EIEF WORKING PAPER SERIES

Einaudi Institute for Economics and Finance

EIEF Working Paper 16/13
May 2013

Social Identity and Punishment

by

Jeffrey V. Butler
(EIEF)

Pierluigi Conzo
(University of Turin)

Martin A. Leroch
(University of Mainz)

Social Identity and Punishment

Jeffrey V. Butler
EIEF

Pierluigi Conzo
University of Turin

Martin A. Leroch
University of Mainz

This version: May 23, 2013

Abstract

Third party punishment is crucial for sustaining cooperative behavior. Still, little is known about its determinants. In this paper we use laboratory experiments to investigate a long-conjectured interaction between group identification and bystanders' punishment preferences using a novel measure of these preferences. We induce minimal groups and give a bystander the opportunity to punish the perpetrator of an unfair act against a defenseless victim. We elicit the bystander's valuation for punishment in four cases: when the perpetrator, the victim, both or neither are members of the bystander's group. We generate testable predictions about the rank order of punishment valuations from a simple framework incorporating group-contingent preferences for justice which are largely confirmed. Finally, we conduct control sessions where groups are not induced. Comparing punishment across treatment and control suggests that third-party punishers tend to treat others as in-group members unless otherwise divided.

JEL Classification : D74, Z1
Keywords: Identity, social norms, culture, cheating, in-group bias, punishment

1 Introduction

The willingness of bystanders to punish transgressions committed against others is an important phenomenon. Enforcement of social norms of cooperation, crucial to the existence of society, may depend on such third party punishment (Fearon and Laitin, 1996; Fehr and Fischbacher, 2004; Carpenter and Matthews, 2010). On the other hand, bystanders entering into disputes and punishing transgressors on the behalf of those directly affected may prolong and extend conflicts beyond an initially limited scope.[1]

One potential determinant that has garnered both theoretical and empirical attention is social/group identity. A handful of existing papers examine particular theoretical conjectures about the interaction between group identification processes and bystander punishment preferences. Long ago, Darwin suggested the logic of group selection hinged upon group-contingent punishment, noting that "...groups with a greater number of courageous, sympathetic and faithful members, who were always ready to warn each other of danger, to aid and defend each other ... would spread and be victorious over other tribes." (1873, quoted in De Dreu, et al., 2010). On the other hand, Choi and Bowles (2007) posit a theory of parochial altruism in which group-directed altruism and a preference for punishing outsiders are necessarily intertwined: neither would survive evolutionary pressures by itself, but combined they do. In the former, one would expect evolutionary processes to deliver a preference to punish transgressions commited against in-group members by out-group members. In the latter, punishing norm violations should be a predominantly within-group affair while punishing outsiders should be more indiscriminate.[2]

Empirically, the results have been mixed and the findings across studies are seemingly contradictory. For instance, using experimentally-induced minimal groups Goette et al. (2012) find that out-group transgressors incur more third party punishment irrespective of the victim's group membership. On the other hand, in an experiment featuring real-world tribal affiliations as the relevant groups, Bernhard, et *al.* (2006) find that it is the group affiliation of the *victim* that matters most for a third party's punishment decisions: harmful behavior affecting in-group fellows is punished more harshly, irrespective of the

[1] The dispute between the Hatfields and the McCoys is one infamous example in the American context. Off the equilibrium path, this amplifying effect may have a silver lining: the spectre of costly prolonged disputes can sustain cooperation in equlibrium (e.g., Fearon and Laitin's "spiral equilibrium.")

[2] An extreme example is the "amoral familism" documented by Banfield (1958) in southern Italy in which moral obligations do not extend outside of the nuclear family.

perpetrator's group affiliation.

Why do results conflict so starkly? One possible reason relates to how punishment is measured. All studies we are aware of use a fixed-price punishment technology: third parties choose how much punishment to levy at a fixed per-unit cost of punishment, where punishment typically takes the form of reducing the transgressors earnings. We would argue that the *amount* of punishment bystanders choose in this setup is a function of at least three components: i) a value judgment about how wrong the act being punished is (moral disgust); ii) the bystander's feeling of responsibility for undoing the injustice; and iii) a desire to deter bad behavior in the first place. The second and third components may be particularly important when the range of feasible punishment is substantial. As bad behavior is typically measured by an unequal money division, reducing the transgressor's payoff sufficiently can restore earnings equality. Prior research suggests the importance of all three components in a fixed-price punishment setting. For example, Lewish, Ottone and Ponzano (2010) document that individuals each levy less punishment when more than one person can punish, ostensibly because responsibility is made more diffuse.

While all three components are interesting in their own right, group identification processes likely affect each of the three in different ways and to different degrees. Group identification may even be synonymous with an enhanced feeling of responsibility for in-group members relative to out-group members. Ideally, to uncover whether and how identity affects punishment preferences one would like to isolate its effect on each component of punishment preferences.

In this paper, we contribute to the literature on the third party punishment and social identity in three ways. First of all, we implement a novel punishment technology allowing us to isolate the moral disgust component of punishment preferences. We fix the amount of punishment third parties can levy at a level that is both a small fraction of the damage inflicted on victims of an unfair act and a small fraction of the perpetrator's potential gain from acting unjustly. We then elicit the bystander's valuation for this fixed amount of punishment in an incentive compatible manner. This punishment mechanism minimizes the scope for responsibility, since the bystander can never undo, in any meaningful way, the injustice that is perpetrated. It also minimizes the scope for deterrence, since the punishment never gets close to wiping out the perpetrator's gains from injustice. It leaves intact, however, a vehicle for expressing moral disgust so that we interpret the bystander's valuation for the punishment opportunity as a relatively clean measure of his or her value judgments about how wrong particular situations are. Secondly, we implement treatment and control sessions featuring minimal groups and no groups, respectively. By comparing punishment preferences across treatment and control, we provide novel evidence on how the *introduction* of groups affects bystander punishment. Our third contribution stems from the treatment sessions, where we vary the group affiliations of the perpetrator, victim and bystander independently and measure the bystander's punishment preferences in each scenario. Here, we are the first to provide clean, incentive compatible, evidence on how the bystander's relationship with the directly involved parties affects his or her value judgment of how wrong an injustice is, i.e., the moral disgust component of punishment preferences.

For concreteness, we consider a dictator game with third party punishment (Fehr and Fischbacher, 2004) and construct a theoretical framework in the spirit of Chen and Li (2010) and Akerlof and Kranton (2005) which allows us to generate testable predictions in this game. Our model assumes that all individuals have a basic preference for justice: they would be willing to incur some personal cost to punish unfair actions. This basic justice preference is modified by group identification processes. We posit that identification causes bystanders to internalize the preferences of in-group members to a greater extent than those of other-group members.

