Social Identity and Punishment

Jeffrey V. Butler
EIEF

Pierluigi Conzo
University of Turin & CSEF

Martin A. Leroch
University of Mainz

This version: March 19, 2014

Abstract

Third party — or, bystander — punishment is crucial for sustaining cooperative behavior. Through laboratory experiments we investigate the interaction between group identification and a bystander's punishment preferences by inducing minimal groups and giving a bystander the opportunity to levy a fixed amount of punishment on the perpetrator of an unfair act towards 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. For predictions, we construct two separate frameworks differing by whether the primary effect of group identity is to create an empathetic bond between in-group members or to affect the weights placed on others' money earnings in a distributional social preferences model. The two frameworks yield starkly different ordinal predictions about the bystander's value for punishment across two cases: i) when the perpetrator and victim are both members of the bystander's group; ii) when only the victim is an in-group member. The empathetic bond framework predicts that punishment will be more highly valued in the latter case, while the distributional preferences framework suggests the opposite. Our data support the predictions of the first. Finally, we conduct control sessions where groups are not induced and find that bystanders tend to treat others as in-group members unless specifically divided into distinct groups.

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

1 Introduction

Theoretically, third party punishment may be beneficial or detrimental to societies. On the one hand, 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[1] entering into disputes to punish transgressors on the behalf of those directly affected may prolong and extend conflicts beyond an initially limited scope.[2]

One potential determinant of punishment patterns 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). More recently, 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. Crucial in both cases is discriminatory behavior towards out-group members.

Empirically, results have been mixed and apparent differences in findings across studies are difficult to interpret. Consider two prominent exemplary studies: Bernhard, et al. (2006) and Goette, et al. (2012). The former study features a strategically simple setting — a dictator game with third party punishment — and players from different real-world indigenous groups. The primary finding is that the *victim's group* matters most for punishment: norm violations harming a member of the bystander's own group are punished more harshly irrespective of the perpetrator's group affiliation. In the latter study, the same pattern is documented using different real-world social groups and a more complex strategic setting.

[1] Throughout the paper, reflecting the state of the literature, we use several terms more or less interchangeably to refer to an individual who is not directly materially affected by a transgression or injustice: third party, bystander and/or observer. Such third party punishment also sometimes goes by the moniker "altruistic punishment."

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

However, a strikingly different pattern emerges in the context of experimentally-induced minimal groups, where "...the group affiliation of the victim has no influence on punishment" (Goette et al. 2012, p. 111). To reconcile these disparate results, the authors of the latter study suggest there is a fundamental difference between real-world social groups and experimentally-induced minimal groups that should give researchers pause in extrapolating from results obtained with minimal groups — group-contingent empathy. The authors posit that social ties among real-world social groups engender an empathetic bond between members, whereas minimal groups which lack social ties by definition do not engender empathy. This is an intriguing claim which warrants closer inspection.

Toward this end, we conduct an experiment to examine bystander punishment preferences. To minimize potential strategic confounds, we use the simplest feasible game — a one-shot anonymous dictator game with third party punishment (Fehr and Fischbacher, 2004). To shed light on whether social ties are a necessary prerequisite for engendering group-contingent empathy, our study features experimentally-induced minimal groups. To understand whether there are testable implications of a model incorporating an empathetic bond as distinct from, e.g., distributional preferences models we construct two frameworks. Our first framework explicitly incorporates empathy. Our second framework adapts inequality aversion to our third party punishment setting. The two frameworks yield starkly different predictions about punishment preferences which we then test using our experimental data. We find that punishment preferences are broadly consistent with the first framework, which allows for group-contingent empathy, and not consistent with the latter distributional preferences framework: bystanders place a higher value on punishing an outsider for treating an insider unfairly than they do on the opportunity to punish an insider for treating an insider unfairly. More generally, punishment levied on outsiders is valued more highly than punishment levied on members of the bystander's own group. Finally, comparing the average value placed on punishment opportunities 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.

In shedding light on this particular question, our design introduces a novel measure of punishment preferences. This measure allows us to ameliorate confounds in the existing literature which may also be driving differences in results across studies. Specifically, one reason different studies of the interaction between bystanders' punishment preferences and

social identity have produced seemingly inconsistent results may be methodological. All experimental 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 transgressors' earnings. We would argue that bystanders' punishment decisions in this setup are a function of at least three components: i) a value judgment about how wrong the act being punished is — i.e., 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. Observed punishment may depend particularly strongly on the second and third components when, as in all studies we know of, the range of feasible punishment is substantial.[3] As bad behavior is typically measured by an unequal money division, reducing the transgressor's payoff sufficiently can restore earnings equality and, moreover, nullify a potential transgressor's individual monetary incentives for malign behavior.[4]

While all three mentioned components of punishment decisions are interesting in their own right, group identification processes likely affect each of the three in different ways and to different degrees. This presents an obvious confound to interpreting results across studies if it is not known whether and to what extent each component is affected by identification processes. Ideally, one would like to isolate the effect of identity on each component of bystanders' punishment preferences — the approach we take here by focusing on the moral disgust component.

The remainder of the paper is organized as follows. First, we discuss closely related literature. Then, the experimental design and procedures are detailed. In Section 4 we discuss competing theories of punishment preferences and construct two separate models yielding contrasting predictions. In Section 5 we provide our formal hypotheses. In Section 6 we present the results of our experiment, which we briefly discuss in Section 7. In the final section, we summarize our findings and provide concluding remarks.

[3]This obviously does not intend to say that not all three components are applicable when punishment is constrained to be small, rather that the latter two should be relatively less important in such circumstances.

[4]Prior research suggests the importance of all three components in a fixed-price punishment setting. Lewish, Ottone and Ponzano (2010), for example, document that individuals each levy less punishment when more than one person can punish, ostensibly because responsibility is made more diffuse even if, as is likely, the moral disgust component is unchanged by the addition of potential punishers.

2 Closely related literature

A result common to many existing studies on social identity is that maintenance and enforcement of social norms and, more generally, altruistic behavior is characterized by in-group bias:[5] 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),[6] 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).[7] 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). Recent findings imply that group-based distributional preferences as well as beliefs and strategic incentives play a crucial role for the existence of in-group bias (Ockenfels and Werner, 2014). In order to single out the role preferences play, in our experimental design (detailed below) we exclude the possibility of dynamic and strategic interests by implementing a simple one-shot structure between anonymous subjects.[8]

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

[5] 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).

[6] 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.

[7] For lab experiments on costly punishment see, among others, Fehr and Fischbacher (2004) and Henrich et al. (2006). For a field experiment on the issue, see Balafoutas and Nikiforakis (2012).

[8] In contrast to Goette et al. (2006, 2012), we therefore not only rule out strategic incentives deriving from real-world groups (e.g., reputation), but also strategic incentives endogenously deriving from repeated play.

