Humans reciprocate by discriminating against group peers

David Hugh-Jones[*] Itay Ron[†] Ro'i Zultan[‡]

Abstract

The evolution of human intergroup conflict is a social science puzzle. Motivated by cycles of intergroup revenge in real-world conflicts, we experimentally test the hypothesis that humans practice group-based reciprocity: if someone harms or helps them, they harm or help other members of that person's group. Subjects played a trust game, then allocated money between other people. Senders whose partners returned more in the trust game gave more to that partner's group members. The effect was about half as large as the effect of direct reciprocity. Receivers' allocations to group members were not affected by their partners' play in the trust game, suggesting that group reciprocity was only triggered when the partner's intentions were unequivocal. We show conditions under which group reciprocity can evolve, and discuss its place in conflict among early humans.

Keywords: Upstream reciprocity, group identity, intergroup conflict.
Word count: 3129

[*]School of Economics, University of East Anglia. E-mail: D.Hugh-Jones@uea.ac.uk.
[†]E-mail: itayron@gmail.com
[‡]Department of Economics, Ben-Gurion University of the Negev. E-mail: zultan@bgu.ac.il.

1 Introduction

Human society is organized in groups, including families, clans, firms and nations. This structure is reflected in individual behaviour and cognition. Humans identify with their ingroup and are altruistic and prosocial towards ingroup members; towards outgroup members, they display stereotyping and prejudice (Balliet, Wu, and De Dreu, 2014; Chen and Chen, 2011; Chen and Li, 2009; De Dreu, Balliet, and Halevy, 2014; Tajfel and Turner, 1979; Yamagishi and Kiyonari, 2000). Group structure provides the backdrop for intergroup conflict—from economic and political competition to inter-ethnic violence and war—which is pervasive in the species (Esteban, Mayoral, and Ray, 2012) and has serious economic costs (World Bank, 2011).

Intergroup conflicts often follow a tit-for-tat logic, in which one group's violence leads to revenge from the other side (Chagnon, 1988; Haushofer, Biletzki, and Kanwisher, 2010; Horowitz, 1985; Horowitz, 2001; Shayo and Zussman, 2010). This suggests that humans practice intergroup *reciprocity*. Reciprocity is a well-known mechanism that may underlie the evolution of cooperation (Nowak, 2006, 2012). While in direct reciprocity, individuals help those who have helped them in the past (and similarly for harm), in indirect reciprocity, individuals help or harm other people than those who have helped them. Indirect reciprocity comes in two flavours: *downstream* reciprocity follows the maxim 'do unto thy neighbour as they have done to others', whereas *upstream* reciprocity follows the maxim 'do unto thy neighbour as others have done unto you'.[1]

In this paper we examine group-based upstream reciprocity, or *group reciprocity*. That is, an individual who is harmed (helped) by a member of an outgroup becomes more likely to harm (help) others from that group. Whereas group-based downstream reciprocity (Bernhard, Fehr, and Fischbacher, 2006; Bernhard, Fischbacher, and Fehr, 2006) follows the maxim 'do unto others as they have done to members of *my* tribe', group-based upstream reciprocity follows the maxim 'do unto others as members of *their* tribe have done to me'

[1] See Greiner and Levati (2005), Güth, Königstein, Marchand, and Nehring (2001), and Tsvetkova and Macy (2014, 2015) for experimental evidence of upstream reciprocity.

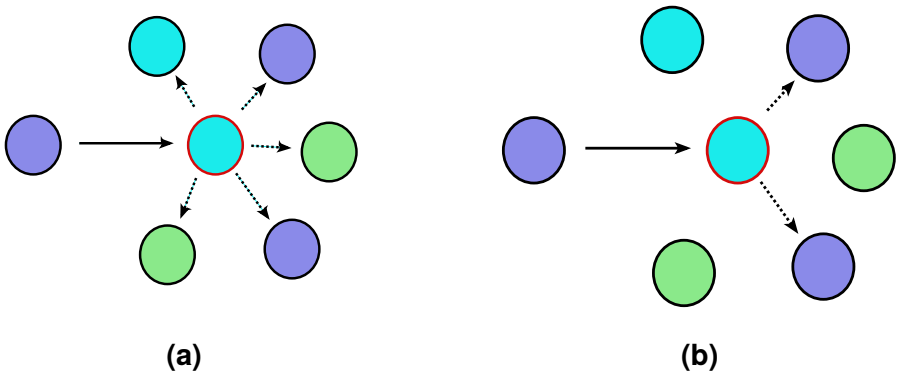


Figure 1: Upstream reciprocity. (*a*) Someone who was helped or harmed becomes more likely to help or harm others. (*b*) Upstream group reciprocity targets people who belong to the same group as the initial partner.

