Group and Similarity Effects on Cooperation and Punishment

Carleen Liu

Advised by: Dr. Yarrow Dunham

Yale University

Correspondence to:

Carleen Liu

Yale University

Email: carleen.liu@yale.edu

Abstract

Social categories, or groups, play a large role in decision making. Ingroup bias strongly influences cooperative decision-making as people tend to share more with ingroup members and differentially enforce fairness norms depending on the group membership of their interaction partners (McAuliffe and Dunham, 2016). We tend to like ingroup members, and also expect more from ingroup members. These two features of ingroup bias in cooperation tend to complement each other until punishment comes into the picture. Are we more likely to punish ingroup members or outgroup members for the same offenses? Are we more likely to punish similar others or dissimilar others for the same offenses? How about when these are crossed? We find that people give more to, expect more from, and punish more for norm violations from ingroup members than outgroup members, which is consistent with maintaining and adhering to group norms. While no evidence was found for ingroup positivity, we found outgroup negativity in punishing norm violations.

Introduction

Human social life depends largely on groups, and people demonstrate strong tendencies to favor the group they are in. Ingroup biases emerge early in development, and even appear for new, arbitrarily assigned social identities (also known as minimal groups), which suggests that basic cognitive abilities needed to identify with social groups are present at a young age and ingroup biases do not require protracted cultural input (McAuliffe and Dunham, 2016). Darwin expressed that group life arises from solving a coordination problem – the question of who to cooperate with (McAuliffe and Dunham, 2016). Human social behavior is uniquely characterized by cooperation, or when groups act together for mutual benefit.

There are two evolutionary biological explanations for the costly cooperative behavior of both humans and animals – kin selection and reciprocal altruism. Kin selection explains cooperation between genetically related actors even when the behavior is costly (Shinada et al., 2004). Reciprocal altruism involves behavior that temporarily reduces one's own fitness while increasing another's fitness, with the expectation that the other will act similarly later on (Trivers, 1971). This helps explain mutual cooperation in repeated dyadic interactions among genetically unrelated actors (Shinada et al., 2004).

Unlike other species, however, humans have developed complex systems of cooperation that extend beyond kin and beyond repeated dyadic interactions with non-kin. Vast literature on human cooperation has shown that people cooperate more with ingroup members, even if they are not kin and even if it is unlikely they will have repeated interactions with them in the future.

Punishment in cooperative contexts benefits the punisher because it gives the punisher a positive reputation and benefits the future cooperation partners of the norm violator, as punishment may alter their behavior. However, there also exists instances of altruistic punishment, or costly punishment that does not benefit other players through reputation effects or through the rehabilitation of a norm violator (Shinada 2004). In this paper, we will examine altruistic punishment and forgiveness directed toward ingroup and outgroup norm violators.

Norms Based Hypothesis and Mere Preferences Hypothesis

McAuliffe and Dunham propose that there are two hypotheses that explain why there are group biases in cooperation. The Norms-Focused Hypothesis claims that enforcement of cooperative norm violations is an evolved mechanism that supports withingroup cooperation.

The Mere Preferences Hypothesis argues group bias in cooperation is a by-product of more general affective preferences for ingroup members.

The Norms Focused Hypothesis argues that an evolutionary mechanism supports withingroup cooperation. Humans are much more flexible in group life than primates, whose social distinctions are limited to biological sex, kinship, and social group (McAuliffe and Dunham, 2016). As reciprocal altruism explains, group norms allow regulated and predictable behavior with partners that one is likely to have repeated interaction with. Aligning personal actions to norms and enforcing compliance to norms in others would make cooperation profitable (McAuliffe and Dunham, 2016).

In contrast, the Mere Preferences Hypothesis is rooted in theories of self-favoritism. This framework is based on social categories' distinct relationship with the self – the idea that "the groups I belong to are closely linked to me, aspects of who I am and where I reside in the social order" (McAuliffe and Dunham, 2016). This has many implications on cognitive and affective processes behind ingroup favoritism, as the ingroup can be a means of self-enhancement and self-defense (Tajfel and Turner, 2004). It is also possible that positive attitudes toward self can spread to the social groups one is in (Greenwald et al., 2002). In the context of cooperation, mere preference of ingroup members can affect behavior toward ingroup members.

There is much evidence that is consistent with the Norms Focused Hypothesis, such as the adherence to ingroup norms and the tendency to choose ingroup members as partners to cooperate with. But the evidence could also support the Mere Preferences Hypothesis – an individual could simply prefer ingroup members and preferences affect behavior. Being more generous to an ingroup member, for example, could support both of these hypotheses.

Though these mechanisms need not be mutually exclusive, determining which mechanisms are at work and how they work together would clarify the nature of group effects on cooperation and punishment.

Mixed Literature regarding Norm Violation

Norm Based Hypothesis Prediction on Norm Violation

A class of cases that begin to parse out these two hypotheses is norm violations. The Norms Based Hypothesis implies ingroup members are subject to norms that outgroup members are not, and that people will be more inclined to enforce norms among ingroup members because they will reap the benefits from maintaining within group cooperation. This means that ingroup member violations of norms are more likely to be punished because of the obligation ingroup members have to one another for the sake of enforcing future cooperation. If the same norm were violated by an outgroup member, the Norm Based Hypothesis predicts relatively lower punishment as the outgroup member has no special obligation to the group member.