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. Despite any lack of material incentives to do so, on average 13% of the choices were to destroy the other group's endowment. 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. 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. Also, other experiments did not replicate this pattern (see e.g. Halevy et al., 2008).

How the group affiliations of perpetrators and victims factor into third parties' punishment decisions is not fully understood, and existing results seem difficult to interpret. 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. As mentioned in the introduction, Goette et al. (2012) explain this difference by appealing to group-contingent empathy which, according to the authors, occurs only in real-world groups and not in minimal groups.

In this paper, we contribute to the literature on third party punishment and social identity in several ways. First of all, we provide evidence on whether experimentally-induced minimal groups *can* give rise to group-contingent empathy.

Our second contribution is to introduce a novel measure of punishment preferences that allows us to isolate the moral disgust component to a greater extent than has been possible in previous work. In all of the mentioned studies, experimenters fixed the price of punishment and participants chose the amount of punishment to levy. We, in contrast, fix

the amount of punishment bystanders 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. Moreover, the amount of punishment that could be levied in the mentioned studies 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 responsibility 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 that we impose marginal punishments. To elaborate, we restrict punishment to be small for two reasons. First, since bystanders cannot unilaterally impose a "fair" outcome for the directly involved parties, restricting punishment in this way should minimize the effect responsibility for others' *outcomes* may exert on punishment preferences. Second, restricting punishment to be small removes bystanders' ability to nullify the strong individual financial incentives for bad behavior, thereby ameliorating the influence that a desire to deter transgressions might otherwise have on preferences. By minimizing both the responsibility and deterrence motives for bystander punishment, we intend to isolate as much as possible the moral disgust component of punishment preferences. Consequently, in our analysis we interpret bystanders' valuations for the opportunity to punish as relatively clean measures of their value judgments about how wrong particular situations are (moral disgust).

Our third contribution comes from implementing 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 bystanders' punishment preferences.

Our fourth 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. 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.[9] 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.[10] 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. To the best of our knowledge, we are the first to provide clean, incentive compatible evidence on how the bystander's relationship with the directly involved parties affects moral disgust.

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: LUISS Guido Carli University and the University of Rome La Sapienza.[11] Six treatment sessions were conducted in which a total of 100 students took part.[12] A total of 96 students took part in six control sessions. In treatment sessions, participants were randomly divided

[9] Note that Choi and Bowles (2007) argue that (in-group) altruism and (inter-group) war may have co-evolved.

[10] 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.

[11] 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.

[12] 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.

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.[13]

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.[14] Each participant drew one chip from the bag which determined his or her group. Participants were then given their experimental packets (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 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."[15] 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

[13] 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.

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

[15] 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, 2009). We chose to avoid this specific technique here in order to avoid confounding group-contingent preferences with generalized reciprocity.

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.[16] 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 sequential moves game of complete and perfect information. The game involves three players: a dictator, a recipient and an observer.[17] 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 euro endowment with the recipient. We restrict the dictator's set of available options to two: i) divide the sum evenly, so 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

[16]The other four games were: i) a dictator game with direct punishment; and ii) three simple games which, together, provide a measure of "betrayal aversion" (Bohnet and Zeckhauser, 2004). These latter three games are: i) a binary trust game; ii) a "risky dictator" game in which one member of a pair chooses between a certain outcome and an outcome determined by a randomizing device; and iii) an individual decision involving a choice between a certain money outcome and a risky outcome.

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

either choice, previous research suggests that the first allocation is viewed as "fair" while the latter, unequal, allocation, is viewed as unfair. 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 draw a number, z, uniformly distributed on the set $\{0.00, \ldots, 1.00\}$. 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.[18]

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, recipient 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 recipient and red/blue observer. In the role of observer, the maximum willingness to pay to punish was elicited in these same four situations.[19]

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

[18]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 online 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).

[19]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 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.

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 Theoretical frameworks and predictions

In this section we provide an intuitive account of our two theoretical frameworks, relegating to the Appendix a more detailed description of each model as well as examples. The importance of these models is that even though they are both quite general they yield alternative testable predictions about how observers' MWP will vary with the group affiliations of the dictator and the recipient. We focus on the observer's valuation for punishment following the dictator's unfair allocation decision. For ease of exposition, in the rest of the paper we subscript MWP by the dictator's group affiliation followed by the recipient's group affiliation. For example, $MWP_{(out,in)}$ denotes that the dictator is not a member of the observer's group while the recipient is.

Our first "empathetic bond" model incorporates empathy directly by allowing the dictator's and recipient's *preferences* to enter into the observer's utility function. We assume all agents have an intrinsic preference for justice so that, putting aside monetary considerations, they would all prefer that unfair acts be punished. We make empathy group-contingent by allowing the weights the observer places on the dictator's and the recipient's preferences to be group-contingent. The central assumption we make is that the dictator prefers not to be punished while the recipient prefers that the observer levy punishment. This allows our empathic bond model to generate the following prediction:

$$MWP_{(out,in)} \geq MWP_{(in,in)}.$$

Intuitively, when both the dictator and the recipient are in-group members their pref-

erences for and against punishment offset each in the observer's utility calculations to a greater extent than when only the recipient shares a group affiliation with the observer. As we will see, this is the key prediction separating our two models.

The second model we construct is "third party inequality aversion." Building on Chen and Li (2009), in this model we assume that group identity affects the weights observers place on others' experimental *earnings*. We allow the weights the observer places on others' earnings to vary by group (in-group/out-group) and, for generality, by role (dictator/recipient). The key assumption we make is that for each role in-group members' earnings receive more weight than out-group members' earnings. Denote role j's earnings by $\Pi_j, j \in \{d, r, o\}$ and the weight the observer places on the dictator's (recipient's) earnings by β_{G_d} (γ_{G_r}), where the subscripts G_d and G_r refer to the group affiliation of the dictator or recipient, respectively. We define *effective inequality* as the absolute value of the difference in the dictator's and recipient's weighted earnings: $|\beta_{G_d}\Pi_d - \gamma_{G_r}\Pi_r|$.[20] To lend empirical content to this model, we assume the observer's utility function is a weighted average of two components: i) utility from the observer's own money earnings; and ii) disutility stemming from effective inequality. We assume this final weighting parameter—$\alpha \in [0, 1]$—is individiual specific so that the observer's utility is: $U_o = (1-\alpha)\Pi_o - \alpha|\beta_{G_d}\Pi_d - \gamma_{G_r}\Pi_r|$. Our primary prediction from this model is that:[21]

$$MWP_{(out,in)} \leq MWP_{(in,in)}.$$

Intuitively, this result can be seen by considering two cases. In the first case, the observer places sufficient weight on the recipient's earnings to make effective inequality always favor the recipient irrespective of the dictator's group affiliation. In this case, lowering the dictator's earnings even further by punishing *increases* effective inequality and lowers the observer's earnings—a pattern the observer will never pay a strictly positive sum to implement. Consequently, in this case $MWP_{(out,in)} = 0 = MWP_{(in,in)}$. In the second case, the observer places a small enough weight on recipient's earnings to make effective inequality favor the dictator irrespective of whether punishment is levied and irrespective of the

[20] Note that we can allow for the observer's earnings to enter into this definition as well without changing our primary prediction, as discussed in the Appendix.