We find that punishment preferences are broadly consistent with our theoretical predictions: most bystanders place the highest value on punishing an outsider for treating an insider unfairly, and least value on punishing an insider for treating an outsider unfairly. Both of these patterns are consistent with a Darwinian view of third party punishment. At the aggregate level, only the first pattern holds, as on average the scenario where both dictator and recipient share the bystander's group affiliation is associated with the lowest valuation for punishment. More generally, punishment levied on outsiders is typically valued more highly than punishment levied on members of the bystander's own group. Finally, comparing average overall punishment in treatment sessions to control session punishment, the data suggest that participants punish others as if they were all in-group members unless they are explicitly divided into distinct groups.

The remainder of the paper is organized as follows. We first discuss closely related literature. Then, the experimental design and procedures are detailed. In Sections 4 and 5 we discuss theories of punishment preferences and obtain predictions. Next, results are presented and discussed (Sections 6 and 7). In the final section, we summarize our findings and provide concluding remarks.

2 Closely related literature

Social identity has been studied for decades by social psychologists and sociologists and, quite recently, has also begun to receive attention from economists.[3] A result common to many existing studies is that maintenance and enforcement of social norms and, more generally, altruistic behavior is characterized by in-group bias: a predilection to favor members of one's own group over members of other groups.

In-group bias or favoritism can take various forms. Being matched with in-group fellows has been shown to increase cooperation (de Cramer and van Vugt, 1999; Guala et al., 2009),[4] increase the level of altruistic giving and reward for good behavior, and decrease punishment for bad behavior (Chen and Li, 2009). Further, Chen and Li (2009) also find that punishment patterns follow the logic of supply and demand. That is, an increase in costs of punishment lowers the propensity to punish, where the punishment of out-group members is more cost-sensitive than punishment of in-group fellows. On the other hand, when particular norms are central to a group's identity, in-group members may be *more* heavily punished for violating these norms than out-group members (McLeish and Oxoby, 2007).[5] Also, individuals may more readily harm members of other groups if this is to the benefit of their in-group (Bornstein 1992, 2003). In general, in-group favoritism has been found in various forms of groups, ranging from tribes (e.g. Bernhard et al., 2006) to other real-world social groups such as army platoons (Goette, et al., 2006, 2012) to minimal and close-to-minimal groups (Tajfel, et al., 1971; Chen and Li, 2009). What is debated, however, is whether in-group favoritism is based on preferences (Guala et al., 2009) or on strategic (individual) interests (Yamagishi and Kiyonari, 2000; Yamagishi and Mifune, 2008, 2009). In our experimental design, which we outline below, we exclude the possibility of strategic interests by implementing a simple one-shot structure.

Although in-group favoritism need not coincide with directly unkind behavior towards

[3] On the importance of identity in economics see Akerlof and Kranton (2000). For an excellent overview of the literature on social identity, see Chen and Li (2009).

[4] See Chen and Chen (2011) for a theoretical argument and experimental support for the increase in cooperation if salient social identity exists. Accordingly, social identity may serve as a coordination mechanism.

[5] For lab or field experiments on costly punishment see, among others, Fehr and Fischbacher (2004) and Henrich et al. (2006).

an out-group (Mumendey, 1992), some experimental results suggest that "vendettas" may evolve rather easily, even in anonymous laboratory settings. Abbink and Herrmann (2009), for instance, gave two opposing groups the possibility to reduce the endowment of the respective other group, at a cost to themselves, over ten subsequent periods. The introduction of a symbolic reward, which did not cover the own expenses of reducing the other group's endowment, tripled the rates of harmful behavior.[6] Because all group members are affected equally, these results seem to imply that subjects have an inclination to also punish others due to their group membership, and not primarily their actions. Experimental designs such as that implemented in Abbink and Herrmann (2009) cannot, however, rule out individual reciprocal attitudes.

How the group affiliations of perpetrators and victims factor into third parties' punishment decisions is not fully understood, and existing results seem to contradict each other. Goette et al. (2012), for instance, find in minimal group settings that out-group transgressors in a third-party punishment game are punished more heavily than transgressors from an in-group, independent of the group membership of the victim. This result was qualified by their findings made in social (i.e. non-minimal) groups, where defections against in-group members were punished more heavily than in minimal groups. Further, Bernhard et al. (2006) find in an experiment with natural groups (tribal affiliation) that it is exactly the group affiliation of the victim that matters for punishment decisions. Harmful behavior towards in-group fellows was punished harder than when out-group members were harmed, irrespective of the violator's group membership. In all of these studies, experimenters fixed the price of punishment and participants chose the amount of punishment to levy. Moreover, the amount of punishment that could be levied was large enough to substantially undo the unfair act. Consequently, the amount of punishment may measure both a value judgment about how unjust an act is and how much responsibility one feels for undoing the wrong: levying a lot of punishment could be the result of feeling a lot of responsibiity for a minimally-wrong action, or little responsibility for a very wrong action. Since we are mostly interested in how group membership affects the value judgment of unjust acts, our study differs from these studies in two main ways. First, by fixing the amount of punishment at a low, largely token, level and eliciting the value participants place on this marginal unit of punishment we seek a money-metric measure of observer's value judgments

[6] However, other experiments did not replicate this pattern (see e.g. Halevy et al., 2008).

of how wrong the same unjust act is perceived when group affiliations of involved parties are varied. Secondly, by implementing control sessions *without* group divisions, we aim to uncover whether, or how, the *introduction* of groups *per se* changes justice perceptions and punishment patterns.

Finally, various theoretical evolutionary arguments have been made about the form third party punishment should take. Some of these arguments predict an in-group bias in punishment, while others predict out-group bias. On the former, it appears to make sense that harm done to in-group members is readily punished, both in order to deter out-group aggression and also to foster in-group bonds.[7] On the latter, it could also make sense to readily punish in-group members who commit transgressions against out-groups in order to prevent costly inter-group conflicts from starting.[8] Since both arguments are *a priori* plausible, whether third party punishment patterns reflect in-group favoritism or, rather, whether the will to prevent intergroup conflicts leads to punishment directed more toward in-group members, is an empirical question. To address this question as cleanly as possible, we use laboratory experiments to focus on one specific aspect of punishment behavior: a money-metric measure of participants' value judgments concerning how wrong an unjust act is.

3 Experimental design and procedures

The experiment was conducted in the laboratory facilities at the Einaudi Institute for Economics and Finance (EIEF) in Rome, Italy, using pen and paper. Participants were recruited from a pre-existing list of individuals who expressed a general willingness to take part in experiments at EIEF. This list consists mainly of students from two nearby universities:

[7] Note that Choi and Bowles (2009) argue that (in-group) altruism and (inter-group) war may have co-evolved.