Mere Preferences Hypothesis Prediction on Norm Violation

In contrast, the Mere Preferences Hypothesis predicts the opposite – an ingroup member is more positively viewed and therefore, more readily forgiven in cases of norm violations. Preference for group members can be separable from our predictions on how people will behave. When participants were given negative information about the prior behavior of group members, Baron and Dunham (2015) found that intergroup preferences were not as affected as inductive generalizations about behavior. That is, ingroup positivity was decreased but not reversed. This could serve to protect the self from negatively evaluating the ingroup, while maintaining the

ability to make accurate predictions about behavior. Therefore, ingroup preference does not prevent people from recognizing ingroup members' norm violations. Still, positive attitudes toward the ingroup do not become negative when learning about ingroup members' non-cooperative behavior. This is consistent with the Mere Preferences Hypothesis, which suggests punishment for ingroup norm violation will be, at least to some degree, offset by positive attitudes toward the ingroup.

The Black Sheep Effect

One specific case of norm violation is the Black Sheep Effect. This occurs when an ingroup member violates a central group norm and becomes disliked more than an outgroup member who performs the same action (Marques and Paez, 1994). It seems the Black Sheep Effect supports the Norms-Focused Hypothesis, as this effect is an outcome of normative pressure. Because there is strong internal differentiation that leads to rejection of socially undesirable ingroup members, it is inconsistent with the Mere Preferences Hypothesis (Marques and Paez, 1994). The Mere Preference Hypothesis holds that ingroup favoritism is based in the idea that maintaining a positive social identity must involve promoting a positive differentiation between the ingroup and outgroup as wholes. Past research on the Black Sheep Effect has focused on ingroup-specific standards though, so it is unclear whether this would also occur in broader cooperative contexts (McAuliffe and Dunham, 2016).

There are also many experiments examining norm violation that are inconsistent with the Black Sheep Effect. This literature has mixed results, as some support the Norms Focused Hypothesis and others support the Mere Preferences Hypothesis. Studies testing group biases in

responses to norm violation often involve second-party or third-party costly punishment economic games, which we will detail in the following sections.

The Ultimatum Game

A commonly used second-party game is the Ultimatum Game. It is a two player game in which the first player proposes how to divide a sum between herself and the second player. The second player then decides whether to accept or reject the offer. If the offer is rejected, the entire sum is lost and neither player gets anything. From a Norms Focused point of view, a low offer by the first player can be seen as a violation of ingroup norm fairness and can be taken more negatively than if an outgroup member were to make the same offer. From a Mere Preferences perspective, the same offer can be viewed more positively if it comes from an ingroup member because it's made by a better liked or preferred partner.

However, using the Ultimatum Game has resulted in many conflicting findings. One study conducted an Ultimatum Game in which participants were paired with racial ingroup and outgroup members and found that participants were more likely to reject marginally unfair offers (80/20 split) from ingroups in comparison to outgroup members, implying that adults are more likely to enforce fairness norm violations within group (Mendoza, 2014). These findings support the Norms Focused Hypothesis.

But another study found that university students were more tolerant of unfair offers from ingroup members. In this Ultimatum Game, participants were more likely to accept a marginally unfair offer ($7.50 out of $20.00) if it came from an ingroup proposer (Valenzuela and Srivastava, 2012). McAuliffe and Dunham (2016) offer a potential reason as to why the Ultimatum Game has yielded inconsistent results – the game creates a tension between favoring

the ingroup and coming to a split that the other party will accept. Therefore, they suggest group bias in cooperation should be studied using third-party punishment games, in which the participant is an observer who can choose to punish players for norm violations, but these violations do not directly affect the observer so the motivation to maximize individual payoff is eliminated.

The Third-Party Punishment Game

A third-party punishment game from Shinada et al. (2004) found those who previously donated in a gift-giving game ("cooperators") were more likely to punish ingroup members than outgroup members who did not donate ("non-cooperators"). This is consistent with the Norms Focused Hypothesis' prediction that norm violations committed by ingroup members are punished more harshly than those commited by outgroup members. Shinada et al. argue that punishment to induce cooperation is more likely to be directed toward ingroup members than outgroup members in the same way cooperative behavior is more likely to be directed toward ingroup members than outgroup members. They extend the group-based nature of cooperation to punishment behavior and raise the issue of second-order cooperation – when should you pay a cost to punish non-cooperators? Punishment benefits all members of a group, but can be costly to the punisher. Costly punishment is worth it when it's directed to an ingroup member, because the punishment will force free riders to change their behavior and cooperate. The benefits of this cooperation is shared among the group, including the punisher.

Another study using a third-party punishment game found opposite punishment patterns – participants punished ingroup members less and also punished to protect ingroup members (Bernhard et al., 2006). There were four conditions: players A (dictator), B (recipient), and C

(observer) all from the same group; only A and B from the same group; only A and C from the same group; only B and C from the same group. If the Norms Focused Hypothesis is correct, then outgroup members do not need to obey the norm nor do they benefit from the enforcement of the norm. However, punishment was found in all four conditions, suggesting egalitarian sharing norms. Punishment was much higher in the ABC and BC conditions. This suggests ingroup victims are better protected by third-party observers. Contrary to the Norms-Focused Hypothesis, the study found that participants were more likely to punish outgroup members for violating a norm, and less likely to punish ingroup members. Norm violations actually occurred more often if the punisher and norm violator belong to the same group because of this leniency. This favoring and forgiveness of ingroup members is consistent with the predictions the Mere Preferences Hypothesis makes.

The Main Question

The main question, then, is whether we are more likely to punish ingroup members or outgroup members for violating cooperation norms. Again, the Norms Focused Hypothesis and the Mere Preferences Hypothesis are not mutually exclusive or incompatible. It could be that preferences are a proximate psychological mechanism that help realize aspects of the Norms Focused Hypothesis (McAuliffe and Dunham, 2016). It is still useful to find out whether ingroup biases stem directly from intergroup cognition that positions ingroup members as reliable cooperation partners or just spill over from general ingroup positivity without expectations of shared norms (McAuliffe and Dunham, 2016). It may be that norms or preferences are the driving factor, or that both interact to bring about cooperation.