(Figure 1). Tit-for-tat conflict looks like negative group reciprocity.

The concept of group reciprocity may help to explain the evolution of intergroup conflict. The current literature includes three differing approaches to understanding this. While cultural theories argue that there is no innate tendency to intergroup aggression, theories of parochial altruism argue that intergroup violence was a driver of within-group altruism via group selection processes; as a result, intergroup violence can involve self-sacrifice for one's group members (Bowles, 2009; Choi and Bowles, 2007). The "chimpanzee model", by contrast, argues that early humans, like chimpanzees, only attack when odds are very favourable; thus a human tendency to kill outgroups evolved by individual selection alone (Wrangham and Glowacki, 2012). This is supported by evidence that both hunter-gatherers and chimpanzees are rarely wounded when they attack.

Kelly (2000) argues that a defining characteristic of war is "social substitutability", whereby members of a perpetrator's group become legitimate targets for revenge. Social substitutability is especially found in segmented societies, which typically feature strong corporate identities such as extended patrilineal families and clans. Some of these societies also have "war/peace systems" featuring well-defined institutions for ending conflict as well as beginning it, such as the Andamanese Peace Dance or the Montenegrin Court of

Good Men for ending feuds (Boehm, 1984). By contrast, while chimpanzees do practice retaliation and reconciliation among alliances within the community, they do not reciprocate towards other groups. Instead, they attack stranger chimpanzees whenever it is safe to do so. The risk of being attacked forces chimps to avoid territory bordering other communities, which limits their available space for foraging (Wilson and Wrangham, 2003). While this fact seems to favour the evolution of peaceful intergroup relations (Kelly, 2005), that ignores the prisoner's dilemma structure of intergroup relations; while both groups would do better not attacking the other, each group does better by attacking when the odds are good enough. Indeed, peaceful unsegmented societies resolve intergroup conflict by avoiding the other group, which entails a loss of access to valuable resources, and hence lower population density. The evolution of group reciprocity could deter opportunistic conflict. When there is group reciprocity, someone who harms an outgroup member brings retaliation on his own group, and this gives his group members an incentive to maintain peace (Boehm, 1984). Group reciprocity could thus have benefited humans by allowing them to range over wider areas and to have more extensive contacts with outgroups. So, the evidence in Kelly (2000) that population density is associated with war can be read two ways: the development of war, particularly of war/peace systems, may allow different groups to live at high densities in peace.

While group reciprocity can benefit the group, to evolve it must increase individual fitness. In the supplementary materials, we report on a series of simulations tracking the evolution of group reciprocators under different environmental parameters. The key result is that when the relative cost of helping (or failing to harm) a target is low enough, relative to the benefit supplied to the target and to the group size, group reciprocity can evolve to fixation in a group. The mechanism is that groups with a high share of group reciprocators learn to cooperate with each other, while not helping groups that have a high share of selfish types. Selfish types in cooperative groups free ride on the group reputation, exploiting members of other cooperating groups, thereby gaining higher fitness then the group reciprocators in their group. Nonetheless, if the share of group reciprocators within the group varies sufficiently

between groups, group reciprocators are overrepresented in the cooperative groups, and therefore have higher fitness overall. The high benefit/cost ratio required for group reciprocity to evolve fits the "chimpanzee model" of conflict where an attack is only launched if it is low risk. It may also explain why most salient examples of group reciprocity are seen in the negative domain of harm and conflict: when the benefit/cost ratio is high, hurting or refraining from helping is a "nasty" action, since it imposes a large loss on the target for a small gain to oneself.

Another factor that could support the evolution of group reciprocity is intra-group dependencies. If group members help each other, e.g. by providing public goods, then punishing a perpetrator's group member indirectly harms the perpetrator. Our simulations show that group reciprocity may emerge even absent such intra-group dependencies, but their presence would lower the benefit/cost threshold needed for it to evolve.

Kelly (2000) argues that group reciprocity emerges in a segmented society, where group affiliations are rigid and salient. However, our simulations show that group reciprocity can evolve even without in-group sanctioning or collective norms of group responsibility and liability. That, and real world examples of apparent intergroup revenge in large modern societies, suggest there may be a universal propensity to group-reciprocate. In this paper, we aim to study the existence and form of the proximate psychological mechanism for group reciprocity in modern humans. Although field observations from conflict are highly suggestive, they are loaded with individual and group context and history. Moreover, in the field, group reciprocity and direct reciprocity may be conflated, because it is difficult to distinguish between retaliatory acts directed at groups that include the perpetrator, and acts directed at unrelated group members. We therefore designed a controlled laboratory experiment to test the psychological mechanism by which humans reciprocate towards groups in a clean way. That is, we test the hypothesis that people reciprocate towards groups.