[21] There is one *a priori* unlikely configuration of parameters in which our primary prediction may not hold. Specifically, when both of the ratios $\frac{\beta_{in}}{\gamma_{in}}$ and $\frac{\beta_{out}}{\gamma_{in}}$ fall within the small interval $(\frac{8}{22}, \frac{8}{21})$, general conclusions about how MWP relates to β_{G_d} and therefore about the relationship between $MWP_{(out,in)}$ and $MWP_{(in,in)}$ cannot be drawn. Rather, these relationships will depend on the specific values of the observer's preference parameters $\alpha, \beta_{in}, \beta_{out}$ and γ_{in}. We discuss this case at length in the Appendix.

dictator's group affiliation. In this case, it can be shown that: $MWP_{(G_d,G_r)} = \frac{\alpha}{1-\alpha}\beta_{G_d}$. Notice that this expression is positively related, and proportional, to the weight placed on the *dictator's* earnings and that the weight placed on the recipient's earnings, γ_{G_r}, plays no role. Because the weight placed on the dictator's earnings is larger for in-group dictators by assumption, it immediately follows that $MWP_{(out,in)} < MWP_{(in,in)}$. This latter result is the most counterintuitive, but follows directly from the fact that since the amount of actual punishment is fixed the amount of effective inequality reduction associated with punishing is proportional to the weight placed on the dictator's earnings.

Summing up, the two models we consider make opposite predictions. The stark contrast in predictions between our two models—the former incorporating empathy, the latter a straightforward adaptation of existing distributional preferences models—allows us to provide evidence for or against the existence of an empathic bond in the context of minimal groups/identities. Before turning to our results, we provide several additional formal testable hypotheses.

5 Formal hypotheses

While our exercise is mostly exploratory, we test several formal and informal hypotheses using 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 evolutionary 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.[22]

Hypothesis 1: Introducing identity increases the observer's valuation of punishment: $MWP_{treatment} \geq MWP_{control}$.

Conditional on an affirmative answer to Hypothesis 1, we can refine the treatment-control comparison a bit more and ask in which treatment cases MWP differs from the control. Two obvious competing hypotheses present themselves. On the one hand, it seems

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

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*: Subjects in the lab natively perceive themselves as sharing commonalities:* $MWP_{control}$ *is indistinguishable from* $MWP_{(in,in)}$.

Hypothesis 2b*: Subjects in the lab natively perceive themselves as not sharing commonalities:* $MWP_{control}$ *is indistinguishable from* $MWP_{(out,out)}$.

For our next pair of hypotheses, we restrict attention to treatment sessions and test the key prediction separating our empathetic bond model from our model based on group-contingent distributional preferences. Fixing the recipient's group, and hence γ, third party inequality aversion predicts the observer's MWP should be weakly increasing in the weight placed on the dictator's earnings, β, for most values of the ratio $\frac{\beta}{\gamma}$. One implication is that we should expect $\text{MWP}_{(\text{in,in})} \geq \text{MWP}_{(\text{out,in})}$. Relative to (out, in), the case (in, in) features a higher β—in-group dictators receive higher weight than out-group dictators—without changing γ. On the other hand, our empathy-based model predicts the opposite relationship between the observer's MWP across these two cases. We thus have two more, competing, hypotheses:

Hypothesis 3a: *In line with the third-party inequality aversion model, punishment of in-group dictators is valued more highly than punishment of out-group dictators when an in-group member is treated unfairly:* $MWP_{(in,in)} \geq MWP_{(out,in)}$.

Hypothesis 3b: *In line with the empathetic bond model, when the recipient is an in-group member observers place a lower value on the opportunity to punish when the dictator is from the in-group dictators than when the dictator is from the out-group:* $MWP_{(in,in)} \leq MWP_{(out,in)}$.

Anticipating the outcome that Hypothesis 3b is accepted and Hypothesis 3a is rejected, we next use our empathetic bond model to generate two more ancillary hypotheses about

how the observer's MWP should vary with the group affiliation of the dictator and recipient. To construct these additional hypotheses, 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 whatever utility increase the observer gets from a marginal increase in justice. This basic tradeoff is tilted in favor of punishment whenever the observer cares about the recipient's utility—who prefers punishment—and tilted against punishing whenever the observer internalizes the dictator's utility, who prefers no punishment. Consequently, punishment should be highest when the observer internalizes the recipient's utility, but not the dictator's utility.

Hypothesis 4a: *Observers will value punishment the most when the dictator is an out-group member and the recipient is an in-group member:* $MWP_{(out,in)} \geq MWP_{otherwise}$.

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 4b: *Observers will value punishment the least when the dictator is an in-group member and the recipient is an out-group member:* $MWP_{(in,out)} \leq MWP_{otherwise}$.

Let us now turn from predictions to results.

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).[23] When

[23]We have only limited demographic information, the major exception being gender. The fact that all subjects were students living in Rome makes us confident that they were relatively homogenous otherwise

looking at the distribution of the stated MWPs by treatment (Figure 1), we find that in all punishment scenarios MWPs are roughly bimodal, with modal values of zero and one. Still, a non-trivial share of MWPs—typically around 30 percent—are interior. We account for subjects' concentration at extreme values in the probit and tobit estimates introduced below.

<<Figure 1 about here >>

In Table 2a 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 2a pools observers' stated MWPs from all four punishment scenarios in the treatment sessions together with observers' MWPs in the control sessions and includes an individual random effect accounting for multiple observations per participant. The main explanatory variable in this most basic regression is an indicator for treatment.

<<Table 2a about here >>

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

Consistent with Hypothesis 1, we find that the coefficient on the treatment indicator is positive, 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. This result is robust to both a tobit regression accounting for censoring in the dependent variable (Table 2b) and to a probit estimation for the probability of a positive MWP, i.e. $Pr(MWP > 0)$ (Table 2c).

<<Table 2b about here >>

<<Table 2c about here >>

Having established that explicitly introducing group divisions changes punishment preferences, columns 2-5 of Tables 2a-c shed some light on how participants may view the situation *sans* group divisions. Does the laboratory environment create a *de facto* shared

— i.e., in terms of age, income, education level, etc. Note that gender composition is quite similar across treatment and control, providing some assurance that randomization into sessions was effective. Nevertheless, we repeat all the empirical analysis including gender as a control and the main findings do not change (results are omitted for reasons of space but are available from the authors upon request).