Mere Similarity and Group Membership

In the midst of mixed literature, Mussweiler and Ockenfels' 2013 study design is distinct because they distinguish mere similarity from group membership. They found that 1) priming similarity versus group identity can differentially affect people's punishment behavior, 2) participants showed less tolerance and more punishment of unfairness generated by those they perceived to be similar to themselves and 3) by contrast, people were more tolerant of uncooperative behavior from ingroup members. This novel comparison between the effects of mere similarity and the effects of group membership on cooperation norms and punishment has the potential to begin addressing the tension between the Mere Preferences Hypothesis and the Norms Focused Hypothesis.

Shared group membership activates a host of motivational and cognitive mechanisms that may operate together in complex ways. For example, people allocate more resources to ingroup members (Tajfel et al., 1971) and have stronger emotional reactions to their distress (Cikara et al., 2011). But they also expect more from ingroup members and often react in anger when an ingroup member intentionally violates a shared norm (Betancourt and Blair, 1992). The conflicting findings so far may be due to the nature of the group manipulation – certain contexts may cue similarity over other aspects of group membership (e.g. future interaction), and varying results may come from the differential impact of similarity and group identity (McAuliffe and Dunham, 2016).

Mussweiler and Ockenfels' finding of more punishment toward similar others and more tolerance toward ingroup members suggests that similarity may be driving norm enforcement. In the process of social projection, people more readily project their own perspectives onto others they perceive to be similar to themselves (Ames, 2004). Thus, they are more likely to assume

shared normative expectations with similar than dissimilar others. Whereas, more tolerance toward ingroup members can be explained as a direct effect of group preference.

The Present Study

No one has yet fully crossed similarity and group membership to identify the relative weight of these two factors, and the present study attempts to address this gap in the literature. The current study is modeled after Mussweiler and Ockenfels (2013). The present study seeks to replicate those findings and also cross mere similarity and group membership to determine how these two factors interact by examining not just how people punish or forgive norm violations from ingroup members versus outgroup members and from similar versus dissimilar others, but also how people punish or forgive norm violations from similar ingroup members, dissimilar ingroup members, similar outgroup members and dissimilar outgroup members.

Overview of Study Design

At the beginning of the study, participants complete a priming task to induce perception of similarity (or dissimilarity) with the other player. Previous studies found that focusing on similarities versus differences in comparing pictures induces a generalized focus on either similarities or differences that carries over to subsequent tasks, including influencing perceived self-other similarity in interactions that follow. In a separate study by Mussweiler (2001), participants who focused on similarities (or differences) in a picture comparison task subsequently judged themselves as more similar to (or different from) a given other. The manipulation in our experiment is designed to induce viewing their partner in the economic game as similar or different. Mussweiler claims this procedure allows a manipulation of self-other

similarity that is independent of group identity, and does not involve influences from information exchange and reduced social distance.

All participants completed the priming task, regardless of whether they are in the no-group or group condition. This is so we can examine how the similarity prime alone (no-group condition) affects subsequent giving and punishment in the following economic game by inducing a similar or dissimilar other, and to examine how the similarity prime interacts with group bias (group condition). Therefore, in the group condition, the other player would fall under one of the following categories: similar/ingroup, similar/outgroup, dissimilar/ingroup, dissimilar/outgroup. For those in the group condition, group identity was manipulated by affiliation with a political party (Democrat or Republican).

The behavioral economics game is designed in two phases. Phase 1 examines how much people give to a partner without knowing how much they would receive in return. Phase 2 measures punishment decisions depending on the other player's cooperation levels in Phase 1. If people punished based on maintaining group norms, ingroup members would be punished more. If people punished based on preferences for ingroup members rather than maintaining norms, they would be more lenient toward ingroup members and punish them less. Likewise, if people held normative expectations for similar others due to social projection, similar others would be punished more. If people simply prefer similar others and have general positive attitudes toward them, similar others would be punished less.

Predictions

Based off of Mussweiler and Ockenfels' findings of greater punishment when similarity (independent of group membership) was evoked and greater leniency toward ingroup members,

we predict that punishment will be highest in the similar/outgroup condition and lowest in the dissimilar/ingroup condition. Based on their findings, it is unclear how punishment decisions might be made toward similar/ingroup members and dissimilar/outgroup members. Punishment patterns across all four conditions will help demonstrate the relative strength and interaction among mere similarity and group membership.

Method

Priming

We recruited 523 participants on Amazon's mechanical Turk. Before they played the economic game, all participants completed a priming task. We replicated Mussweiler and Ockenfel's priming task in order to manipulate perceived mere similarity without evoking group membership. Participants completed a task in which they compared two images. Half the participants were asked to list all the similarities they could find between the two pictures, and the other half were asked to list all the differences they could find.

The prime task instructions were as follows: "Part 1. Please have a close look at the two pictures below. Try and determine in what way the two pictures resemble each other and write down as many similarities as possible on the lines provided. In doing so it is important that you compare the pictures as accurately as possible and that you name as many similarities as possible (at least 5). Please take a few minutes for this comparison. What similarities between the two pictures were you able to find?" If participants were in the differences condition, the instructions are the same except it asks them to seek differences.

The following images from Mussweiler and Ockenfels (2013) were used for the prime.

