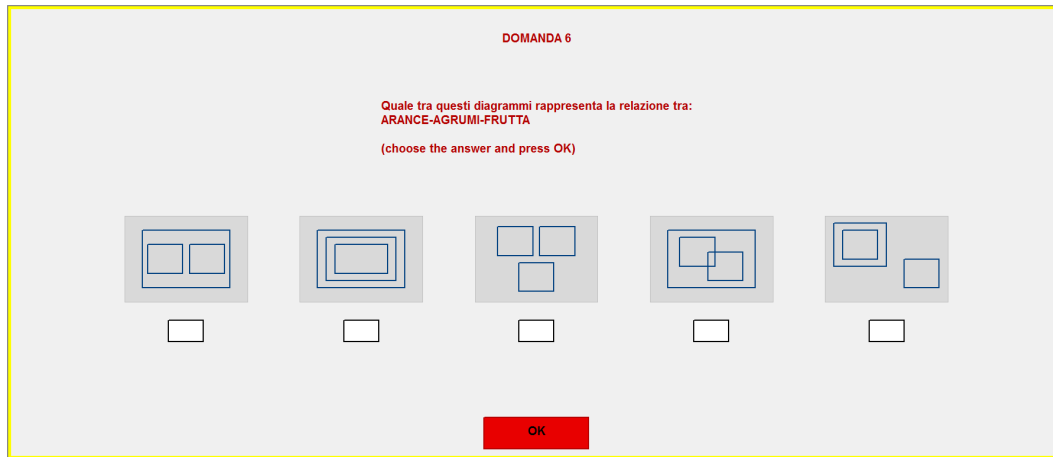


Figure C.3:



Figure C.4:

DOMANDA 8

Una mazza e una pallina costano euro 1.10 in total. La mazza costa euro 1.00 piu' della pallina.
Quanto costa la pallina?

Euro

Figure C.5:

DOMANDA 9

Se 5 macchine impiegano 5 minuti per fare 5 oggetti,
quanto tempo impiegheranno 100 macchine per fare 100 oggetti?

Minuti

Figure C.6:

In un lago c'è una chiazza di orchidee. Ogni giorno, questa chiazza raddoppia in dimensioni.
Se la chiazza di orchidee impiega 48 giorni per coprire completamente il lago, quanto tempo impiegherà per coprire la metà del lago?

Giorni

Figure C.7:

References

- Akerlof, G. A., 1980. A theory of social custom, of which unemployment may be one consequence. *The quarterly journal of economics* 94 (4), 749–775.
- Akerlof, G. A., 1989. The economic of illusion. *Economics & Politics* 1 (1), 1–15.
- Akerlof, G. A., Yellen, J. L., Katz, M. L., 1996. An analysis of out-of-wedlock childbearing in the united states. *The Quarterly Journal of Economics* 111 (2), 277–317.
- Asch, S., 1951. Effects of Group Pressure upon the Modification and Distortion of Judgments.
- Asch, S., 1956. Studies of Independence and Conformity: A minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied* 70 (9), 1–70.
- Axelrod, R., 1986. An Evolutionary Approach to Norms. *American political science review* 80 (04), 1095–1111.
- Becker, G. S., 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76 (2), 169–217.
- Becker, G. S., 1991. A note on restaurant pricing and other examples of social influences on price. *Journal of Political Economy* 99 (5), 1109.
- Bendor, J., Mookherjee, D., 1990. Norms, Third-party Sanctions, and Cooperation. *JL Econ & Org.* 6, 33.
- Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial Altruism in Humans. *Nature* 442 (7105), 912–915.
- Bernheim, B., 1994. A Theory of Conformity. *Journal of political Economy*, 841–877.

- Bond, R., Smith, P., 1996. Culture and Conformity: A Meta-analysis of Studies using Asch's (1952b, 1956) Line Judgment Task. *Psychological Bulletin*; *Psychological Bulletin* 119 (1), 111.
- Bose, P., 1995. Regulatory Errors, Optimal Fines and the Level of Compliance. *Journal of Public Economics* 56 (3), 475–484.
- Buckholtz, J., Asplund, C., Dux, P., Zald, D., Gore, J., Jones, O., Marois, R., 2008. The Neural Correlates of Third-party Punishment. *Neuron* 60 (5), 930–940.
- Burnkrant, R. E., Cousineau, A., 1975. Informational and normative social influence in buyer behavior. *Journal of Consumer research*, 206–215.
- Carpenter, J., 2007. The Demand for Punishment. *Journal of Economic Behavior & Organization* 62 (4), 522–542.
- Casal, S., Mittone, L., 2014. Social esteem versus social stigma: the role of anonymity in an income reporting game. CEEL Working Papers 1401, Cognitive and Experimental Economics Laboratory, Department of Economics, University of Trento, Italia.
- Casari, M., Luini, L., 2009. Cooperation under Alternative Punishment Institutions: An experiment. *Journal of Economic Behavior & Organization* 71 (2), 273–282.
- Cason, T. N., Mui, V.-L., 1998. Social influence in the sequential dictator game. *Journal of Mathematical Psychology* 42 (2), 248–265.
- Christakis, N. A., Fowler, J. H., 2007. The spread of obesity in a large social network over 32 years. *New England journal of medicine* 357 (4), 370–379.
- Cialdini, R. B., 1993. *Influence: the Psychology of Persuasion*. Collins, NY.
- Coffman, L., 2011. Intermediation Reduces Punishment (and Reward). *American Economic Journal: Microeconomics* 3 (4), 77–106.