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?

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

All three tables provide an unequivocal answer. Average punishment in the control sessions does not differ significantly from the (in, in) case (column 2). Meanwhile there is a substantial and significant difference between punishment in the control sessions and punishment in the (out, out) treatment scenario (column 5).

Next, we 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. To account for the fact that we have multiple observations per participant, 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 the participant's willingness to pay to punish an in-group dictator for an unfair action that affects the participant's own earnings directly.[24] We report both specifications in columns 1-2 of Table 3.[25]

<<Table 3 about here >>

[24]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. This game features only two players: the dictator and the recipient. The dictator chooses between a fair and unfair division, then the recipient him/herself decides whether to punish the dictator. 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.

[25]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.

Result 3: *Behavior is consistent with hypothesis 3b (empathetic bond model) and not consistent with hypothesis 3a (third party inequality aversion model). In our data* $MWP_{(out,in)} > MWP_{(in,in)}$.

In all specifications estimated in Table 3, observers value the marginal punishment opportunity more highly when the dictator is from the out-group and the recipient is from the in-group than when to both dictator and recipient are in-group members. This is true both in terms of valuation for punishment (columns 1-4) as well as for the probability of placing a strictly positive value on the punishment opportunity (columns 5-6). The magnitude of the difference is substantial as well. The 0.14 euro increase in MWP when only the recipient is an in-group member compared to the case where both dictator and recipient are in-group members represents a 42 percent increase in the value of punishment.[26]

Having tested the key prediction separating the empathetic bond model from a model based on distributional preferences, we now investigate the ancillary predictions from our empathetic bond model. Turning again to Table 3, we examine how MWP varies across the remaining scenarios.

Result 4: *Behavior is consistent with hypothesis 4a: 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 4b.*

The estimates in columns 1-4 of Table 3 are consistent with Hypothesis 4a. We indeed find the highest average valuation for the marginal punishment opportunity in the case where the dictator is an out-group member and the recipient is an in-group member. However, on average, we do not find support for Hypothesis 4b. The lowest average value for punishment is associated with the case of an in-group dictator and an in-group recipient. This finding is robust to a tobit regression accounting for censoring (columns 3-4) and to a probit estimation of a positive MWP (columns 5-6); in all the specifications participants on average value punishment the most in the (out, in) treatment. This result is a puzzle that neither the empathetic bond model nor the alternative model we consider explains well.[27]

[26] That is to say, $\frac{0.14}{0.33} = 0.42$.

[27] To see that third party inequality aversion does not explain this pattern, notice that going from the (in-group dictator, out-group recipient) case to the (in-group dictator, in-group recipient) case, the dictator's group is constant while the recipient's group changes. As we have seen, third party inequality aversion predicts the recipient's group should (almost) never directly affect MWP. It could matter by changing the

7 Discussion

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 for instance 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, social identity theory—starting with Tajfel et al. (1971)—suggests there may be an inherent bias toward in-group members which, intuitively, should constrain punishment of in-group members. Predictions in this case depend on how in-group bias is modeled. If the primary effect of a shared group affiliation is to create an empathetic bond among group members (*cf.* Goette et al., 2012), one may want to model in-group bias as an enhanced internalization of the *preferences* of in-group members relative to out-group members. This is our first framework above, in which observers' internalization of others' justice preferences is group-contingent. If, to the contrary, a shared group affiliation primarily affects the weights placed on others' *earnings*, a plausible way to incorporate these effects yields our third party inequality aversion model above. As we have seen, these two approaches differ markedly in their predictions about the relationship between punishment preferences in two cases: i) when both the perpetrator and victim share the observer's group affiliation; vs. ii) when only the victim shares the observer's group affiliation. Our data tend to support the former approach. This is the first direct evidence we know of suggesting that an empathetic bond may be engendered even among minimal-group members.

To answer the more general question of whether a shared social identity enhances or ameliorates the moral disgust component of punishment preferences, our data suggest that both patterns may be at work. First of all, in line with several other studies, pooling over recipients' group affiliations, observers in our experiment generally reported a lower maximum willingness to pay to punish in-group dictators than to punish out-group dictators.[28]

relative $\frac{\beta}{\gamma}$ ratio. Since $\frac{\beta_{in}}{\gamma_{in}} < \frac{\beta_{in}}{\gamma_{out}}$, it could, e.g., move MWP from the case of being strictly positive ($\frac{\beta}{\gamma} > \frac{8}{21}$) to the case where it is always zero ($\frac{\beta}{\gamma} < \frac{8}{22}$). However, this would also imply that imply $MWP_{(out,in)} = 0$, since $MWP_{(out,in)} \leq MWP_{(in,in)} = 0$, which is contrary to our data as well.

[28] Note that Goette et al. (2006) compare punishment behavior towards in- and out-group norm violators

Furthermore, on average, observers' 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 pattern in our data—that punishment is valued by third parties most highly when an out-group member treats an in-group member unfairly—is consistent with 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

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.

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, 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 the moral disgust component of an observer's preference for punishment varies according to the group affiliation of both the transgressor and the victim by minimizing the role for other factors which may obviously affect punishment behavior: responsibility for *undoing* the injustice; and/or a deterrence motive. Specifically, in our setting observers cannot levy anywhere near the amount of punishment necessary to restore equality of others' earnings. Nor can the fixed amount of punishment available substantially alter the transgressor's monetary incentives to select the unfair action. As evidence that we have successfully minimized the deterrent capability of punishment in our experiment, 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.[29]

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. We find that the introduction of group divisions significantly increases the average willingness of bystanders to punish unfair choices.

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. We construct two plausible models of how bystanders' incorporate others' outcomes—internalizing others' preferences or, alternatively, their money earnings—and test a prediction which separates these two models. Our findings suggest identity matters for punishment preferences since introducing artificial group divisions significantly increases the willingness to punish unfair acts.

[29]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).

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. This latter pattern supports an *empathetic bond* framework over a distributional preferences story.

Consistent with the both the literature on in-group bias and group defensive behavior, 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, K., Benedikt, H., 2009. Pointless Vendettas. Manuscript.

[2] Akerlof, G., Kranton, R., 2000. Economics and Identity. The Quarterly Journal of Economics, 115(3), 715-773.

[3] Balafourakis, L., Nikiforakis, N., 2012. Norm Enforcement in the City: A Natural Field Experiment. European Economic Review 56, 1773-1785.

[4] Banfield, E., 1958. The Moral Basis of a Backward Society. Glencoe, IL: The Free Press.

[5] Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial Altruism in Humans. Nature 442(24), 912-915.

[6] Becker, G., DeGroot, M., Marschak, J., 1964. Measuring utility by a single-response sequential method. Behavioral Science 9(3), 226-232.