After participants completed the priming task, those randomly assigned to the no-group condition continued straight to the economic game. Those who were randomly assigned to the group condition were asked if they identified as a Democrat or Republican, then randomly matched with participants from their same political party or different political party.

Game Instructions

After similarity priming (and reporting political affiliation for those in the group condition), participants played a two-phase economic game. The game involves real monetary stakes as the payoffs the participants earn by playing the game are real and will be rewarded to them as a bonus. All participants receive a baseline payment for completing the survey. The participants received the following instructions:

"This game has two players: Player 1 vs. Player 2. The game also has two distinct phases.

In Phase 1: Sending:

- Both players start with 20 cents each.

- Both players then decide how many cents, if any, to send to the other player.
- Any money sent from one player to the other will be doubled: for every one cent you send, the other player will receive 2 cents. For every one cent the other player sends, you will receive 2 cents.

For example:

- If both players send all 20 cents, both players will keep nothing and will receive 40 cents each (20 cents times two) and so both will have more than they started with.
- If Player 1 sends 20 cents and Player 2 sends nothing, then Player 2 will keep 20 cents and receive 40 cents (for a total payoff of 60 cents) and Player 1 will receive nothing.

After Phase 1 is Phase 2: Punishment:

- Both players receive an additional 10 cents. Each player can either keep that money or use any amount of it to reduce the payoff of the other player depending on the amount they sent you in Phase 1.
- Thus, in Phase 2 both players decide how much, if at all, to reduce the other player's payoff.
- Any money spent on punishment will take twice that amount away from the other player: in other words, if you spend one cent on punishment, two cents will be deducted from the other player’s payoff.

For example:

- If Player 1 decides not to reduce Player 2's payoff at all, Player 1 will keep all 10 cents.
- If Player 1 decides to reduce Player 2's payoff by 20 cents, Player 1 will spend all 10 cents.”

Comprehension Check

After receiving instructions on the economic game, there were two comprehension check questions that asked for the highest payoff strategy in Phase I and Phase 2.

1. Imagine that the players are in the first phase of the game: sending. Which set of circumstances will result in Player 1 earning the highest payoff?
   A) Player 1 sends everything and Player 2 sends everything
   B) Player 1 sends everything and Player 2 sends nothing
   C) Player 1 sends nothing and Player 2 sends everything (correct answer)
2. Imagine that the players are in the 2nd phase of the game: punishment. Which decision will result in Player
   1 having the highest payoff?
   A) Deciding not to punish Player 2 at all (correct answer)
   B) Deciding to punish Player 2 as much as he or she can
   C) Deciding to punish Player 2 a little bit

All participants must answer the questions correctly before starting the game.

Conceptual Background of the Game

Our second-party game avoids the pitfalls that McAuliffe and Dunham point out about the Ultimatum Game. This study is set up such that personal material interest is at odds with sending any money to the other player (a prosocial cooperative action) and at odds with punishing at all (a costly altruistic action). All participants know sending no money and not punishing at all is the most profit-maximizing strategy before they begin the game. Whereas in the Ultimatum Game, it would be more profitable to gear towards a more equitable split because

if the other player rejects your offer, you get nothing. This is why the Ultimatum Game renders mixed results regarding group bias.

Our study was a replication and adaptation of the Mussweiler and Ockenfels study, so each participant played a one-shot game with a nonkin, randomly determined, anonymous other player on Amazon's Mechanical Turk. The interaction involved real financial stakes and excluded any potential direct benefits of punishing, such as reputation gains or higher future payoffs. Thus, the amount of money participants spent to reduce the payoff of the other player is a valid measure of altruistic punishment.

Results

Expectation of reciprocity was found across the board, not just for ingroup members.

What a participant sends and what a participant expects are strongly positively correlated in both the no-group condition ($r = 0.63$, $p < 0.001$) and the group condition ($r = 0.64$, $p < 0.001$). This expectation of reciprocity is present across conditions. In other words, if you send less, you expect less and if you send more, you expect more.

Punishment was significantly though weakly correlated with the amount sent back in both the group condition ($r = -0.17$, $p < 0.001$) and in the no-group condition ($r= -0.13$, $p < 0.001$).

No-group, Prime Only Condition (Similar/Dissimilar)

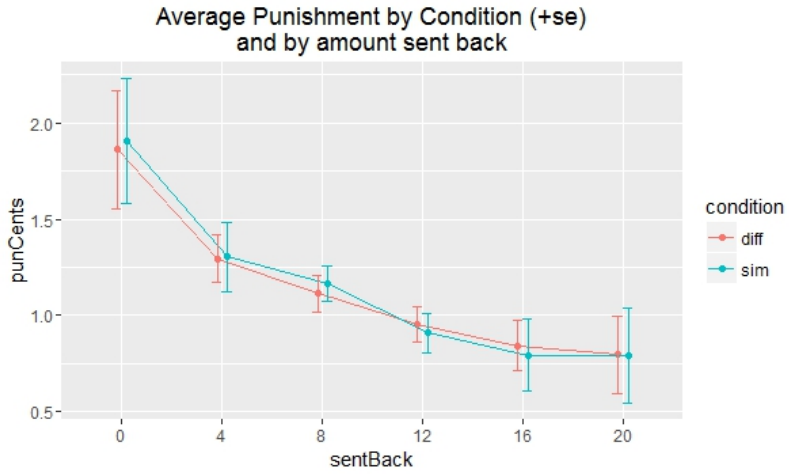
Participants did not send more or expect more when primed with similarity.