[8] Such reasoning finds support in the theoretical results of Fearon and Laitin (1996). They model inter-ethnic encounters as a repeated Prisoners' Dilemma, in which the possibility to build individual reputations across groups is limited by the low number of encounters. In this setting they find two punishing equilibria which may sustain cooperation within group boundaries and peace across group boundaries. In the first equilibrium, members of either group ignore transgressions committed by members of the other group affecting their own group, because they trust in the other group's punishment of perpetrators in their own ranks (which will indeed take place in equilibrium). In the second equilibrium, members of each group hold all members of the other group they can get hold of responsible for transgressions. In this case, cooperation is sustained by the fear of ending up in a vicious cycle of mutual "punishment" for earlier harm inflicted by the respective other group.

LUISS Guido Carli University and the University of Rome La Sapienza.[9] Six treatment sessions were conducted in which a total of 100 students took part.[10] A total of 96 students took part in six control sessions. In treatment sessions, participants were randomly divided into two groups before playing any games. In the control sessions participants were not divided into groups. An even number of participants took part in each session.[11]

An important design consideration we faced was whether to use real-world identity categories (university affiliation, favorite soccer team, etc.) or to use identities artificially induced in the laboratory. Because we wanted to be able to isolate the effects of categorization from obvious confounds associated with real-world divisions such as reputation or reciprocity stemming from previous interactions or expected future interactions, we decided to use artificial identities induced in the lab. In particular, the identities we induce fall within the minimal group paradigm of social psychology, where "...there is neither a conflict of interests nor previously existing hostility between the 'groups.' No social interaction takes place between the subjects, nor is there any rational link between economic self-interest and the strategy of in-group bias ...these groups are purely cognitive, and can be referred to as 'minimal.' " (Tajfel and Turner 1986, p. 14).

3.1 Minimal group inducement

At the beginning of each treatment session, participants were divided into two groups of equal size. This was accomplished by placing an equal number of red and blue chips into a bag: if there were n participants in a particular session, $\frac{n}{2}$ red chips and $\frac{n}{2}$ blue chips were placed into an opaque bag in full view of all participants.[12] Each participant drew one chip from the bag which determined his or her group. Participants were then given their experimental packet (described below) and a red or blue pen. The color of their assigned pen matched the color of their chip. They were instructed to use only this pen during the

[9] We do not exploit these university affiliations as a source of group identity. In particular, participants were not made aware of others' university affiliations. We recruit from these two populations because they are both situated in close proximity to EIEF. LUISS Guido Carli is a small private university in Rome, while La Sapienza is the largest public university in Rome, with a diverse student population totaling nearly 100,000.

[10] One person in one treatment session failed to respond to any of the questions about third-party punishment. Additionally, we were unable to match one treatment session participant to our demographic data. Consequently, our analyses incorporate only 98 observations from the treatment sessions.

[11] If an odd number of participants showed up, we randomly selected one person to be sent home and paid that person a show-up fee as is standard practice.

[12] These colors do not have political connotations in Italy as they might in, e.g., the U.S.

experiment. Finally, participants were seated, by color group, on opposite sides of the lab. The group-colored pen and group-contingent seating were meant to reinforce a sense of shared fate which previous research has shown to be crucial to engendering "groupness."[13] Within each color group, seats were assigned randomly. Each participant was separated from all other participants by an opaque divider, effectively creating a personal cubicle for each individual, to maintain anonymity of responses. Which side of the room was reserved for the red group and which was for the blue group was randomly determined before each session.

Once all participants were seated, general experimental instructions (do not talk, no cell phones, etc.) were read aloud and participants were given a few minutes to look through their experimental packet and ask questions if necessary. Any questions were answered privately by the experimenters. After all questions were answered, participants began the experiment.

Each participant's experimental packet contained instructions and response sheets for five simple games. Among these games was a binary dictator game with third party punishment, which we describe in detail below. Participants were to fill out the response sheet for each game. They were informed that only one of the games would be randomly chosen to count and that each game had the same probability of being chosen. The order in which the five games appeared in each packet was randomized to ameliorate order effects. We focus here mainly on the dictator game with third party punishment and leave for future work the analysis of the other four games in the packet.

3.2 The dictator game with third-party punishment

The binary dictator game with third party punishment is a complete and perfect information sequential game involving three parties: a dictator, a recipient and an observer.[14] Only the dictator and the observer make decisions, with the dictator moving first and the observer second. The dictator is endowed with 30 euros, the recipient with nothing and the observer with 15 euros. The dictator decides how to split his or her 30 euros endowment with the recipient. We restricted the set of available options to two: i) divide the sum evenly, so

[13] Another common technique that has been shown to enhance group-contingent behavior is to implement pre-play communication and cooperation on a group-specific task (*cf.*, Chen and Li, 2010). We chose to avoid this specific technique here in order to avoid confounding group-contingent preferences with generalized reciprocity

[14] The dictator role was called "the proposer," a more neutral term.

the dictator and the recipient both earn 15 euros; ii) divide the sum quite unevenly, so that the dictator retains 22 euros while the recipient earns only 8 euros. After observing the dictator's choice, the observer reports how much he or she is willing to spend to levy a (token) punishment on the dictator: a 1 euro reduction in the dictator's earnings. While the observer can reduce the dictator's earnings following either choice, previous research suggests that the first allocation is viewed as "fair" while the latter, unequal, allocation, is viewed as unfair (see, e.g., Bohnet and Zeckhauser, 2004 or Butler, Giuliano and Guiso, 2011). Because our aim is to study punishing transgressions, we focus on the observer's punishment decision conditional on the unequal allocation.

To elicit the observer's maximum willingness to pay to punish (MWP) the dictator's unfair behavior, we use a Becker-DeGroot-Marschak mechanism which provides proper incentives for truthful reporting (Becker, DeGroot and Marschak, 1964). The mechanism proceeds in two steps: first we ask the observer to state the maximum amount of money they are willing to pay to levy the 1 euro punishment on the dictator; next, we randomly draw a number, z, between 0 and 1, inclusive. If the observer's stated MWP is at least z, the dictator's earnings are lowered by 1 euro and the observer's earnings are lowered by z euro—i.e., the observer is charged the price z and the dictator is punished. If the observer's stated MWP to pay is below z, neither the dictator's nor the observer's earnings are lowered.[15]

Participants' decisions were collected using the strategy method. Before knowing with whom they were matched—two red group members, two blue group members or one of each—and before knowing which roles would be assigned to their co-players or themselves—dictator, dictatee or observer—each participant submitted their complete contingent strategy in each role. In the role of the dictator, participants chose the equal split or the unfair split for all four possible combinations of red/blue dictatee and red/blue observer. In the role of observer, the maximum willingness to pay to punish was elicited in these same four situations.[16]

[15] To enhance the credibility of this mechanism, participants were informed that if this game were chosen to count, the random draw would be performed in full view of all participants using the on-line randomizing service random.org. To strengthen incentives for truthful reporting, the draw utilizes the full range of *a priori* plausible values for the punishment—i.e., 0.00 euros to 1.00 euros. For a discussion of why these considerations are important, see Plott and Zeiler (2005) and Harrison and Rütstrom (2008).