[7] Bohnet, I., Zeckhauser, R., 2004. Trust, Risk and Betrayal. Journal of Economic Behavior and Organization 55, 467-484.

[8] Bolton, G.E., Ockenfels, A., 2000. ERC: A Theory of Equity, Reciprocity and Competition. American Economic Review 90(1), 166-193.

[9] Bornstein, G., 1992. The Free Rider Problem in Intergroup Conflicts over Step-Level and Continuous Public Goods. Journal of Personality and Social Psychology 62, 597-606.

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

[11] Carpenter, J., Matthews, P.H., 2010. Norm Enforcement: The Role of Third Parties. Journal of Institutional and Theoretical Economics 166, 239-258.

[12] Charness, G., Rabin, M., 2002. Understanding Social Preferences with Simple Tests. The Quarterly Journal of Economics 117(3), 817-869.

[13] Chen, R., Chen, Y., 2011. The Potential of Social Identity for Equilibrium Selection. American Economic Review 101(6), 2562-2589.

[14] Chen, Y., Li, S.X., 2009. Group Identity and Social Preferences. American Economic Review 99(1), 431-457.

[15] Choi, J.-K., Bowles, S., 2007. The Coevolution of Parochial Altruism and War. Science 318, 636-640.

[16] Darwin, C., 1873. The Descent of Man. New York: Appleton.

[17] De Cremer, D., van Vugt, M., 1999. Social Identification Effects in Social Dilemmas: A Transformation of Motives. European Journal of Social Psychology 29, 871- 893.

[18] De Dreu, C.K.W., Greer, L.L., Handgraaf, M.J.J., Shalvi, S., Van Kleef, G.A., Baas, M., Ten Velden, F.S., Van Dijk, E., Feith, S.W.W., 2010. The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans. Science 328, 1408-1411.

[19] Fearon, J.D., Laitin, D.D., 1996. Explaining Interethnic Cooperation. American Political Science Review 90(4), 715-35.

[20] Fehr, E., Schmidt, K., 1999. A Theory of Fairness, Competition, and Cooperation. The Quarterly Journal of Economics 114(3), 817-868.

[21] Fehr, E., Fischbacher, U., 2004. Third-Party Punishment and Social Norms. Evolution and Human Behavior 25, 63-87.

[22] Goette, L., Huffman, D., Meier, S., 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), 212-216.

[23] Goette, L., Huffman, D., Meier, S., 2012. The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups. American Economic Journal: Microeconomics 4(1), 101-115.

[24] Guala, F., Mittone, L., Ploner, M., 2009. Group Membership, Team Preferences, and Expectations. University of Trento, CEEL Working Papers 0906.

[25] Halevy, N., Bornstein, G., Sagiv, L., 2008. In-group love and out-group hate as motives for individual participation in intergroup conflict: A new game paradigm. Psychological Science 19, 405-411.

[26] Harrison, G., Rutström, E.E., 2008. Risk Aversion in the Laboratory. in Cox, J.C., Harrison, G.W. (eds.), Risk Aversion in Experiments. Bingley: Emerald.

[27] Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, JK.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly Punishment Across Human Societies. Science 312, 1767 - 1770.

[28] Lewisch, P., Ottone, S., Ponzano, F., 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.

[29] McLeish, K.N., Oxoby, R., 2007. Identity, Cooperation, and Punishment. IZA Discussion Paper No. 2572.

[30] Mumendey, A., Otten, S., 1998. Positive-negative asymmetry in social discrimination. European Review of Social Psychology 9, 107-143.

[31] Ockenfels, A., Werner, P., 2014. Beliefs and Ingroup Favoritism. Journal of Economic Behavior and Organization. http://dx.doi.org/10.1016/j.jebo.2013.12.003

[32] Plott, C., Zeiler, K., 2005. The Willingness to Pay Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations. American Economic Review 95, 530-545.

[33] Tajfel, H., Flament, C., Billig, M.G., Bundy, R.F., 1971. Social Categorization and Intergroup Behavior. European Journal of Social Psychology 1, 149-177

[34] Tajfel, H., Turner, J., 1986. The Social Identity Theory of Intergroup Behavior. in: Worchel, S., Austin, W. (eds). The Social Psychology of Intergroup Relations. Chicago: Nelson- Hall.

[35] Yamagishi, T., Kiyonari, T., 2000. The Group as the Container of Generalized Reciprocity. Social Psychology Quarterly 63(2), 116-132.

[36] Yamagishi, T., Mifune, N., 2008. Does Shared Group Membership Promote Altruism?. Rationality and Society 20(1), 5-30.

[37] Yamagishi, T., Mifune, N., 2009. Social Exchange and Solidarity: In-Group Love or Out-Group Hate?. Evolution and Human Behavior 30(4), 229-237.

Tables and Figures

Table 1: Descriptive Statistics

Average MWP	All Subjects		Conditional on MWP > 0	
	Control	Treatment	Control	Treatment
$MWP_{(Control)}$	0.31		0.51	
	(0.04)		(0.05)	
$MWP_{(in, in)}$		0.33		0.54
		(0.04)		(0.05)
$MWP_{(out, in)}$		0.47		0.70
		(0.04)		(0.05)
$MWP_{(in, out)}$		0.39		0.62
		(0.04)		(0.05)
$MWP_{(out, out)}$		0.40		0.61
		(0.04)		(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 2a: Treatment Effect on Observer's Punishment Preferences (OLS)

	(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*
	(0.041)	(0.035)	(0.053)	(0.043)	(0.043)
Constant	0.31***	0.31***	0.31***	0.31***	0.31***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Observations	488	194	194	194	194
R-squared	0.01	0.00	0.01	0.04	0.01

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 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 2b: Treatment Effect on Observer's Punishment Preferences (TOBIT)

	(1)	(2)	(3)	(4)	(5)
		Control vs. Treatment Scenario			
	All (Pooled)	(in, in)	(in, out)	(out, in)	(out, out)
Treatment (dummy)	0.26**	0.08	0.23*	0.42***	0.25**
	(0.110)	(0.096)	(0.136)	(0.108)	(0.114)
Constant	0.01	0.01	0.03	0.03	0.02
	(0.045)	(0.036)	(0.032)	(0.031)	(0.030)
Observations	488	194	194	194	194
R-squared	0.005	0.001	0.007	0.023	0.008

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. [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 TOBIT regression accounting for censoring at the extreme values of MWP (i.e. 0 and 1). [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 2c: Treatment Effect on Observer's Punishment Preferences (PROBIT)