Using an independent samples t-test, we compared cents sent by no-group participants in the similarity prime ($M = 11.41$) and in the differences prime ($M = 11.16$) and found no significant differences in the amount sent, $t(170) = 0.20$, $p = 0.84$. An independent samples t-test

was also conducted on the expected amount from no-group participants in the similarity prime (M = 10.56) and in the differences prime (M = 11.82), and there were no significant differences in expectation of the amount received, t(166) = 1.01, p = 0.32. Therefore the prime did not seem to evoke significant effects in sending cents or expectations of cents.

Participants did not punish more when primed with similarity.

Using an independent samples t-test, we compared punishment decisions by no-group participants in the similarity prime (M = 1.21, amount spent to punish) and in the differences prime (M = 1.08, amount spent to punish) and found no significant differences in the punishment decisions, t(1020) = -0.86, p = 0.39. There is also no difference in expectation of punishment for no-group participants in the similarity prime (M=2.93) or the differences prime (M=2.22), t(165)=-0.89, p = 0.37.



Participants' explicit attitudes about how similar they are to the other player were unaffected by the prime.

At the end of the survey, we asked participants in the no-group condition to rate themselves on a sliding scale (0-100, 0 being very dissimilar and 100 very similar) how much they think they are similar to the other player. There was no significant difference between those who were primed in similarity (M = 57.76) and primed in differences (M = 58.09), t(168) = 0.11, p = 0.91. Therefore the prime did not seem to evoke significant effects in explicit perception of similarity.

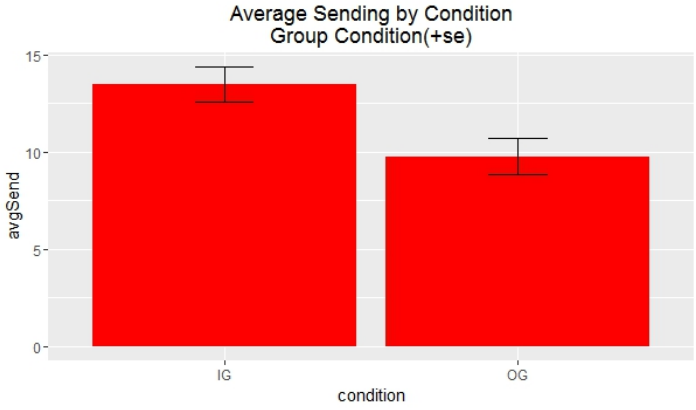
Both Prime and Group Condition (Similar/Ingroup, Similar/Outgroup, Dissimilar/Ingroup, Dissimilar/Outgroup)

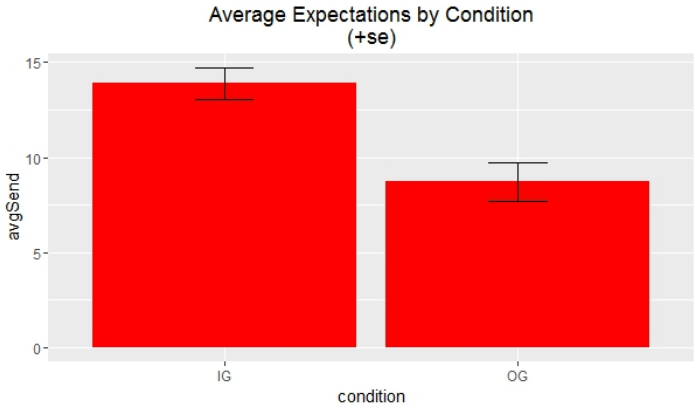
Participants sent more to ingroup members than to outgroup members. Again, the similarity prime had no effect.

A two-way ANOVA was conducted on the amount sent in group condition (ingroup or outgroup) and in the prime condition (similarity or differences prime) to see if there was a significant interaction. There was a significant main effect of group condition, F(1, 264) = 4.68, p = 0.03. Participants sent significantly more to ingroup members (M =12.24) than to outgroup members (M = 9.80), t(263) = 2.37, p = 0.02.There was no significant main effect of similarity prime, F(1,264) = 0.02, p = 0.87. There was no significant interaction between group condition and similarity prime condition, F(1,264) = 1.16, p = 0.28.

Participants expected more from ingroup members than from outgroup members. Again, the similarity prime had no effect.

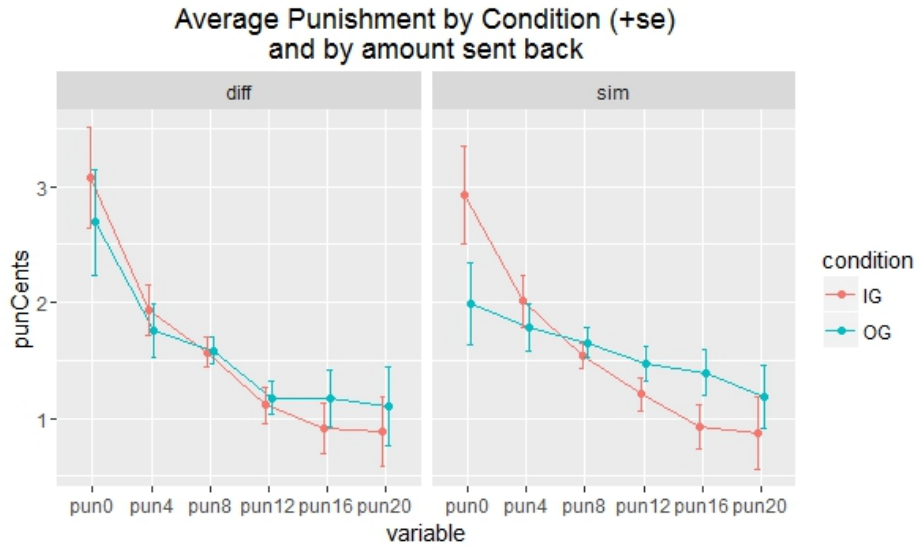
A two-way ANOVA was conducted on amount expected in group condition (ingroup or outgroup) as well as in the prime condition (similarity or differences prime) and yielded similar results. There was a significant main effect of group condition, $F(1,264) = 3.92$, $p = 0.04$. Participants expected more from ingroup members ($M = 11.91$) than outgroup members ($M = 9.80$), $t(263) = 2.05$, $p = 0.04$. There was no significant main effect of similarity prime on expectation, $F(1, 264) = 0.48$, $p = 0.49$. There was no significant interaction between group condition and similarity prime on expectation, $F(1, 264) = 0.10$, $p = 0.74$.