Cleanly identifying group reciprocity requires controlling for three confounds: individual level reciprocity, e.g. if subjects' actions affect an entire group including the original actor who helped or harmed them; generalized reciprocity,

where subjects reciprocate not specifically towards the original actor's group, but towards other people in general; and strategic interactions, where apparent reciprocity is driven by reputation-building. Our experiment fulfils all three: subjects can differentiate the original actor from his or her group members, they interact both with these group members and with members of other groups, and we minimize strategic concerns by not giving feedback about subjects' actions.

While previous studies looked at retaliation towards groups, this retaliation does not necessarily reflect group reciprocity as defined here. Gaertner, Iuzzini, and O'Mara (2008) found that rejection by one group member leads to more hostility towards the group when the group is perceived as a unified entity. Since hostility was directed towards the whole group, individual and group level reciprocity were confounded. Similarly, Böhm, Rusch, and Gürerk (2016) examine intergroup retaliation using the intergroup prisoner's dilemma paradigm, but cannot distinguish between individual and group reciprocity. Stenstrom, Lickel, Denson, and Miller (2008) manipulated entitativity by making the original perpetrator (a political analyst) an official affiliate of the group (a presidential campaign). Thus, holding the group accountable for its member's action is justified without resorting to group reciprocity. In contrast, we look at how people reciprocate a clear individual act by one group member to an unrelated other group member, where group structure is minimal and free of existing social context.

Our experimental set up was the following. After an initial group-formation stage, participants interacted in two strategic stages. The upstream action, in which the individual could be helped or harmed by another person, was represented by a Trust Game (TG) (Berg, Dickhaut, and McCabe, 1995). In this game, the Sender (S) receives 150 money-equivalent tokens, and chooses how many of them to send to the Responder (R). The amount sent is multiplied by a factor of 3, so that R receives between 0 and 450 tokens, of which he can send any number back to S. While Rs' actions clearly have a benefit/cost ratio of 1 (money returned to S is lost to R), Ss may send money in the expectation of having money returned. In addition, not returning money in the trust game violates a social norm (Kimbrough and Vostroknutov, 2015). For this reason,

we expected R's actions to elicit stronger reciprocity, although we test the effect of both S's and R's actions.

The upstream action was followed by the reciprocal action, in which the individual could help others. We implemented this as an Allocation Game in which subjects divided a fixed amount between two recipients. In Direct Reciprocity rounds, the recipients included the TG partner; in Group Reciprocity rounds, a member of the TG partner's group; and in Ingroup Favoritism rounds, a member of the allocator's group. The other recipient was always a member of a third, neutral, group. Baseline rounds included two neutral recipients, to test whether the TG experience leads to arbitrary discrimination in the absence of any reciprocal or group motivations.

Our expectations were as follows. First, in Direct Reciprocity rounds, individuals' allocations to their TG partner should positively covary with the amount the partner sent (or returned) in the Trust Game. This simply comes from from the well-known theory of direct reciprocity. Second, if group reciprocity is present, then allocations to the TG partner's group member, in Group Reciprocity rounds, should also covary with the amount sent or returned by the TG partner. We also measured participants' social value orientation (Van Lange, 1999). It is plausible that willingness to group-reciprocate should be linked to other social preferences. We were not certain *a priori* whether group reciprocity would be stronger among selfish or among prosocial types. On the one hand, both prosociality and group reciprocity can be seen as actions that benefit the group, by providing support to ingroup members or protecting it from outgroups. On the other hand, negative reciprocity in general may be linked to spite (Johnstone and Bshary, 2004). So we test a non-directional hypothesis here.

2 Material and methods

Each session consisted of 24 participants, randomly allocated into six *teams* of four. Each participant was identified throughout the experiment by team colour and individual number (1–4) within the team. At the beginning of the

experiment, participants were informed that the experiment had five distinct stages, and that they might interact with the same people in different stages. Specific instructions for each stage were distributed and read aloud at the beginning of the stage. The five stages were a group formation stage, the TG stage, the Allocation Game stage, a social value orientation elicitation stage (Murphy, Ackermann, and Handgraaf, 2011) and a collectivism scale measurement stage (adapted from the horizontal collectivism scale in Singelis, Triandis, Bhawuk, and Gelfand, 1995). Other than the collectivism measurement, all decisions were incentivized.

Following (Chen and Li, 2009), we created group identity in the first stage by allowing participants to consult each other by anonymous chat while solving a simple task. Participants solved five Raven matrices (see supplementary material). Each matrix was presented on screen for 120 seconds, during which each participant could both send written messages to the team and update her own answer. The final answer submitted at the end of the 120 seconds determined payoffs, with 10 tokens paid for each correct answer. To further boost group identity through a common goal, team members each earned an additional bonus of 5 tokens if all four team members answered correctly.