- Cohen, J. B., Golden, E., 1972. Informational social influence and product evaluation. *Journal of Applied Psychology* 56 (1), 54.
- Coleman, S., 1996. The minnesota income tax compliance experiment: State tax results.
- Cooper, D., Rege, M., 2008. Social Interaction Effects and Choice under Uncertainty: an Experimental Study. Tech. rep., University of Stavanger.
- Cooter, R., 1998. Expressive law and economics. *The Journal of Legal Studies* 27 (2), 585–608.
- Cox, J. C., Friedman, D., Gjerstad, S., 2007. A tractable model of reciprocity and fairness. *Games and Economic Behavior* 59 (1), 17–45.
- Deutsch, M., Gerard, H. B., 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology* 51 (3), 629.
- Devenow, A., Welch, I., 1996. Rational Herding in Financial Economics. *European Economic Review* 40 (3), 603–615.
- Ela, J. S., 2008. Law and norms in collective action: Maximizing social influence to minimize carbon emissions. *UCLA Journal of Environmental Law & Policy* 27 (1).
- Ellickson, R., 1999. The Evolution of Social Norms: a Perspective from the Legal Academy. Yale Law School, Program for Studies in Law, Economics and Public Policy, Working Paper (230).
- Engel, C., 2011. Dictator games: a meta study. *Experimental Economics* 14 (4), 583–610.
- Falk, A., Fischbacher, U., 2002. “Crime” in the Lab-Detecting Social Interaction. *European Economic Review* 46 (4), 859–869.

- Falk, A., Fischbacher, U., Gächter, S., 2010. Living in Two Neighborhoods—Social Interaction Effects in the Laboratory. *Economic Inquiry*.
- Falk, A., Ichino, A., 2006. Clean Evidence on Peer Effects. *Journal of Labor Economics* 24 (1), 39–57.
- Fehr, E., Fischbacher, U., 2004a. Social norms and human cooperation. *Trends in cognitive sciences* 8 (4), 185–190.
- Fehr, E., Fischbacher, U., 2004b. Third-party Punishment and Social Norms. *Evolution and human behavior* 25 (2), 63–87.
- Fehr, E., Fischbacher, U., et al., 2003. The Nature of Human Altruism. *Nature* 425 (6960), 785–791.
- Fehr, E., Gächter, S., 2002. Altruistic Punishment in Humans. *Nature* 415, 137–140.
- Fischbacher, U., 2007. z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics* 10 (2), 171–178.
- Fortin, B., Lacroix, G., Villeval, M., 2007. Tax Evasion and Social Interactions. *Journal of Public Economics* 91 (11), 2089–2112.
- Fowler, J. H., 2005. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America* 102 (19), 7047–7049.
- Gächter, S., Fehr, E., 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90 (4), 980–994.
- Galbiati, R., Zanella, G., 2012. The tax evasion social multiplier: Evidence from Italy. *Journal of Public Economics* 96 (5), 485–494.
- Glaeser, E. L., Sacerdote, B., Scheinkman, J. A., 1996. Crime and social interactions. *The Quarterly Journal of Economics* 111 (2), 507–548.

- Greiner, B., 2004. An Online Recruitment System for Economic Experiments.
- Güth, W., Schmittberger, R., Schwarze, B., 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization* 3 (4), 367–388.
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J., Gurven, M., Gwako, E., Henrich, N., et al., 2006. Costly Punishment across Human Societies. *Science* 312 (5781), 1767–1770.
- Hetcher, S., 2004. Norms in a Wired World. Cambridge University Press.
- Hirshleifer, D., Hong Teoh, S., 2003. Herd Behaviour and Cascading in Capital Markets: a Review and Synthesis. *European Financial Management* 9 (1), 25–66.
- Hoff, K., Kshetramade, M., Fehr, E., 2011. Caste and Punishment: the Legacy of Caste Culture in Norm Enforcement. *The Economic Journal* 121 (556), F449–F475.
- Kahan, D. M., 1997. Social influence, social meaning, and deterrence. *Virginia Law Review*, 349–395.
- Klick, J., Parisi, F., 2008. Social networks, self-denial, and median preferences: Conformity as an evolutionary strategy. *The Journal of Socio-Economics* 37 (4), 1319–1327.
- Krupka, E., Weber, R., 2009. The Focusing and Informational Effects of Norms on Pro-social Behavior. *Journal of Economic Psychology* 30 (3), 307–320.
- Krupka, E. L., Weber, R. A., 2013. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association* 11 (3), 495–524.