Punishment depended on the amount the participant received from the other player and whether the other player was an ingroup member or outgroup member.

Punishment was significantly though weakly correlated with the amount sent in both the group condition ($r = -0.17$, $p < 0.001$) and in the no-group condition ($r= -0.13$, $p < 0.001$). The less the other player sent, the greater the punishment. For expectation of punishment, there was no significant effect of group condition or prime nor was there a significant interaction between the two variables. A two-way ANOVA was conducted on punishment based on group condition and amount participant received from the other player. There was no significant main effect on punishment based on group condition, but there was a significant main effect based on amount the other player sent, $F(1,1604) = 47.90$, $p<0.001$, and a significant interaction between group condition and the amount the other player sent, $F(1,1604) = 4.71$, $p = 0.03$.



When participants were sent little (0 cents or 4 cents) from the other player, they punished ingroup players more than outgroup players. When participants were sent more (16 cents or all 20 cents) from the other player, they punished outgroup players more than ingroup players.

To examine this significant interaction between group condition and the amount the other player sent, we performed post hoc t-tests for each possible amount the other player could send to the participant (0, 4,8,12,16, or all 20 cents). This was done to see how the participants' punishment decisions differed for each amount depending on whether the other player was an ingroup member or outgroup member. For all of these individual t-tests, there were no statistically significant differences found in punishment toward ingroup members and outgroup members at each amount.

However, there is an emerging trend. If the other player sent nothing (0 cents), participants punish ingroup players more (M = 2.99) than outgroup players (M = 2.29), t(259)=1.39, p = 0.16. This is also true when the other players sent 4 cents, participants punish ingroup players more (M = 1.97) than outgroup players (M = 1.78), t(261)= 0.51, p = 0.61. At 8 cents and at 12 cents (amounts in adherence to fairness norms), participants punished ingroup and outgroup members at almost identical levels. If the other player sent 16 cents, participants punish outgroup players (M=1.30) more than ingroup players (M=0.91), t(256)=-1.20, p = 0.23. This also occurs if the other player sent all 20 cents -- outgroup members are punished more (M=1.16) than ingroup members (M=0.87), t(261)=-0.81, p=0.42.

Using the No-group, Prime Only Condition as a Control

Using the no-group condition for baseline comparison reveals ingroup norms and outgroup negativity.

Since the prime did not work, we treated the "no-group, prime only" condition as a no-group condition. Therefore, we could use the data from this no-group condition as a baseline control to compare ingroup data and outgroup data to. Participants sent more to the ingroup (M=12.24) than to the no-group control (M=11.29), p = 0.32. Participants also expected more from ingroup players (M= 11.91) than from no-group players (M=11.19), p = 0.45. Participants punished ingroup players who sent 0 (M=2.99) cents or 4 cents (M=1.97) significantly more than no-group players who sent 0 cents (M=1.89) or 4 cents (M=1.29); t(250)= 0.02, p=0.01 and t(243)= -1.97, p=0.05, respectively. This supports the prediction the Norms Focused Hypothesis makes regarding norm violation, in which ingroup members are punished more. There does not seem to be evidence of ingroup favoritism, as the Mere Preferences Hypothesis predicts, at the other amounts sent (8, 12, 16, and 20 cents). At these cooperation levels, participants punished ingroup members at similar rates as no-group members (and even at slightly greater rates, though high p-values indicate this could largely be due to chance).

Participants sent less to outgroup players (M=9.96) than to no-group players (M=11.29), p=0.16. They also expected less from outgroup players (M=9.81) than no-group players (M=11.20), p=0.15. There were no statistically significant differences in punishment toward outgroup players and no-group players, but across all amounts sent, participants punished outgroup players more than no-group players. This trend of greater punishment toward outgroup players regardless of the amount they send to the participant implies outgroup negativity.

Belief that the Other Player is Real

At the end of the game, we asked participants if they believed they were playing the game with a real person. We took many precautions to ensure the participants that they were in fact, matched up with another real person on mTurk. At each phase of the game, we added reminders that said, "NOTE: The game is not played in real time. The other players are REAL, and your decision will determine how much bonus you and the other player actually receive. Once the HIT is over, we will use your decision from this page to determine how much to reduce the other player's payoff, and the cost to you for doing so. Then, we will use your decisions and the other player's decisions from phase 1 and phase 2 to calculate your bonus and the other player's bonus." However, in the no-group condition, 60 out of 114 players did not believe they were matched with a real person. Similarly, in the group condition, 137 out of 332 players did not believe they were matched with a real person.

Disbelief that the other player was real did not affect sending, expectation or punishment across conditions.