[16] Because restricting the number of participants in a session to be divisible by 2 (which four of the five games required) and by 3 (which the game analyzed here required) was impractical, participants were instructed that if the dictator game with third party punishment were chosen to count we would randomly

After all participants had completed all five games in their packet, all experimental materials were collected, and the game that was randomly chosen to count was publicly revealed. For this game, participant matchings were then randomly formed, game roles were randomly assigned and outcomes and earnings were determined. Earnings were paid in cash to each participant, separately. Each session lasted approximately one hour.

3.3 Control sessions

In control sessions, participants were not divided into groups and the roles in the games participants played did not involve group distinctions. In all other respects, the sessions were conducted exactly as described above. The strategy method was used, the experiments were conducted using pen and paper, seating was randomized, red and blue pens were provided and the realization of randomness involved in determining the outcome of the game chosen to count was publicly conducted.

4 A simple framework

To obtain testable predictions, we consider a simple framework where: i) all agents derive (weakly) positive utility from unfair behavior being punished (justice preference); ii) the observer internalizes others' preferences; iii) the extent to which observers internalize others' preferences is group-contingent. To simplify notation while capturing the intution of group-contingent preferences, we assume extreme in-group bias in preference internalization: observers completely ignore the preferences of out-group members when making decisions and put strictly positive weight on the preferences of in-group members.

In symbols, use the subscripts d, r and o, to denote dictator, recipient and observer, respectively. Let $\phi_j, j \in \{d, r, o\}$, be the positive utility and agent of type j derives from justice due to punishment being levied against a dictator choosing the unfair allocation. Write an agent's total utility as $U_j = u_j + \phi_j, j \in \{d, r, o\}$ and assume that u_j is a function only of the vector of material payoffs $\Pi = (\pi_r, \pi_d, \pi_o)$. Let $c(p)$ be the cost of one euro of punishment. Let $\alpha \in (0, 1)$ be a parameter capturing how much the observer generally weights other preferences when making decisions ("other-regardingness") and $\beta \in (0, 1)$ be a

form as many 3-person groups as possible to determine outcomes, while the (at most) remaining two participants would be paid a fixed fee of 15 euros. Since participants had no control over whether they would be in a 3-person group, this procedure is still incentive compatible.

parameter capturing how much the observer weights the dictator relative to the recipient.[17] To allow for group-contingent preferences, abusing notation slightly let $\mathbf{G_d}$ and $\mathbf{G_r}$ be indicator functions taking the value of one whenever the dictator or recipient, respectively, share the same group affiliation as the observer.

Restricting attention to the case where the dictator selects the unfair option, we can write the observer's utility from not punishing as:

$$U_{o|\text{not punish}} = \alpha\{\mathbf{G_d}\beta[u_d(8,22,15)] + \mathbf{G_r}(1-\beta)[u_r(8,22,15)]\} + (1-\alpha)[u_o(8,22,15)]$$

The observer's utility from punishing the dictator for the unfair allocation is:

$$\begin{aligned} U_{o|\text{punish}} = \alpha\{\mathbf{G_d}\beta[u_d(8,21,15-c(p)) + \phi_d] + \mathbf{G_r}(1-\beta)[u_r(8,21,15-c(p)) + \phi_r]\} + \\ (1-\alpha)[u_o(8,21,15-c(p)) + \phi_o] \end{aligned}$$

We leave unspecified the precise functional functional forms of u_r, u_d and u_o, making only the following assumptions: i) utility is increasing in own monetary payoffs; ii) ϕ_d is small enough that the dictator prefers to not be punished—i.e., $u_d(8,22,15) > u_d(8,21,15) + \phi_d$; and iii) the recipient's utility is not so increasing in its other arguments to make the recipient prefer no punishment. The last assumption seems justified in light of studies, including our own, where recipients reveal a substantial willingness to spend *their own* money to *directly* punish unfair acts committed against them.

As a simple example, suppose that for the dictator and recipient total utility is simply monetary payoff plus justice utility: $u_j = \pi_j + \phi_j, j \in \{d, r\}$. Then the observer's utility conditional on punishing is:

$$U_{o|\text{punish}} = \alpha[\beta\mathbf{G_d}(21 + \phi_d) + (1-\beta)\mathbf{G_r}(8 + \phi_r)] + (1-\alpha)[15 - c(p) + \phi_o]$$

Conditional on not punishing, the observer's utility is:

$$U_{o|\text{not punish}} = \alpha[\beta\mathbf{G_d} \times 22 + (1-\beta)\mathbf{G_r} \times 8] + (1-\alpha) \times 15$$

[17] This latter parameter might vary because, e.g., the dictator earns the most money irrespective of the observer's punishment decision, or because the recipient is powerless.

To calculate the observer's MWP in this example for each of the four possible combinations of the dictator's and the recipient's group affiliation, we find $c(p)$ at which $U_{o|\text{punish}} = U_{o|\text{not punish}}$. For clarity, in the rest of the paper we subscript MWP by the dictator's group affiliation followed by the recipient's group affiliation—e.g., $MWP_{(in,out)}$ denotes that the dictator was a member of the observer's group, while the recipient was not. The observer's MWP in the four cases are given by:

$$\begin{aligned}
MWP_{(out,in)} &= \tfrac{\alpha}{(1-\alpha)}(1-\beta)\phi_r + \phi_o \\
MWP_{(out,out)} &= \phi_o \\
MWP_{(in,out)} &= \tfrac{\alpha}{(1-\alpha)}[\beta(\phi_d - 1)] + \phi_o \\
MWP_{(in,in)} &= \tfrac{\alpha}{(1-\alpha)}[\beta(\phi_d - 1) + (1-\beta)\phi_r] + \phi_o
\end{aligned}$$

Briefly, notice that since $\phi_d \leq 1$ in this example by assumption—otherwise the dictator would prefer being punished to not being punished—$MWP_{(in,out)} \leq MWP_{(out,out)}$. Next, since ϕ_r, $1-\beta$ and $\frac{\alpha}{1-\alpha}$ are all positive, $MWP_{(out,in)} \geq MWP_{(out,out)}$. Finally, notice that MWP in the case (in, in) can be written as MWP in the case (in, out) plus one other positive term: $\frac{\alpha}{1-\alpha}(1-\beta)\phi_r$. Therefore, we can rank $MWP_{(in,in)} \geq MWP_{(in,out)}$. Putting these rankings together, in this example we have $MWP_{(out,in)} \geq [MWP_{(out,out)}, MWP_{(in,in)}] \geq MWP_{(in,out)}$. Where $MWP_{(in,in)}$ stands in relation to $MWP_{(out,out)}$ depends on the relationship between the dictator's and recipient's justice utilities. If $\beta(\phi_d - 1) + (1-\beta)\phi_r \leq 0$ then $MWP_{(out,out)} \geq MWP_{(in,in)}$, otherwise $MWP_{(out,out)} \leq MWP_{(in,in)}$.