Next, participants were rematched into pairs to play the one-shot TG. To facilitate understanding, participants played five practice rounds, in which they entered decisions both as S and as R. In the actual interaction, participants could see their TG partner's team colour and individual number.

The third stage Allocation Game consisted of six rounds. In each round, participants interacted in groups of three. Individuals in each group were identified to each other by team colour and number. Each round consisted of a random dictator game, as follows. Each player in the group of three had to allocate 100 tokens within the group. The allocator always received exactly 30 tokens, and could freely allocate the remaining 70 tokens between the other two players. Previous research has found that people do not harm, but refrain from helping negatively perceived outgroups (Weisel and Böhm, 2015). Accordingly, we set the parameters of the game so that an equal division between the other two players provides them with 35 tokens each, more than the allo-

Table 1: Matching example

Round	Allocates to			Treatment
1	Red 1	/	Yellow 1	Group reciprocity (GR)
2	Yellow 4	/	Brown 2	Group reciprocity (GR)
3	Green 3	/	Yellow 2	Direct reciprocity (DR)
4	Red 1	/	Brown 1	Baseline (B)
5	Brown 2	/	Brown 4	Baseline (B)
6	Blue 3	/	Green 2	Ingroup (IG)

Note: Example treatments shown for player Blue 2, who played the TG with Yellow 2 (see the supplementary material for the full matching scheme).

cator's own share.[2]

One round of the six rounds was randomly chosen for payment. In that round, the payoffs of the members of each group were determined by the allocation decision of one randomly chosen player in the group. No feedback was provided between rounds. At the end of the stage, players learned the payoff round, whether their allocation was chosen to determine payoffs, and their payoff for the round. Thus, all allocation decisions were completely independent of each other, both within and between participants.

Table 1 shows the matching scheme over the six rounds. Each participant was matched to be in the same group of three with a member of her own team in one of the six rounds (*ingroup* condition), with her TG partner in another round (*direct reciprocity* condition), and in two other rounds with other members of the TG partner's team (*group reciprocity* condition). The remaining two rounds served as the baseline condition. Note that the matching is not independent. For example, if one player is in the direct reciprocity condition, then one other player is in the direct reciprocity condition and the third player is in either the baseline or group reciprocity condition.

The fourth stage implemented the slider measure of social value orienta-

[2] The allocator's decision is not costly, which might have introduced additional confounding considerations. As our aim in this paper is to identify and study the qualitative characteristics of group reciprocity, we accept the limitations that this choice imposes on the ability to estimate the *magnitude* of preferences for group reciprocity.

tion (Crosetto, Weisel, and Winter, 2012; Murphy, Ackermann, and Handgraaf, 2011), in which participants choose nine allocations between themselves and another person. For consistency with the previous stages, the team identity of the partner was known. To keep the decision independent of previous experience with the different teams, we matched participants within teams. Therefore, this measure captures within-group social value orientation. Payoffs were determined by one randomly chosen decision of the nine decisions made by one randomly chosen player in each dyad. The decisions yielded a social orientation angle for each participant, with 0° corresponding to selfishness, 45° to pure altruism, and negative angles to spitefulness.

After the fifth and final stage (a non-strategic and non-incentivised collectivism measurement), participants learned their cumulative payoff in tokens and were paid in private. One hundred and ninety two participants, recruited using ORSEE (Greiner, 2015) participated in eight sessions conducted between June 2014 and January 2015. The experiment was programmed in z-Tree (Fischbacher, 2007). The average payment was approximately $18) for a duration of 70 minutes. The lowest and highest payments were approximately $6 and $32, respectively.

The key outcomes in this design are based on the allocation decisions made in the third stage. Direct and group reciprocity can be both positive and negative, and therefore are not hypothesized to have a systematic effect on the the amount allocated to either the TG partner or to his team mates. Nonetheless—while there is arguably no reason to discriminate between two neutral players—we hypothesize that direct and group reciprocity will lead the allocator to discriminate either for or against the TG partner or his team mates. Consequently, we predict that the absolute difference between the two allocations will be larger in all treatments compared to the baseline. This difference is measured in our 'Discrimination' outcome.

We measure reciprocity directly by looking at the effect of the TG experience in the second stage on allocations made in the third stage. We define the experience with the TG partner in two ways. For responders, this is the amount sent to them by their partner. For senders, we calculate the amount returned to them by their partner as a fraction of the money available to the responder.