Yet when independent sample t-tests were conducted, there was no significant effect on the amount sent in the no-group condition, $t(117) = 0.10$, $p = 0.919$ nor in the group condition, $t(275) = 1.64$, $p = 0.10$. In addition, there was no significant effect on the amount expected in the no-group condition, $t(111) = 0.83$, $p = 0.41$, nor in the group condition $t(286) = 0.94$, $p = 0.35$. Group effects in sending and expectation still emerged even when participants did not believe the other player was real. Disbelief that the other player was real also did not affect punishment decisions, $t(687) = -0.94$, $p=0.35$. Oddly, believing the other player is real does not seem to matter.

Discussion

Similarity Prime

We did not successfully replicate the similarity prime effects from Mussweiler and Ockenfels (2013). In both no-group and group conditions, we did not find any significant effect of the similarity prime on the amount sent to the other player, expectation of what the other player will send, or on costly punishment of the other player. The way the prime was executed was almost identical to their study – we used the same images with the same instructions. There were few and subtle differences. Their study took place in Germany, whereas we recruited exclusively in the United States because of the political nature of the group manipulation (Democrat or Republican). They carried out their study out in crowded lecture halls with undergraduate students, whereas we recruited online adult participants of a wider age range on Amazon's Mechanical Turk.

We asked participants to list as many similarities (or differences) as possible and left 10 spaces, and required a minimum of 5 responses. We determined 5 would be an appropriate minimum number of responses because we did not want to require too many similarities. As it becomes more difficult to come up with more similarities, participants may actually develop a belief that the two pictures are more different than similar, thus defeating the purpose of the similarity prime. The majority of participants gave the minimum of 5 responses. Mussweiler and Ockenfels did not ask for a minimum number of responses like we did. Rather, they gave the participants time until the last student finished the task before allowing everyone to move on to the economic game. While it is a novel way to study mere similarity without ties to group membership, it is quite possible that Mussweiler's priming task may not be strong enough to

induce attitudes of similarity toward the other player in the game and future attempts to replicate this finding should be taken.

Ingroup Norms and Outgroup Negativity

While we found no effects on mere similarity because the similarity prime manipulation did not hold, we found significant effects of group condition on the amount sent to the other player, on expectation of what the other player will send, and on costly punishment of the other player. As hypothesized and supported by past literature, participants gave more to ingroup members in the sending phase and also expected to receive more from ingroup members (McLeish and Oxoby, 2011). Regardless of whether they were in the no-group or group condition, there was still a strong correlation between the amount the participant sent and the amount they expected to receive. It seems reciprocity is anticipated across ingroup and outgroup membership, though more is sent to and more is expected from the ingroup member. There are a number of hypotheses that attempt to explain this favoring of ingroup members and higher expectations of ingroup members when sharing resources. Both the Mere Preferences Hypothesis and Norms Focused Hypothesis are consistent with this finding.

We found an interesting trend of group effects on punishment: if the other player sent only 0 cents or 4 cents, punishment for ingroup members was higher than for outgroup members. If the other player sent 8 or 12 cents, the punishment levels are about the same for both ingroup and outgroup members. If the other player sent 16 or 20 cents, punishment for outgroup members is higher. This aligns with both the Norms Focused Hypothesis and the Mere Preferences Hypothesis, and our findings nuance their predictions. As predicted by the Norms Focused hypothesis, ingroup members who gave little were punished more than outgroup

members who gave little. Therefore, fairness norms are more enforced within groups. However, outgroup members who gave more were more likely to be punished than ingroup members who gave more. These are cases where participants carried out costly punishment, even when the other player sent more than half their resources. It seems to be driven not by a response to norm violation or failure to cooperate, but rather by a general dislike of outgroup members. So a potential driving factor behind this behavior is not mere preference for ingroup members, but mere dislike for outgroup members.

Moral Outrage and Competitive Social Motivation

A possible explanation for this punishment pattern is that monetary punishment involves two motivational bases: 1) moral outrage against norm violators and 2) competitive social motivation (Masclet et al., 2003). Past studies have found that the main motivational basis for ingroup punishment is moral outrage, whereas moral outrage is not related to outgroup punishment (Shinada et al., 2004). Ingroup punishment driven by moral outrage toward norm violators is consistent with the Norms Focused Hypothesis, which argues that norms are enforced in within group contexts.

Outgroup punishment may be explained by competitive social motivation, as participants reduced the payoff of outgroup players regardless of their cooperation levels to enhance the relative standing of the self against others. This would be consistent with self-favoritism, a theoretical underpinnings of the Mere Preferences Hypothesis. In our study, any money spent on punishment will take twice that amount away from the other player. Therefore, the cost/benefit factor is 2. This ratio allows for punishment to be used as a means to enhance personal standing in comparison to the outgroup. If the cost/benefit ratio is lowered to 1 (in which case, any money

spent on punishment will take the same amount away from the other player), the punisher loses as much as the punished and cannot improve relative standing by punishing an outgroup member (Shinada et al., 2004). When Shinada et al. replicated their experiment with a cost benefit ratio of 1, outgroup punishment decreased greatly and ingroup punishment was only marginally affected. This supports our finding that outgroup punishment is driven by self-enhancement and outgroup negatvity, which is consistent with the Mere Preferences account.

Ingroup Positivity, Outgroup Negativity, or Both?

In many studies regarding group bias in cooperation, it is unclear whether the bias is for the ingroup, against the outgroup or both. In the framework of the Mere Preferences Hypothesis, it is unclear whether our positive attitude toward our ingroup is necessarily coupled with a negative attitude toward outgroups. Schiller et al. (2014) found both ingroup favoritism and outgroup negativity are at work in third-party punishment games by comparing the punishment of ingroup and outgroup norm violators to a baseline of unaffiliated norm violators.