5 Hypotheses

Although our exercise is mostly exploratory, we test several formal and informal hypotheses in our data. First and foremost, by comparing observer's MWP in our control sessions—$MWP_{control}$—to MWP in our treatment sessions, we test whether introducing identity increases the observer's MWP. Since in several dynamic models the maintenance of social norms and cooperation depend on the observers' willingness to punish, this hypothesis sheds light on which environment—fractionalized or homogenous—is more conducive to the survival of such norms.

Hypothesis 1: Introducing identity increases the observer's MWP

Conditional on an affirmative answer to Hypothesis 1, we can refine the treatment-control comparison a bit more and ask in which cases is MWP different in the treatment than in the control. Two obvious competing hypotheses present themselves. On the one hand, it seems intuitively plausible that being thrown together into an unfamiliar, stressful environment like the laboratory could create a *de facto* shared social identity among participants even without explicitly dividing them into groups, in which case one would expect $MWP_{control} = MWP_{(in,in)}$. On the other hand, the relative sterility of the laboratory environment and the explicitly individual monetary incentives may serve to isolate participants from one another, leading participants to define *everybody else* as the out-group, in which case we would expect $MWP_{control} = MWP_{(out,out)}$. This leads to two more, competing, hypotheses:

Hypothesis 2A: $MWP_{control}$ is indistinguishable from $MWP_{(in,in)}$

Hypothesis 2B: $MWP_{control}$ is indistinguishable from $MWP_{(out,out)}$

For our third hypothesis, we restrict attention to treatment sessions and predict a partial ranking of the observer's MWP over the four cases considered there. To construct this ranking, first notice that in all four cases — in-group/out-group dictator/recipient — there is a common tradeoff the observer faces: utility lost from paying the price to punish, $c(p)$, versus the justice utility benefit to the observer, ϕ_o, from a marginal increase in justice. This basic tradeoff is tilted in favor of punishment whenever the observer cares about the recipient's utility and tilted against punishing whenever the observer internalizes the dictator's utility. Consequently, punishment should be highest when the observer internalizes the recipient's utility, but not the dictator's utility.[18]

Hypothesis 3a: *Observers will value punishment the most when the dictator is an out-group member and the recipient is an in-group member.*

Similarly, in the case where the dictator is an in-group member and the recipient is an out-group member, then internalizing the dictator's preferences but not the recipient's tilts the observer's preferences towards *not* punishing. We should therefore expect the least value for punishment in this case.

Hypothesis 3b: *Observers will value punishment the least when the dictator is an in-group member and the recipient is an out-group member.*

Let us now turn from predictions to results.

[18] In our simple example above, Hypothesis 1 can be seen by comparing the expressions for MWP directly. For example, $\mathrm{MWP}_{(\mathrm{out,in})} - \mathrm{MWP}_{(\mathrm{out,out})} = \frac{\alpha}{(1-\alpha)}(1-\beta)\phi_r + \phi_o - \phi_0 = \frac{\alpha}{(1-\alpha)}(1-\beta)\phi_r > 0$ since $0 < \alpha, \beta, \phi_r < 1$ by assumption.

6 Results

In Table 1, we provide descriptive statistics for the treatment and control sessions. Consistent with previous studies using different methodologies and subject pools, we find clear evidence that third parties prefer to punish unfair behavior: a majority of participants in both control and treatment sessions report a strictly positive valuation for the marginal unit of punishment. On average, this valuation ranges widely from around 30 cents (control) to just below 50 cents (treatment, out-group dictator and in-group recipient). We have only limited demographic information, the major exception being gender.[19] Overall, gender composition is quite similar across treatment and control, providing some assurance that randomization into sessions was effective. Nevertheless, in our main analyses we include gender as a control whenever possible.

In Table 2 we report a series of simple OLS regressions related to our first two hypotheses. To account for potential within-session correlation of behavior, in all regressions standard errors are clustered by session unless otherwise noted. The first column of Table 2 pools observers' stated MWPs from all four punishment scenarios in the treatment sessions together with observers' MWPs in the control sessions. The main explanatory variable in this most basic regression is an indicator for treatment. Consistent with Hypothesis 1, we find that the coefficient on the treatment indicator is positive and significant—and substantial in magnitude. The coefficient suggests that introducing group divisions increased observers' value for the opportunity to levy one unit of punishment by 25 percent.

Result 1: *Hypothesis 1 is confirmed. Introducing explicit group divisions significantly increases punishment.*

Having established that explicitly introducing group divisions changes punishment preferences, the Columns 2-5 of Table 2 shed some light on how participants may view the situation *sans* group divisions. Does the laboratory environment create a *de facto* shared social identity so that third-party punishment behavior resembles the (in, in) case in the treatment sessions (Hypothesis 2A)? Or, do individual monetary incentives isolate individuals so that third-party punishment behavior resembles the (out, out) treatment case? The table provides an unequivocal answer: average punishment in the control sessions does not

[19] The fact that all subjects were students living in Rome makes us confident that they were relatively homogenous otherwise—i.e., in terms of age, income, education level, etc.

differ significantly from the (in, in) case (column 2), while there is a substantial and significant difference between punishment in the control sessions and punishment in the (out, out) treatment scenario (column 5).

Result 2: *Hypothesis 2A (2B) is confirmed (rejected).* $MWP_{(in,in)}$ *does not significantly differ from* $MWP_{(control)}$*, while* $MWP_{(out,out)}$ *does.*

We now restrict attention to the treatment session data, and consider how punishment preferences vary within the treatment across the four punishment scenarios. Toward this end, we pool the data from the treatment session scenarios and construct a dataset with four observations per participant: for each individual, the resulting data contain one observation pertaining to each of $MWP_{(in,in)}$, $MWP_{(in,out)}$, $MWP_{(out,in)}$ and $MWP_{(out,out)}$. We then run a simple OLS regression including as explanatory variables a set of dummies for the four separate MWPs—$MWP_{(in,in)}$ being the excluded category. We control for gender by inserting an indicator for being male and, to account for the fact that we have multiple observations per subject, we cluster robust standard errors by session. As an additional check, to account for the notion that it may be particularly aversive, for whatever reason, to punish in-group members, we insert a control for a participant's willingness to pay to directly punish an in-group dictator for an unfair action.[20] We report both specficiations in Table 3.[21]

The estimates in Table 3 are consistent with Hypothesis 3a. We indeed find the highest average valuation for the marginal punishment opportunity in the case where the dictator is an out-group member, but the recipient is an in-group member: $MWP_{(out,in)}$ is about 37% larger than $MWP_{(in,in)}$, the excluded category in the simplest specification (top row). However, on average, we do not find support for Hypothesis 3b. Seemingly contrary to our predictions, the lowest average value for punishment is associated with the in-group dictator, in-group recipient case.