	(1)	(2)	(3)	(4)	(5)
Dep. Var:		Control vs. Treatment Scenario			
Pr(MWP)>0	All (Pooled)	(in, in)	(in, out)	(out, in)	(out, out)
Treatment (dummy)	0.11*	0.03	0.11	0.19***	0.10*
	(0.056)	(0.055)	(0.069)	(0.055)	(0.057)
Observations	488	194	194	194	194
R-squared	0.006	0.001	0.009	0.030	0.008

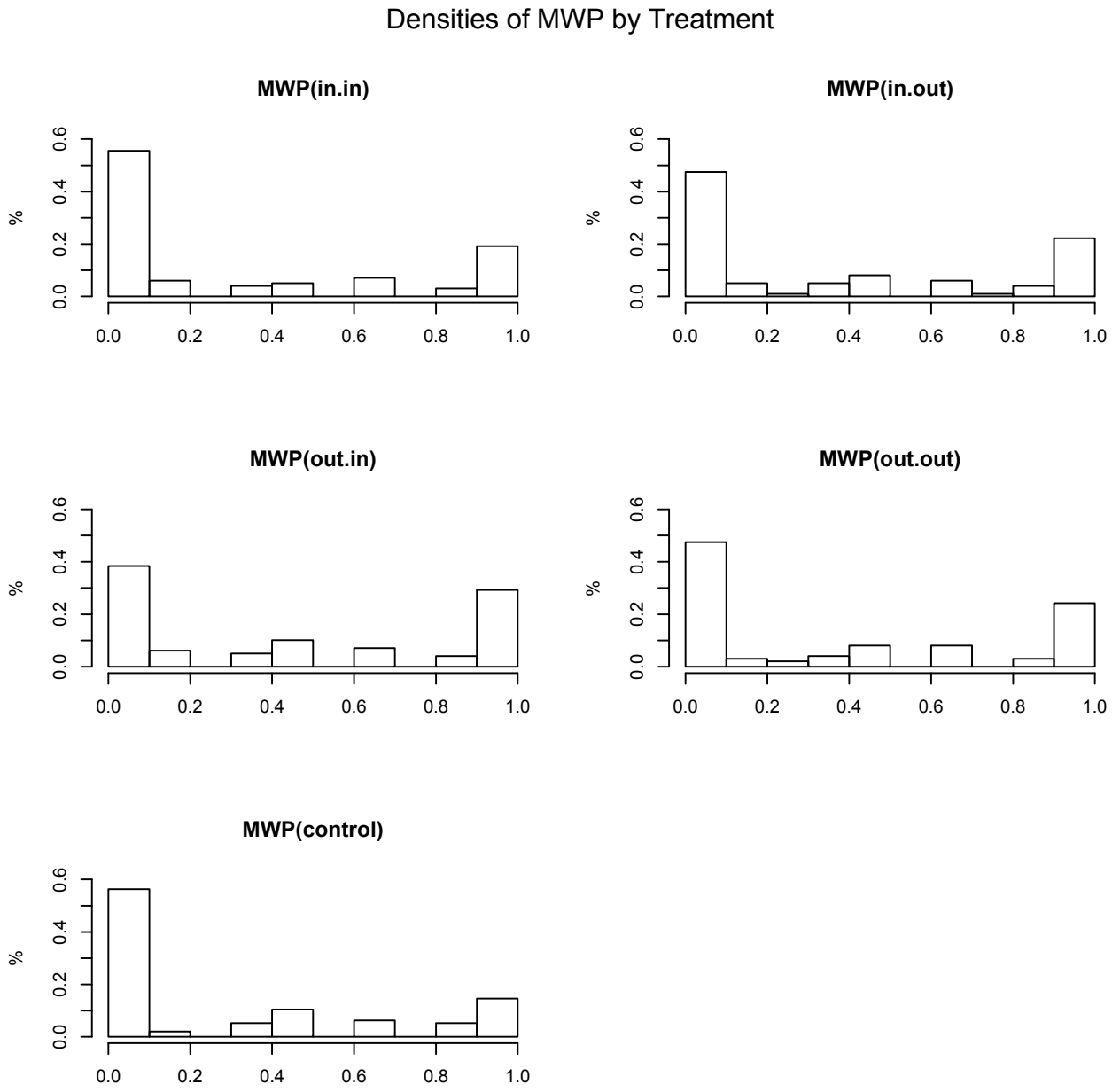
Notes: [1] Column 1 pools data from the control sessions together with observations from all four punishment scenarios in the treatment sessions, resulting in four observations per treatment session participant. [2] Columns 2-5 include only one observation from one punishment scenario for each individual, with the specific scenario listed in the column heading. Accordingly, we report marginal effects after a PROBIT estimation of a positive MWP, i.e. *Pr(MWP)>0*. [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

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Var:	*MWP*				*Pr(MWP>0)*	
Model:	OLS		TOBIT		PROBIT	
in_out	0.06*	0.06*	0.16*	0.15*	0.08**	0.08**
	(0.028)	(0.028)	(0.084)	(0.078)	(0.034)	(0.036)
out_in	0.14***	0.14***	0.35***	0.34***	0.16***	0.17***
	(0.020)	(0.020)	(0.076)	(0.071)	(0.017)	(0.020)
out_out	0.07*	0.07*	0.17*	0.17*	0.07*	0.07**
	(0.035)	(0.035)	(0.090)	(0.085)	(0.036)	(0.037)
MWP for direct punishment of own group		0.17*		0.99***		0.34***
		(0.070)		(0.166)		(0.070)
Observations	392	392	392	392	392	392
R-squared	0.014	0.140	0.007	0.122	0.011	0.085

Notes: [1] The table reports 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 specifications in columns 1 and 2, but the results were similar so we report only the simpler OLS models. [2] In columns 1-2 we report results from a simple OLS estimation of the willingness to pay to punish (MWP); in columns 3-4 we report results from a TOBIT estimation accounting for censoring in the dependent variable (MWP); columns 5-6 finally reports the marginal effects after a PROBIT estimation of the probability of a positive MWP, i.e. *Pr(MWP)>0*. [3] Controls include: a set of dummies for the four possible (dictator group, recipient group) punishment scenarios—the excluded category being (in, in). [4] In specifications 2, 4 and 6 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. [5] Robust standard errors, clustered by session, appear in parentheses.

Figure 1: Distribution of Punishment Preferences by Treatment



Appendix

A Empathetic bond model and example

In our empathetic bond model, we assume: i) all individuals derive weakly positive utility from the perpetrators of unfair acts being punished (justice preference); ii) observers internalize others' preferences; iii) the extent to which observers internalize others' preferences is group-contingent. To simplify notation while capturing the intuition 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.

We use the subscripts d, r and o to denote the dictator, recipient and observer, respectively. Let $\phi_j, j \in \{d, r, o\}$, be the utility an 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 monetary cost the observer must pay to levy one euro of punishment. Let $\alpha \in (0, 1)$ be a parameter capturing how much the observer generally weights others' preferences when making decisions ("other-regardingness") and $\beta \in (0, 1)$ be a parameter capturing how much the observer weights the dictator relative to the recipient.[30] 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:

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

[30] 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.