- Kurzban, R., DeScioli, P., O'Brien, E., 2007. Audience Effects on Moralistic Punishment. *Evolution and Human behavior* 28 (2), 75–84.
- Lewisch, P., Ottone, S., Ponzano, F., 2011. Free-Riding on Altruistic Punishment? An Experimental Comparison of Third-Party. *Review of Law and Economics* 7 (1).
- Lieberman, D., Linke, L., 2007. The Effect of Social Category on Third-party Punishment. *Evolutionary Psychology* 5 (2), 289–305.
- List, J. A., 2007. On the interpretation of giving in dictator games. *Journal of Political Economy* 115 (3), 482–493.
- Lotz, S., Baumert, A., Schlösser, T., Gresser, F., Fetchenhauer, D., 2011. Individual Differences in Third-Party Interventions: How Justice Sensitivity Shapes Altruistic Punishment. *Negotiation and Conflict Management Research* 4 (4), 297–313.
- Manski, C., 2000. Economic Analysis of Social Interactions. *Journal of Economic Perspectives* 14 (3), 115–136.
- Marlowe, F., Berbesque, J., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., et al., 2008. More "Altruistic" Punishment in Larger Societies. *Proceedings of the Royal Society B: Biological Sciences* 275 (1634), 587–592.
- Mas, A., Moretti, E., 2009. Peers at work. *American Economic Review* 99 (1), 112–145.
- Mathew, S., Boyd, R., 2011. Punishment Sustains Large-scale Cooperation in Prestate Warfare. *Proceedings of the National Academy of Sciences* 108 (28), 11375–11380.
- Nelissen, R., 2008. The Price You Pay: Cost-dependent Reputation Effects of Altruistic Punishment. *Evolution and Human Behavior* 29 (4), 242–248.

- Nelissen, R., Zeelenberg, M., 2009. Moral Emotions as Determinants of Third-party Punishment: Anger, Guilt, and the Functions of Altruistic Sanctions. *Judgment and Decision Making* 4 (7), 543–553.
- Newcomb, T. M., Koenig, K. E., Flacks, R., Warwick, D. P., 1967. Persistence and change: Bennington College and its students after twenty-five years. Wiley New York.
- Nikiforakis, N., 2008. Punishment and Counter-punishment in Public Good Games: Can We Really Govern Ourselves? *Journal of Public Economics* 92 (1), 91–112.
- Okimoto, T., Wenzel, M., 2011. Third-party Punishment and Symbolic Intra-group Status. *Journal of Experimental Social Psychology* 47 (4), 709–718.
- Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86 (02), 404–417.
- Perkins, H., Linkenbach, J. W., Lewis, M. A., Neighbors, C., 2010. Effectiveness of social norms media marketing in reducing drinking and driving: A statewide campaign. *Addictive behaviors* 35 (10), 866–874.
- Piazza, J., Bering, J., 2008. Concerns about Reputation via Gossip Promote Generous Allocations in an Economic Game. *Evolution and Human Behavior* 29 (3), 172–178.
- Posner, E. A., 2000. Law and social norms: The case of tax compliance. *Virginia Law Review*, 1781–1819.
- Raihani, N., Grutter, A., Bshary, R., 2010. Punishers Benefit from Third-party Punishment in Fish. *Science* 327 (5962), 171–171.
- Scharfstein, D., Stein, J., 1990. Herd Behavior and Investment. *The American Economic Review*, 465–479.

- Seymour, B., Singer, T., Dolan, R., 2007. The Neurobiology of Punishment. *Nature Reviews Neuroscience* 8 (4), 300–311.
- Shinada, M., Yamagishi, T., Ohmura, Y., 2004. False Friends are Worse than Bitter Enemies: “Altruistic” Punishment of In-group Members. *Evolution and Human Behavior* 25 (6), 379–393.
- Slemrod, J., Yitzhaki, S., 2002. Tax Avoidance, Evasion, and Administration. *Handbook of public economics* 3, 1423–1470.
- Sunstein, C. R., Schkade, D., Ellman, L. M., Sawick, A., 2006. Are judges political?: an empirical analysis of the federal judiciary. *Brookings Institution Press*.
- Swope, K., Cadigan, J., Schmitt, P., Shupp, R., 2008. Social position and distributive justice: Experimental evidence. *Southern Economic Journal*, 811–818.
- Thaler, R. H., Sunstein, C. R., 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Topa, G., 2001. Social Interactions, Local Spillovers and Unemployment. *The Review of Economic Studies* 68 (2), 261–295.
- Turner, J. C., 1991. *Social influence*. Thomson Brooks/Cole Publishing Co.
- Zasu, Y., 2007. Sanctions by Social Norms and the Law: Substitutes or Complements? *The Journal of Legal Studies* 36 (2), 379–396.