Table 2: Allocations and Discrimination

	Allocation	Discrimination	Reciprocity
Senders			
Baseline	35.00 (—)	4.15 (1.13)	—
Direct Reciprocity	33.98 (2.30)	22.00 (1.51) ***	15.64 (5.12)**
Group Reciprocity	34.39 (0.77)	8.08 (1.61) ***	7.78 (2.37)**
In-Group	38.98 (1.11) ***	15.46 (2.99) ***	0.20 (5.50)
Responders			
Baseline	35.00 (—)	2.25 (0.51)	—
Direct Reciprocity	35.38 (1.08)	22.17 (2.30) ***	20.87 (6.04)***
Group Reciprocity	34.79 (0.62)	6.12 (1.51) **	1.20 (2.08)
In-Group	42.13 (1.99) ***	17.20 (3.40) ***	4.72 (7.62)

Mean allocation, mean discrimination, and reciprocity (marginal effect of TG partner's kindness on allocation) by condition. Robust standard errors clustered on sessions. Significance of comparison to Baseline is marked.*, **, and *** indicate p $<$ 0.05, p $<$ 0.01, and p $<$ 0.001, respectively.

Thus, an equal split of the pie implies a value of 1/2, and compensating the sender for his investment implies a value of 1/3. We subsequently define (direct or group) reciprocity as the slope of the allocation made to the TG partner or his team mates on the TG experience.

3 Results

We report results on allocations, discrimination between recipients (measured as the absolute difference between the two recipients' allocations), and direct and group reciprocity. All reported statistical tests are based on linear regressions with standard errors clustered by session.

The first data column in Table 2 presents the mean allocations. Participants gave significantly more to members of their own team at the expense of the neutral recipient ($z = 3.58, p < 0.001$ for senders, $z = 3.59, p < 0.001$ for responders), establishing that our group formation manipulation was successful in inducing group identity and triggering ingroup favouritism. Allocations to the TG partner and his team mates were not significantly different to the baseline 35 ($p > 0.43$ for all comparisons). This result suggests that the experience

with the TG partner is, on average, neutral, such that positive and negative experiences balance each other overall.

Nonetheless, both positive and negative treatment of the TG partner or his team mates increase the absolute difference between the two allocations. Indeed, column of Table 2 shows that allocators discriminated significantly more than in the baseline both when interacting with their TG partner ($z = 9.08, p < 0.001$) and with his team mates ($z = 3.93, p < 0.001$). This effect was not significantly different between TG senders and receivers (F test $0.50, p = 0.68$).

3.1 Direct and group reciprocity

The third column of Table 2, *Reciprocity*, reports the slope of allocations regressed on the subjects' experience with their TG partners. The responder's experience with the sender is measured as the share of the endowment that the sender chose to send. The sender's experience with the responder is measured as the share of the received amount that the responder chose to send back. The sender's experience was not defined for the six (out of 96) senders who did not send any money. There is strong direct reciprocity: allocations to the TG partners increase with the TG experience both for senders ($z = 3.06, p < 0.01$) and for responders ($z = 3.46, p < 0.001$).

Group reciprocity, however, is only observed for senders, who allocate less to team mates of a responder who returned less. Responders, although directly reciprocating the TG partner's action, do not systematically discriminate against team mates of a sender who sent little. The regression analysis shows no significant effect of the responder's TG experience on her allocation to the sender's team mates ($z = 0.58, p = 0.56$). The sender's TG experience, on the other hand, significantly increases the allocations made to the responder's team mates ($z = 3.29, p < 0.01$).[3] The estimated ratio of the group and direct reciprocity coefficients is 50%, so that for every allocation dollar a responder

[3] Responders who receive higher amounts also return a higher share. As a result, senders with a more positive TG experience are, on average, those who sent more in the TG, creating a potential confound. There is no reason, however, why different senders should systematically discriminate between groups in a non-costly way. Indeed, the results hold when we control in the regression for the amount sent and its interaction with the share returned.

loses due to an unkind action in the TG, his team mates lose 50 cents. This relationship is shown graphically in Figure 2 (the corresponding figure for direct reciprocity is included in the supplementary material).

Senders' group reciprocity was related to their social value orientation. The slope of the effect of the TG experience on allocations was 15.97 for those with less than median SVO, and -1.06 for those with median or greater SVO (interaction, $p = 0.061$).These results should be interpreted cautiously, since both scores were affected by the TG experience.

4 Discussion

Our results show that upstream reciprocity is moderated by social boundaries. Humans respond to harms from outgroup members by discriminating against others in that specific outgroup.