[20] This measure is taken from a dictator game with direct, but no third party, punishment that was one of the four other games in each participant's packet. The game was otherwise identical. In particular, each participant's valuation was elicited, using the same BDM mechanism described, for the opportunity to reduce the dictator's earnings by one euro following an (unfair) unequal money division decision. Each participant's stated maximum willingness to pay for this opportunity to punish is the control we insert.

[21] As an alternative method for handling the issue of multiple observations per participant, we also estimated otherwise-identical individual random effects models (not reported). In these models, the estimated coefficients were identical, and significance levels similar, except that the (in, out) and (out, out) dummy coefficients became significant at the 5% level in both specifications.

However, notice that our hypotheses 3a and 3b are about individual-level rankings. Obviously, averaging over all individuals may be misleading. As a second step, we compute the proportion of individuals whose MWPs over the four scenarios are consistent with each of our hypotheses. The results are reported in Table 4, where both hypotheses 3a and 3b find support: the vast majority—about 85% of participants—valued the punishment opportunity the most when the dictator was an out-group member and the recipient was an in-group member. A similarly large fraction of participants were also consistent with hypothesis 3b: about 83% of them valued least the opportunity to punish an in-group dictator treating an out-group recipient unfairly.

Result 3: *Behavior is consistent with hypothesis 3a on the aggregate level and at the individual level: by both of these measures, the marginal punishment opportunity is the most valued when an out-group member treats an in-group member unfairly. The data are (not) consistent with hypothesis 3b at the individual (aggregate) level.*

7 Discussion

Understanding how third party punishment preferences are shaped by the presence or absence of pre-existing group divisions is an important undertaking. Group-contingent third party punishment, for example, plays a central role in several theoretical models of the the evolution an maintenance of social norms, cooperation and, at least off the equilibrium path, conflict. If third parties have fundamental preferences over punishment aside from the punishment incentives arising from dynamic strategic forces such as (group) reputation, this may determine which equilibria are likely to be played.

A priori, how group affiliation modifies punishment preferences—enhancing or ameliorating inter-group punishment—is not clear. On the one hand, if the scope and expectation of normative behavior is confined within one's group boundaries, as argued by Bernhard et al. (2006) and documented by Banfield (1958), then punishment of out-group members for norm violations may be less severe or even wholly lacking since out-group members violate no covenant through untoward behavior. The implication is that the only case in which we would expect costly, moralistic, punishment would be when all parties to a dispute are members of a common group. On the other hand, if, as social identity theory—starting with Tajfel et al. (1971)—suggests, there is an inherent bias toward in-group members, then a straightforward extrapolation of Chen and Li's (2009) group-contingent preferences model to the context of third party group-contingent punishment suggests individuals will more readily punish out-group members. However, it is not clear in this case what punishment patterns are to be expected when in-group members commit transgressions against out-group members.

Our results lend partial support to both of these stories. On the one hand, and in line with several other studies, our participants clearly exhibit a form of in-group bias in punishment: similar to the findings of e.g. Goette et al. (2006), pooling over recipients' group affiliations, the observers in our experiment generally reported a lower maximum willingness to pay to punish in-group dictators than to punish out-group dictators.[22] Furthermore, on average, willingness to pay to punish was the largest in the case where an out-group dictator treated an in-group recipient unfairly. This latter pattern can be interpreted as group-based defensive behavior. Though necessarily speculative, group-based evolution may have supported such a behavioral trait. The intuition is the familiar folk theorem logic: as long as punishment is harsh enough, levied by *someone*, and conditional on bad behavior crossing group boundaries, peace can be sustained in equilibrium. Punishment, even of random members of an offender's group, may then, in turn, induce this group to begin enforcing peaceful behavior of its members to prevent the escalation of conflict.

On the other hand, seemingly inconsistent with the notion that in-group bias is the whole story, we find a substantial willingness to spend money to punish in-group dictators who treat out-group recipients unfairly. From an evolutionary point of view, even such behavior could make sense: group conflict could be prevented if groups managed to convince each other that offenders are sufficiently punished to deter further potential transgressors within their own group. Although in equilibrium both punishment strategies will induce peace among groups, behavioral patterns off the equilibrium path differ dramatically. Harsh punishment of out-group offenders may lead to inter-group reprisals and conflict spiraling out of control, while containing intra-group punishment leads to inter-group docility.

The strongest, most consistent, pattern in our data—that punishment is valued by third parties when an out-group member treats an in-group member unfairly—is consistent with

[22] Note that Goette et al. (2006) compare punishment behavior towards in- and out-group norm violators on the one hand, and in- and out-group victims on the other. They do not compare, for instance, the case of an in-group violator matched with an in-group victim to the case of an in-group violator matched with an out-group victim.

the former of these stories, i.e. inter-group conflict may spiral out of control off the equilibrium path. Future research can directly test in a repeated-game setting if group contingent punishment preferences, most obviously revealed by our finding that the maximum willingness to punish unfair out-group members in general exceeds that to punish in-group fellows, fuel conflicts from an initially inter-personal level to escalate to group conflict.

8 Conclusion

Through a laboratory experiment we introduce artificial group identity in a one-shot dictator game with third-party punishment and test if and how preferences for justice, measured as willingness to pay for punishment of an unfair act, are influenced by identification into minimal groups (Tajfel and Turner, 1986).

The first novelty of our paper hinges on the punishment mechanism we adopt. Differently from many related studies where punishers may undo the unfair dictator's decision (see, for instance, Goette et al., 2012 and Bernhard et al., 2006), we elicit the willingness to pay to levy an amount of punishment which has been fixed at a low level. This strategy allows us to capture how observers' preference for justice varies according to the group affiliation of the transgressor and the victim by ruling out other confounding factors influencing punishment (e.g., responsibility for undoing the injustice and/or desire to deter unfair behavior). Specifically, in our setting observers cannot levy an amount of punishment necessary to restore justice, nor can the fixed amount of punishment available substantially alter the transgressor's payoff.[23]

A second contribution of our paper consists in the comparison between treatment sessions where group identity is induced *vis-a-vis* control sessions in which it is not. Such comparison allows us to isolate the impact the *introduction* of artificial minimal groups on preferences for punishment. As a third contribution, within the treatment sessions we vary all players' group affiliations independently and look at how the desire to punish changes when the perpetrator, the victim, both or neither are members of the bystander's group.

[23] In our data, dictators' behavior does not seem to be driven by actual group-contingent punishment patterns. Conditional on the group affiliation of the recipient, the proportion of dictators choosing the unfair allocation does not vary with the group affiliation of the observer: when the recipient is an out-group member, exactly 55 percent of dictators choose the unfair allocation both when the observer is from the out-group and when the observer is from the in-group; when the recipient is an in-group member, 45 (42) percent of dictators choose the unfair allocation when the observer is from the in-group (out-group) (p=0.235).