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.[31]

Given these assumptions, consider $MWP_{(out,in)}$ compared to $MWP_{(in,in)}$. The latter case is identical to the former case except that observer puts positive weight on the dictator's preferences in the latter case but not the former case. By assumption, the dictator prefers no punishment. Consequently, it must be the case $MWP_{(out,in)} < MWP_{(in,in)}$.

Similarly, consider $MWP_{(out,in)}$ vs. $MWP_{(out,out)}$. The latter case is identical to the former case except the observer puts positive weight on the recipient's preferences in the former case but not the latter. Since the recipient, by assumption, prefers punishment we would expect $MWP_{(out,in)} > MWP_{(out,out)}$.

Finally, consider $MWP_{(out,in)}$ vs. $MWP_{(in,out)}$. These two cases differ in two respects. First of all, the observer places weight on the dictator's preferences in the latter case but not the former, lowering the observer's MWP in the latter case. Secondly, the observer places positive weight on recipient's preferences in the former case but not the latter, again lowering the observer's MWP in the latter case relative to the former. Since both of these effects go in the same direction, we would expect $MWP_{(out,in)} > MWP_{(in,out)}$.

Putting all of these arguments together, we would expect the observer to place the largest value on the opportunity to punish in the case dictator where the dictator is an out-group member and the recipient is an in-group member:

$$max\{MWP_{(out,in)}, MWP_{(in,out)}, MWP_{(in,out)}, MWP_{(in,in)}\} = MWP_{(out,in)}.$$

A.1 An illustrative example of the empathetic bond model

As a motivating example, consider the case where the dictator and recipient care only about their own money earnings and justice: $u_j = \pi_j + \phi_j, j \in \{d, r\}$. Further, suppose the observer cares about his own money utility, justice and the others' utility. In this special case, the observer's utility conditional on punishing is:

[31] 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.

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

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

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

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 $U_{o|\text{punish}} = U_{o|\text{not punish}}$. The observer's MWP in the four cases in this simple example are given by:

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

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 $MWP_{(in,in)}$ can be written as $MWP_{(in,out)}$ 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 all of 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 $MWP_{(out,out)} \leq MWP_{(in,in)}$.

B Third party inequality aversion model and example

The second model we construct allows us to address the question of whether our empathetic bond model offers any insights or testable predictions beyond those provided by more standard distributional preferences models like inequality aversion (Fehr and Schmidt 1999; Bolton and Ockenfels, 1999) or social welfare preferences (Charness and Rabin, 2002) when suitably modified to incorporate group or social identity (Chen and Li, 2009). There are

multiple ways to construct a distributional preferences model. We take a straightforward and transparent route and construct a model of "third party inequality aversion." Generally, we assume the observer's utility is a weighted average of two components: i) his or her own money earnings; and ii) the "effective inequality" embodied by the distribution of others' money earnings.

We choose inequality aversion as our base model for two reasons. First, because the observer can only ever destroy surplus, inequality aversion will yield predictions similar to those of other popular models of distributional social preferences such as Social Welfare Preferences (Charness and Rabin, 2002). Secondly, because the observer can only lower the dictator's earnings and cannot affect the recipient's earnings, a distributional preferences model based on the difference in earnings between the dictator and recipient—such as inequality aversion—has the best chance of providing concrete predictions. For example, suppose observers care only about the earnings of others without taking into account inequality. Then since "punishing" only lowers the observer's earnings without affecting the recipient's earnings, we would expect the recipient's group affiliation to have no effect on "punishment." Anticipating that we will find such variation implies using a model like inequality aversion as the base model.

To incorporate group-contingent preferences, we construct a measure of *effective inequality* which depends on the dictator's and recipient's group affiliations. We compute effective inequality by multiplying the dictator's and recipient's money earnings by group-contingent weights and then taking the absolute value of the difference in these weighted earnings. Let $0 \leq \beta_{G_d} \leq 1$ be the weight placed on the dictator's money earnings and $0 \leq \gamma_{G_r} \leq 1$ be the weight associated with the recipient's earnings, where $G_j \in \{\text{in, out}\}$ indicates the group of agent $j \in \{d, r\}$ relative to the observer's group. To be consistent with previous research we assume that the weights assigned to in-group members' earnings are larger than the weights assigned to out-group members' earnings: $\beta_{\text{in}} \geq \beta_{\text{out}}$ and $\gamma_{\text{in}} \geq \gamma_{\text{out}}$. For ease of exposition, for the moment let us suppress the dependence of these weights on agents' groups and write:[32]

[32]Note that we omit the observer's money earnings from our measure of effective inequality. We do so because, intuitively, when comparing her own earnings to those of the dictator and the recipient, respectively, the observer implicitly also compares the other two's earnings to each other. Including the observer's earnings would thus only complicate the expression we derive for MWP below without providing any additional insight relevant for our analysis, especially as we restrict ourselves to the case of unfair dictator choices. Here, the dictator will always have higher monetary earnings than other two, and the observer will always have higher earnings than the recipient.

$$\text{Effective inequality} = |\beta\Pi_d - \gamma\Pi_r|$$

Let $0 \leq \alpha \leq 1$ be the weight the observer places on effective inequality, so that $0 \leq (1-\alpha) \leq 1$ is the weight placed on the observer's own money earnings. In this third party inequality aversion model, an observer's utility can be written:

$$U_o = (1-\alpha)\Pi_o - \alpha|\beta\Pi_d - \gamma\Pi_r|$$

Like our empathetic bond model, this third party inequality aversion model is simple and flexible. Additionally, it provides strong predictions about observers' behavior. Denote by $c(p)$ the price of punishment. We can write the observer's utility from punishing as:

$$U_{o|\text{punish}} = (1-\alpha)[\Pi_o - c(p)] - \alpha|\beta(\Pi_d - 1) - \gamma\Pi_r|$$

The observer's utility from not punishing is simply:

$$U_{o|\text{no punish}} = (1-\alpha)\Pi_o - \alpha|\beta\Pi_d - \gamma\Pi_r|$$

For predictions, it will prove useful to divide the parameter space into three intervals:

$$\frac{\beta}{\gamma} \in \begin{cases} [0, \frac{\Pi_r}{\Pi_d}], \\ (\frac{\Pi_r}{\Pi_d}, \frac{\Pi_r}{\Pi_d - 1}), \\ [\frac{\Pi_r}{\Pi_d - 1}, \infty] \end{cases}$$

Consider the first the case, where $\frac{\beta}{\gamma} \leq \frac{\Pi_r}{\Pi_d} = \frac{8}{22}$. Intuitively, in this case the observer cares relatively little about the dictator's earnings, valuing a marginal unit of the dictator's earnings no more than about 36% as much as a marginal unit of the recipient's earnings. In this case effective inequality already favors the recipient so that paying a positive amount to reduce the dictator's earnings *increases* effective inequality and decreases the observer's own money earnings. Consequently, here the observer's $MWP = 0$.