Group reciprocity as a proximate mechanism bears implications for human social cognition. While ingroup altruism and group-based downstream reciprocity require people to differentiate their own group from outsiders—"us" from "them"—upstream group reciprocity requires them to differentiate between different outgroups—between "them and them"—and to keep a mental account of outgroups' reputation. Thus, group reciprocity could provide a cognitive foundation for the phenomena of intergroup prejudice and stereotyping (Allport, 1954).[4] Furthermore, by making group reputation a valuable asset, group reciprocity could encourage groups to differentiate themselves symbolically from others, and to police their members' behaviour towards outgroups—both behaviours that we indeed observe in humans (Fearon and Laitin, 1996).

We observed group reciprocity only towards receivers, not senders. On the one hand, we find group reciprocity towards receivers, confirming that the experiment was successful in setting up the type of group interactions that triggers group reciprocity. On the other hand, we find *direct* reciprocity towards senders, indicating that responders perceived the TG interaction as meaningful and relevant for the later allocation decisions. We therefore conclude that

[4]This argument is a between-group parallel to Yamagishi and Kiyonari (2000), which argues that expectations of generalized reciprocity lie behind altruism within a group.

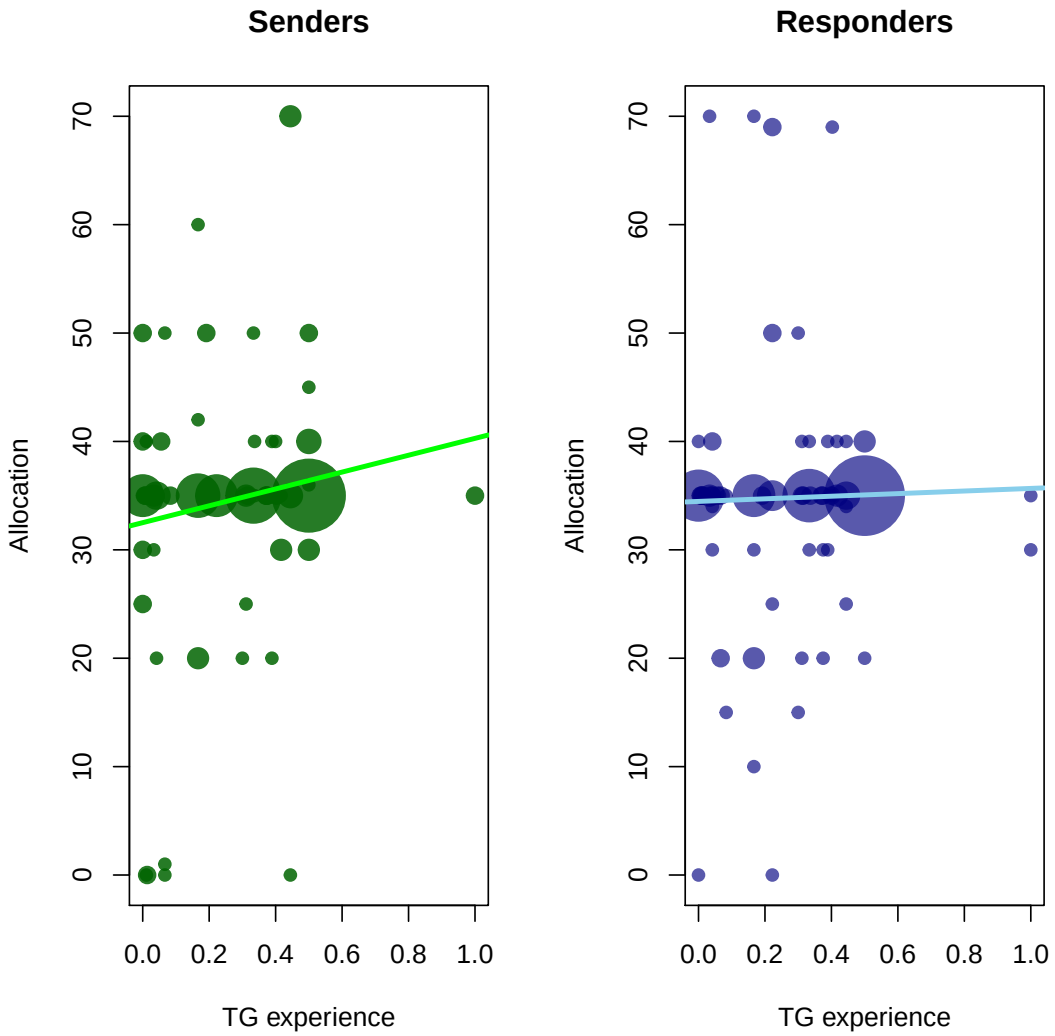


Figure 2: Allocations in the Group Reciprocity condition versus the TG experience. Circles show individual data points (circle size proportional to number of observations). Lines show linear regressions.

it is some characteristic of the responder decision, not shared with the sender decision, that triggers group reciprocity.