Our findings suggest identity matters for punishment preferences since introducing artificial group divisions significantly increases the willingness to punish unfair acts. Restricting the analysis to sessions where group identity is induced, we find punishment is valued the most when an out-group member treats an in-group member unfairly while its is valued the least when both the perpetrator and the victim belong to the same group of the observer.

Consistent with the both the literature on in-group bias (see, among others, Chen and Li, 2009 and Choi and Bowles, 2009) and group defensive behavior (Goette et al., 2012 and Bernhard et al., 2006), participants in our experiment prefer to punish out-group perpetrators more than in-group perpetrators. However, we also show evidence of a non-vanishing (although lower-ranked) preference for punishing in-group participants who behave unfairly towards out-group victims. The two findings are not inconsistent since, under an evolutionary perspective, the punishment of harmful behavior of in-group fellows towards out-group members prevents the escalation of costly inter-group conflicts while in-group favoritism sustains group bonds and deters out-group aggressions of in-group fellows (Fearon and Laitin, 1996).

References

[1] Abbink, Klaus, and Benedikt Herrmann (2009), "Pointless Vendettas", Manuscript.

[2] Akerlof, George and Rachel Kranton (2000). "Economics and Identity." *The Quarterly Journal of Economics*, 115, pp. 715-773.

[3] Akerlof, George and Rachel Kranton (2005). "Identity and the Economics of Organizations." *The Journal of Economic Perspectives*, 19, pp. 9-32

[4] Banfield, Edward (1958). *The Moral Basis of a Backward Society.* Glencoe, IL: The Free Press.

[5] Bernhard, Helen, Urs Fischbacher and Ernst Fehr (2006). "Parochial Altruism in Humans." *Nature*, 442(24), pp. 912-915.

[6] Becker, Gordon, Morris DeGroot and Jacob Marschak (1964), "Measuring utility by a single-response sequential method." *Behavioral Science*, 9(3), pp. 226-232.

[7] Bohnet, Iris and Richard Zeckhauser (2004), "Trust, Risk and Betrayal." *Journal of Economic Behavior and Organization*, 55, pp. 467-484.

[8] Bornstein, Gary (1992), "The Free Rider Problem in Intergroup Conflicts over Step-Level and Continuous Public Goods", *Journal of Personality and Social Psychology* 62: 597-606.

[9] Bornstein, Gary (2003), "Intergroup Conflict: Individual, Group, and Collective Interests", *Personality and Social Psychology Review* 7(2): 129-145.

[10] Butler, Jeffrey and Paola Giuliano and Luigi Guiso (2011), "Cheating in the Trust Game", Working Paper, Einaudi Institute for Economics and Finance.

[11] Carlsmith, Kevin M., John M. Darley, and Paul H. Robinson (2002), "Why do we punish? Deterrence and Just Deserts as Motives for Punishment", *Journal of Personality and Social Psychology*, 83(2), pp. 284-299.

[12] Carpenter, Jeffrey, and Peter Hans Matthews (2010), "Norm Enforcement: The Role of Third Parties," *Journal of Institutional and Theoretical Economics*, 166, pp 239-258.

[13] Charness, Gary, Luca Rigotti, and Aldo Rustichini (2007). "Individual behavior and group membership." *American Economic Review*, 97, pp. 1340-1352.

[14] Chen, Roy and Yan Chen (2011), "The Potential of Social Identity for Equilibrium Selection", *American Economic Review*, 101(6), pp. 2562-2589.

[15] Chen, Yan and Sherry Xin Li (2009), "Group Identity and Social Preferences." *American Economic Review*, 99(1), pp. 431-457.

[16] Choi, Jung-Kyoo and Samuel Bowles (2007), "The Coevolution of Parochial Altruism and War", *Science* 318, pp. 636-

[17] Darwin, Charles (1873). *The Descent of Man.* Appleton, New York.

[18] De Cremer, David and van Vugt, Mark (1999), "Social Identification Effects in Social Dilemmas: A Transformation of Motives", *European Journal of Social Psychology* 29: 871- 893.

[19] De Dreu, Carsten K. W., Lindred L. Greer, Michel J. J. Handgraaf, Shaul Shalvi, Gerben A. Van Kleef, Matthijs Baas, Femke S. Ten Velden, Eric Van Dijk, and Sander W. W. Feith (2010). "The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans." *Science*, 328: pp. 1408-1411.

[20] Dufwenberg, Martin, Uri Gneezy, Werner Güth and Eric van Damme (2001), "Direct versus Indirect Reciprocity." *Homo Oeconomicus*, 18, pp. 19-30.

[21] Eckel, Catherine C., and Philip J. Grossman (2005), "Managing Diversity by Creating Team Identity." *Journal of Economic Behavior and Organization.* 58 (3), pp. 371-392.

[22] Fearon, James D. and David D. Laitin (1996), "Explaining Interethnic Cooperation". *American Political Science Review* 90:4 (December), pp. 715-35

[23] Fehr, Ernst and Urs Fischbacher (2004), "Third-Party Punishment and Social Norms." *Evolution and Human Behavior*, 25, pp. 63-87.

[24] Goette, Lorenz, David Huffman and Stephan Meier (2006), "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups." *The American Economic Review*, 96(2), pp. 212-216.

[25] Goette, Lorenz, David Huffman and Stephan Meier (2012), “The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups.” *American Economic Journal: Microeconomics*, 4(1), pp. 101-115.

[26] Guala, Francesco, Luigi Mittone and Matteo Ploner (2009), “Group Membership, Team Preferences, and Expectations”, University of Trento, CEEL Working Papers 0906.

[27] Harrison, Glenn and E. Elisabet Rutström (2008), “Risk Aversion in the Laboratory,” in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments.*

[28] Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwind Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker (2006). “Costly Punishment Across Human Societies”. *Science*, 23:312, 5781, pp. 1767 - 1770

[29] Herrmann, Benedikt, Christian Tho´oni and Simon G´achter (2008), “Antisocial Punishment Across Societies.” *Science*, 319, pp. 1362-1367

[30] Hugh Jones, David and Martin A. Leroch (2011). “Reciprocity towards Groups,” *Competitive Advantage in the Global Economy (CAGE) Online Working Paper Series n. 51.*

[31] Lewisch, Peter, Stefania Ottone and Ferruccio Ponzano (2010), “Free-riding on altruistic punishment? An experimental comparison of third-party punishment in a stand-alone and in an in-group environment,” *Institute of Public Policy and Public Choice (POLIS) Working Papers.*

[32] McLeish, Kendra N. and Robert Oxoby (2007). “Identity, Cooperation, and Punishment,” IZA Discussion Paper No. 2572

[33] Mumendey, Amélie, and Sabine Otten (1998), “Positive-negative asymmetry in social discrimination”, *European Review of Social Psychology* 9, pp. 107-143.