Because our similarity prime was unsuccessful, we were able to use the "no-group, prime only" condition as a no-group baseline to compare the ingroup and outgroup conditions to. We found evidence for outgroup negativity because though the differences in punishing outgroup and no-group players were not statistically significant, across all amounts sent, participants punished outgroup members more than no-group players. If ingroup positivity were to be present, we would predict that punishment levels for ingroup players would be lower than no-group players, especially when there are no norm violations. However, we found no pattern of ingroup positivity. Outside of norm violations (in which participants punished ingroup players

more than no-group players), participants punished ingroup players at very similar levels as no-group players.

Limitations

Similarity Prime

The similarity prime was not successfully replicated, but regardless, it is unclear whether this is a strong enough manipulation to be paired with explicit group identity manipulation. Rather than priming for similarity, perhaps another method of evoking mere similarity without group membership can be considered for future studies.

Political Polarization in Election Season

This study was conducted in 2016, which is a presidential election year in the United States. We used political party as the group manipulation, and it is possible that increased political polarization due to election season induced or magnified group effects.

Strategy Method

We asked for punishment decisions using a strategy method, in which participants make conditional decisions for each possible amount sent to them by the other player. Some literature has found this method leads to different experimental results than the standard direct-response method (in which the participant learns the action of the other player and then chooses how to respond), while some literature has not found a significant difference (Brandts and Charness, 2011). We decided to use the strategy method because it allowed us to pair real partners outside of real time using an online platform. It was also more compatible with our study design of a one-shot game because repeated games (which would be more compatible with a direct-response method) must take into account reputation effects.

Difficulty Distinguishing Group and Similarity Effects

While Mussweiler and Ockenfels argued that similarity is tied to social projection and expectation of shared norms, it is also arguable that preference for ingroup members is tied to attraction to others similar to self. Social networks are often homogenous in many ways, and this tendency of individuals to associate and bond with similar others is called homophily, or "love of the same" (Mussweiler and Ockenfels, 2013). Homophily can be found along the developmental timeline. Adults, 3-year-olds (Fawcett and Markson, 2010), and even prelinguistic infants prefer similar others (Mahajan and Wynn, 2012). This would predict greater leniency toward similar others in norm violation, contrary to what Mussweiler and Ockenfels found. Given this overlap, effects of group membership and similarity in terms of norms and preferences can be hard to parse.

Conclusion and Future Directions

People favor ingroup members when sharing resources, expect more from them, and punish them more than outgroup members for norm violations. This finding supports the Norms Focused Hypothesis as cooperation norms were enforced within the group. However, outgroup members who did not violate norms were punished more than ingroup members who did not violate norms. This seems to be driven by competitive social motivation and outgroup negativity, while evidence for ingroup positivity was not found.

It is still unclear whether mere similarity or group membership is operating. No one has yet fully crossed mere similarity and group membership to identify the relative weight of these two factors, though understanding the mechanisms underlying group biases in cooperation and punishment could help us understand why people choose to cooperate, defect, or punish when

interacting with an economic partner in their ingroup or outgroup. Parsing out mere similarity and group membership can be difficult, as shared group membership is an indicator of meaningful shared similarity. Future research should continue testing similarity primes or find other avenues to examine mere similarity's role in cooperation.

Acknowledgements

Many thanks to Yarrow Dunham for his guidance, Shaina Coogan for her support, the post-doctoral fellows and graduate students affiliated with the Social Cognitive Development Lab for their constructive feedback, and Antonio Alonso from the Human Cooperation Lab for training on mTurk command line tools. I am also grateful to my peers in the Cognitive Science major, for peer reviewing my thesis and for undergoing this process together.

References

Ames, D. R. (2004). Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of personality and social psychology*, *87*(5), 573.

Baron, A. S., & Dunham, Y. (2015). Representing 'Us' and 'Them': Building Blocks of Intergroup Cognition. *Journal of Cognition and Development*, *16*(5), 780-801.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912-915.

Betancourt, H., & Blair, I. (1992). A cognition (attribution)-emotion model of violence in conflict situations. *Personality and Social Psychology Bulletin*,*18*(3), 343-350.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, *14*(3), 375-398.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them social identity shapes neural responses to intergroup competition and harm. *Psychological science*.

Fawcett, C. A., & Markson, L. (2010). Similarity predicts liking in 3-year-old children. *Journal of experimental child psychology*, *105*(4), 345-358.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological review*, *109*(1), 3.

Mahajan, N., & Wynn, K. (2012). Origins of "us" versus "them": Prelinguistic infants prefer similar others. *Cognition*, *124*(2), 227-233.

Marques, J. M., & Paez, D. (1994). The 'black sheep effect': Social categorization, rejection of ingroup deviates, and perception of group variability. *European review of social psychology*, *5*(1), 37-68.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review*, *93*(1), 366-380.

McAuliffe, K., & Dunham, Y. (2016). Group bias in cooperative norm enforcement. *Phil. Trans. R. Soc. B*, *371*(1686), 20150073.

McLeish, K. N., & Oxoby, R. J. (2011). Social interactions and the salience of social identity. *Journal of Economic Psychology, 32*(1), 172-178.

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For members only ingroup punishment of fairness norm violations in the ultimatum game. *Social Psychological and Personality Science*, *5*(6), 662-670.

Mussweiler, T. (2001). 'Seek and ye shall find': antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology*,*31*(5), 499-509.

Mussweiler, T., & Ockenfels, A. (2013). Similarity increases altruistic punishment in humans. *Proceedings of the National Academy of Sciences*,*110*(48), 19318-19323.

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, *35*(3), 169-175.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. *Evolution and Human Behavior*, *25*(6), 379-393.