One possible interpretation for this difference between senders and responders stems from the distinction between intention-based and outcome-based motives in reciprocal behaviour (Falk and Fischbacher, 2006; Stanca, Bruni, and Corazzini, 2009). In this sense, senders' intentions are more ambiguous, as they do not know what the responder will do. Responders who do not return money, in contrast, are clearly intentionally harming the senders. It is possible that humans generalize *intentions* across group members. That is, if group member 1 takes an action that deliberately harms them, they predict that group member 2 wishes to harm them also. If not returning money is seen as deliberately harmful, while not sending money can be explained by caution or mistrust, then this would generate the difference in group reciprocity that we observe.

Another distinction made in the literature between trust (sender behavior) and trustworthiness (responder behavior) is based on norms and rules of conduct. In their analylsis of Adam Smith's *A Theory of Moral Sentiments*, Wilson and Smith (2017) argue that trust is benificient act, while breaking trust is misconduct. Accordingly, Wilson and Smith (2017) found that people punish responders but not senders. Similarly, Kimbrough and Vostroknutov (2015) found that 'rule followers' are more trustworthy, but not more trusting, than other individuals. We view these interpretations of our results as tentative. Further research will be necessary to map and understand the boundaries of the group reciprocity phenomenon.

We mention some caveats and limitations. First, since our study was conducted with students from a rich industrialized democracy, results may not generalize to all cultures (Henrich, Heine, and Norenzayan, 2010). In particular, the link between intentions and moral judgment may vary across cultures (Barrett, Bolyanatz, Crittenden, Fessler, Fitzpatrick, Gurven, Henrich, Kanovsky, Kushnick, Pisor, et al., 2016), and this could affect how group reciprocity plays out in different societies. Second, our experiment did not differentiate between positive and negative group reciprocity: we leave this for future work.

We have argued that group reciprocity could help explain why some groups

have relatively peaceful intergroup relations. It may also provide a step from the "chimpanzee model" of conflict towards the large-scale, organized intergroup conflicts observed in tribal and state-level societies. For example, Wrangham, Wilson, and Muller (2006) provide evidence that hunter-gatherers and farmers have similar levels of lethal violence to chimpanzees but much less non-lethal violence. This could be because the threat of high-level violence can contain low-level violence. A further step could be provided by "third party" group reciprocity. That is, in many ethnic conflicts, a harm from one group to another is revenged by the entire second group, leading to cycles of intergroup violence. Third party group reciprocity could result from organized groups taking collective action to maintain their reputation as reciprocal (and hence, dangerous to attack).

Upstream reciprocity is notoriously difficult to understand in evolutionary terms (Boyd and Richerson, 1989; Nowak and Roch, 2007). Group reciprocity may provide another piece of the puzzle. Group reicprocity allows individuals to use reciprocal strategies based on group reputation. Consequently, upstream reciprocity can direct group-level selection in ways parallel to those by which direct reciprocity direct individual-level selection. We acknowledge, though do not develop here, two other ways by which group reciprocity may evolve. First, group members are interdependent, especially in the small groups that were the norm during most of human evolutionary history. Punishing a perpetrator's group member therefore indirectly harms the perpetrator, who is dependent on his peers for , e.g., public goods provision. Thus, group reciprocity may bridge upstream indirect reciprocity and direct reciprocity through intra-group dependencies. Second, the evolution of indirect reciprocity acts by way of chains of reciprocal actions, which return with some probability to the original instigator of the chain (Nowak and Roch, 2007). In a population organised in groups, such that individuals interact more frequently with their own group members, group reciprocity may increase the likelihood of successful reciprocal chains, facilitating the evolution of upstream reciprocity. These ideas could be formalized in future work.

References

Allport, Gordon W (1954). *The nature of prejudice*. Cambridge: Addison-Wesley.

Balliet, Daniel, Junhui Wu, and Carsten KW De Dreu (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychol Bull* 140(6), pp. 1556–1581.

Barrett, H Clark, Alexander Bolyanatz, Alyssa N Crittenden, Daniel MT Fessler, Simon Fitzpatrick, Michael Gurven, Joseph Henrich, Martin Kanovsky, Geoff Kushnick, Anne Pisor, et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *P Natl Acad Sci Usa* 113(17), pp. 4688–4693.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). Trust, reciprocity, and social history. *Game Econ Behav* 10(1), pp. 122–142.

Bernhard, Helen, Ernst Fehr, and Urs Fischbacher (2006). Group affiliation and altruistic norm enforcement. *Am Econ Rev* 96(2), pp. 217–221.

Bernhard, Helen, Urs Fischbacher, and Ernst Fehr (2006). Parochial altruism in humans. *Nature* 442(7105), pp. 912–915.