Consider the third case next, where $\frac{\beta}{\gamma} \geq \frac{\Pi_r}{\Pi_d - 1} = \frac{8}{21}$. Here, the observer cares at least 38% as much about a marginal unit of the dictator's earnings as he or she cares about a marginal unit of the recipient's earnings. After a bit of algebraic manipulation, the observer's MWP in this case can be written as:[33]

[33]The observer prefers punishing to not punishing whenever $U_{o|\text{punish}} = (1-\alpha)(\Pi_o - c(p)) - \alpha\beta\Pi_d + \alpha\beta + \alpha\gamma\Pi_r \geq (1-\alpha)\Pi_o - \alpha\beta\Pi_d + \alpha\gamma\Pi_r = U_{o|\text{no punish}}$. This condition simplifies to $c(p) \leq \frac{\alpha}{(1-\alpha)}\beta$.

$$MWP = \frac{\alpha}{1-\alpha}\beta$$

Importantly, notice that MWP is positively related with β. Since β is larger for in-group dictators than for out-group dictators by assumption, in this area of the parameter space observers will place a *higher* value on punishing *in-group* dictators than out-group dictators. Intuitively, here punishment always reduces effective inequality by a fixed amount proportional to β. Observers are willing to pay a higher price for a larger reduction in effective inequality. This phenomenon would hold more generally and would still obtain if, e.g., the observer incorporated his or her own money earnings into the definition of effective inequality.

The remaining case is when $\frac{8}{22} < \frac{\beta}{\gamma} < \frac{8}{21}$. What happens in this case is less clear and more dependent on functional form assumptions. The subtlety arises because levying punishment will change the sign of effective inequality from benefitting the dictator towards benefitting the recipient so that how much the observer is willing to pay depends on the precise tradeoff between these two sides of inequality. Because we consider this case to be *a priori* unlikely to drive our experimental results generally — it constitutes a small portion of the parameter space with no obviously focal qualities — we do not examine this case any further. We only note that the observer's MWP on this interval should fall in between the MWPs in the previous two cases: MWP will always be weakly above 0 which is the MWP when $\frac{\beta}{\gamma} \leq \frac{8}{22}$; on the other hand, the amount of effective inequality reduction here is unambiguously (weakly) lower than when $\frac{\beta}{\gamma} > \frac{8}{21}$, resulting in a weakly lower MWP.

We are now in a position to generate a testable prediction from this model. Begin by fixing α. For the starkest contrast between this model and the previous framework, consider $MWP_{(out,in)}$ vs. $MWP_{(in,in)}$. Moving from the former to the latter increases the relevant β, because $\beta_{in} > \beta_{out}$ by assumption, while leaving the relevant $\gamma = \gamma_{in}$ unchanged. Consequently, as long as it is not the case that both of the ratios $\frac{\beta_{in}}{\gamma_{in}}$ and $\frac{\beta_{out}}{\gamma_{in}}$ lie within the small interval $(\frac{8}{22}, \frac{8}{21})$, we can confidently predict:

$$MWP_{(out,in)} \leq MWP_{(in,in)}$$

Too see this, note that by assumption $\frac{\beta_{in}}{\gamma_{in}}$ and $\frac{\beta_{out}}{\gamma_{in}}$ do not *both* lie within the interval $(\frac{8}{22}, \frac{8}{21})$. If both $\frac{\beta_{in}}{\gamma_{in}}$ and $\frac{\beta_{out}}{\gamma_{in}}$ are in the interval $(-\infty, \frac{8}{22}]$, then $MWP_{(out,in)} = MWP_{(in,in)} = 0$, as we have seen, so the prediction obtains. If $\frac{\beta_{in}}{\gamma_{in}} \leq \frac{8}{22}$ and $MWP_{(in,in)} >$

$\frac{8}{21}$, then $MWP_{(out,in)} = 0 \leq \frac{\alpha}{1-\alpha}\beta_{\text{in}} = MWP_{(in,in)}$ and the prediction holds. Finally, supppose both $MWP_{(out,in)} \geq \frac{8}{21}$ and $MWP_{(in,in)} \geq \frac{8}{21}$. Then $MWP_{(out,in)} = \frac{\alpha}{1-\alpha}\beta_{\text{out}} \leq \frac{\alpha}{1-\alpha}\beta_{\text{in}} = MWP_{(in,in)}$ and the prediction again obtains.

B.1 An illustrative example of the third party inequality aversion model

At this point, it may prove helpful to compute a numerical example. So, let $\alpha = \frac{1}{2}$ so that the observer cares as much about his or her own money earnings as about effective inequality. Suppose that the observer does not care about roles, *per se* but only about the in-group/out-group distinction and that the observer cares twice as much about in-group members' earnings: $\beta_{in} = \gamma_{in} = 1$; $\beta_{out} = \gamma_{out} = \frac{1}{2}$.

Consider first the case where both dictator and recipient are in-group members. The observer's utility from punishing or not punishing in this case is:

$$\begin{array}{lcl} U_{o|\text{punish}} & = & \frac{1}{2} \times [15 - c(p)] - \frac{1}{2} \times |21 - 8| \\ U_{o|\text{no punish}} & = & \frac{1}{2} \times 15 - \frac{1}{2} \times |22 - 8| \end{array}$$

Solving for the largest $c(p)$ which still leaves $U_{o|\text{punish}} \geq U_{o|\text{no punish}}$ implies that the observer's $MWP_{(in,in)} = 1$. The observer would pay the largest feasible amount of 1 euro in our experiment to levy one euro of punishment.

Next, consider the case where only the recipient is an in-group member. The expressions for the observer's utility with and without punishment are:

$$\begin{array}{lcl} U_{o|\text{punish}} & = & \frac{1}{2} \times [15 - c(p)] - \frac{1}{2} \times |\frac{1}{2} \times 21 - 8| \\ U_{o|\text{no punish}} & = & \frac{1}{2} \times 15 - \frac{1}{2} \times |\frac{1}{2} \times 22 - 8| \end{array}$$

Solving again for the largest $c(p)$ satisfying $U_{o|\text{punish}} \geq U_{o|\text{no punish}}$ yields $MWP_{(out,in)} = \frac{1}{2}$. The observer would be willing to spend only half as much to punish an out-group dictator for treating an in-group recipient unfairly compared to the case where both dictator and recipient are in-group members. This illustrates our primary prediction from this model: $MWP_{(out,in)} \leq MWP_{(in,in)}$.