[34] Plott, Charles and Kathryn Zeiler (2005), “The Willingness to Pay Willingness to Accept Gap, the ‘Endowment Effect,’ Subject Misconceptions, and Experimental Procedures for Eliciting Valuations,” *American Economic Review*, 95, pp. 530-545

[35] Tajfel, Henri, Claude Flament, Michael G. Billig, and Robert F. Bundy (1971), "Social Categorization and Intergroup Behavior," *European Journal of Social Psychology*, 1, pp. 149-177

[36] Tajfel, Henri and John Turner (1986), "The Social Identity Theory of Intergroup Behavior," in Stephen Worchel and William Austin, eds., *The Social Psychology of Intergroup Relations*, Chicago: Nelson- Hall.

[37] Yamagishi, Toshio and Toko Kiyonari (2000), "The Group as the Container of Generalized Reciprocity", *Social Psychology Quarterly* 63(2): 116-132.

[38] Yamagishi, Toshio and Nobuhiro Mifune (2008), "Does Shared Group Membership Promote Altruism?", *Rationality and Society* 20(1): 5-30.

[39] Yamagishi, Toshio and Nobuhiro Mifune (2009), "Social Exchange and Solidarity: In-Group Love or Out-Group Hate?", *Evolution and Human Behavior* 30(4): 229-237.

Tables

Table 1: Descriptive Statistics

	Control	Treatment
$MWP_{(Control)}$ > 0 (dummy)	0.51	
	(0.05)	
$MWP_{(in, in)}$ > 0 (dummy)		0.54
		(0.05)
$MWP_{(out, in)}$ > 0 (dummy)		0.70
		(0.05)
$MWP_{(in, out)}$ > 0 (dummy)		0.62
		(0.05)
$MWP_{(out, out)}$ > 0 (dummy)		0.61
		(0.05)
$MWP_{(Control)}$	0.31	
	(0.04)	
$MWP_{(in, in)}$		0.33
		(0.04)
$MWP_{(out, in)}$		0.47
		(0.04)
$MWP_{(in, out)}$		0.39
		(0.04)
$MWP_{(out, out)}$		0.40
		(0.04)
Male (dummy)	0.55	0.57
	(0.05)	(0.05)
Observations	96	98

Notes: [1] Standard errors in parentheses. [2] $MWP_{(Control)}$ is the observer's stated maximum willingness to pay to levy a one euro punishment on the dictator following the unfair division of money between the dictator and recipient in the control sessions. The other subscripts refer to the particular combination of (dictator group, recipient group) considered in the treatment sessions. A subscript of "out" denotes not being a member of the observer's group, while a subscript of "in" denotes belonging to the observer's group. [3] In a student sample such as this, demographics are *a priori* unlikely to be powerful predictors of behavior. The major exception is gender, which we include in our analyses.

Table 2: Treatment Effect on Observer’s Punishment Preferences

	(1)	(2)	(3)	(4)	(5)
		Control vs. Treatment Scenario			
	All (Pooled)	(in, in)	(in, out)	(out, in)	(out, out)
Treatment (dummy)	0.09**	0.02	0.08	0.16***	0.09**
	(0.04)	(0.03)	(0.05)	(0.04)	(0.04)
Male (dummy)	-0.07	-0.07	-0.08	-0.09	-0.05
	(0.05)	(0.04)	(0.06)	(0.06)	(0.04)
Constant	0.35***	0.35***	0.35***	0.36***	0.34***
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)
Observations	488	194	194	194	194
R-squared	0.02	0.01	0.02	0.05	0.02

Notes: [1] Column 1 pools data from the control sessions together with observations from all four punishment scenarios in the treatment sessions, resulting in four observations per treatment session participant. To account for multiple observations per individual, while still allowing us to control for gender, a fixed trait, we estimate and report in Column 1 an individual random-effects model. [2] Columns 2-5 include only one observation from one punishment scenario for each individual, with the specific scenario listed in the column heading. Accordingly, we estimate and report simple OLS regressions. [3] Robust standard errors, clustered by session, in parentheses. [4] Each punishment scenario is labeled with the convention of (dictator group, recipient group) relative to the observer so that, e.g., (in, out) denotes the scenario where the dictator and observer are members of the same group (in-group), while the recipient is not a member of the observer’s group (out-group).

Table 3: Punishment Preferences in Treatment Sessions Only

$\text{MWP}_{(\text{recipient group, dictator group})}$ Within Treatment Sessions, Relative to (in, in)							
(in, out)	(out, in)	(out, out)	Male	MWP for direct punishment of own group	Constant	Obs	R^2
0.06*	0.14***	0.07*	-0.08	--	0.38***	392	0.02
(0.03)	(0.02)	(0.04)	(0.08)		(0.06)		
0.06*	0.14***	0.07*	-0.04	0.45***	0.16**	388	0.24
(0.03)	(0.02)	(0.04)	(0.06)	(0.06)	(0.05)		

Notes: [1] The table reports simple OLS estimates using treatment session data pooled over scenarios to generate a dataset containing one observation per individual per punishment scenario. To account for multiple (4) observations per individual, we cluster standard errors by session. We also estimated individual random effects models of both specifications, but the results were similar so we report only the simpler OLS models. [2] Controls include: a set of dummies for the four possible (dictator group, recipient group) punishment scenarios—the excluded category being (in, in); a dummy for gender. [3] In the second specification reported, we add a control for how aversive punishing one’s own group, generally. The variable “MWP for direct punishment of own group” is the participant’s stated willingness to pay to reduce the dictator’s earnings by one-euro when playing the role of recipient in a dictator game with direct punishment - i.e., where the recipient him/herself is the sole punisher. We lose one individual, or four observations, by inserting this control. [4] Robust standard errors, clustered by session, appear in parentheses.

Table 4: Proportion of Individuals Consistent with Hypotheses 3a, 3b

	(1)	(2)	(3)
	Max MWP = MWP$_{(out, in)}$	Min MWP = MWP$_{(in, out)}$	Max MWP = MWP$_{(out, in)}$ **and** Min MWP = MWP$_{(in, out)}$
	0.85	0.83	0.78
	(0.04)	(0.04)	(0.04)
Obs	98	98	98

Notes: [1] Table reports the proportion of individuals in treatments sessions whose valuations for the opportunity to punish across the four different scenarios considered are consistent with Hypotheses 3a and 3b. An individual is consistent with Hypothesis 3a if his or her MWP$_{(out, in)}$ = max{MWP$_{(out, in)}$, MWP$_{(in, in)}$, MWP$_{(in, out)}$, MWP$_{(out, out)}$}; an individual's punishment valuations are consistent with Hypothesis 3b if MWP$_{(in, out)}$ = min{ MWP$_{(out, in)}$, MWP$_{(in, in)}$, MWP$_{(in, out)}$, MWP$_{(out, out)}$}. [2] Standard errors, clustered by session, appear in parentheses.