Boehm, Christopher (1984). *Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies*. Philadelphia, PA: University of Pennsylvania Press.

Böhm, Robert, Hannes Rusch, and Özgür Gürerk (2016). What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evol Hum Behav* 37(1), pp. 29–34.

Bowles, Samuel (2009). Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science* 324(5932), pp. 1293–1298.

Boyd, Robert and Peter J. Richerson (1989). The evolution of indirect reciprocity. *Soc Networks* 11(3), pp. 213–236.

Chagnon, Napoleon A. (1988). Life Histories, Blood Revenge, and Warfare in a Tribal Population. *Science* 239(4843), pp. 985–992.

Chen, Roy and Yan Chen (2011). The Potential of Social Identity for Equilibrium Selection. *Am Econ Rev* 101(6), pp. 2562–2589.

Chen, Yan and Sherry X. Li (2009). Group identity and social preferences. *Am Econ Rev* 99(1), pp. 431–457.

Choi, Jung-Kyoo and Samuel Bowles (2007). The coevolution of parochial altruism and war. *Science* 318(5850), pp. 636–640.

Crosetto, Paolo, Ori Weisel, and Fabian Winter (2012). A Flexible z-Tree Implementation of the Social Value Orientation Slider Measure (Murphy et al. 2011): Manual. *Jena Economic Research Papers* (2012-062). Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.

De Dreu, Carsten K.W., Daniel Balliet, and Nir Halevy (2014). "Chapter One – Parochial Cooperation in Humans: Forms and Functions of Self-Sacrifice in Intergroup Conflict". In: *Advances in Motivation Science*. Ed. by Andrew J. Elliot. Vol. 1. Elsevier. Chap. 1, pp. 1–47.

Esteban, Joan, Laura Mayoral, and Debraj Ray (2012). Ethnicity and conflict: Theory and facts. *science* 336(6083), pp. 858–865.

Falk, Armin and Urs Fischbacher (2006). A theory of reciprocity. *Game Econ Behav* 54(2), pp. 293–315.

Fearon, James D and David D Laitin (1996). Explaining interethnic cooperation. *American political science review* 90(4), pp. 715–735.

Fischbacher, Urs (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Exp Econ* 10(2), pp. 171–178.

Gaertner, Lowell, Jonathan Iuzzini, and Erin M. O'Mara (2008). When rejection by one fosters aggression against many: Multiple-victim aggression as a consequence of social rejection and perceived groupness. *J Exp Soc Psychol* 44(4), pp. 958–970.

Greiner, Ben (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. English. *Journal of the Economic Science Association* 1(1), pp. 114–125.

Greiner, Ben and M. Vittoria Levati (2005). Indirect reciprocity in cyclical networks: An experimental study. *J Econ Psychol* 26(5), pp. 711–731.

Güth, W., M. Königstein, N. Marchand, and K.D. Nehring (2001). Trust and Reciprocity in the Investment Game with Indirect Reward. *Homo Oeconomicus* 18, pp. 241–262.

Haushofer, Johannes, Anat Biletzki, and Nancy Kanwisher (2010). Both sides retaliate in the Israeli–Palestinian conflict. *P Natl Acad Sci Usa* 107(42), pp. 17927–17932.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan (2010). Most people are not WEIRD. *Nature* 466(7302), pp. 29–29.

Horowitz, D. L (1985). *Ethnic Groups in Conflict*. Berkeley: University of California Press.

Horowitz, D. L. (2001). *The deadly ethnic riot*. University of California Press.

Johnstone, Rufus A and Redouan Bshary (2004). Evolution of spite through indirect reciprocity. *P Roy Soc B-biol Sci* 271(1551), pp. 1917–1922.

Kelly, Raymond C (2005). The evolution of lethal intergroup violence. *P Natl Acad Sci Usa* 102(43), pp. 15294–15298.

Kelly, Raymond Case (2000). *Warless societies and the origin of war*. University of Michigan Press.

Kimbrough, Erik O and Alexander Vostroknutov (2015). Norms make preferences social. *J Eur Econ Assoc* 14(3), pp. 608–638.

Murphy, Ryan O., Kurt A. Ackermann, and Michel J. J. Handgraaf (2011). Measuring Social Value Orientation. *Judgm Decis Mak* 6(8), pp. 771–781.

Nowak, Martin A. (2006). Five rules for the evolution of cooperation. *Science* 314(5805), pp. 1560–1563.

— (2012). Evolving cooperation. *J Theor Biol* 299, pp. 1–8.

Nowak, Martin A. and Sébastian Roch (2007). Upstream reciprocity and the evolution of gratitude. *P Roy Soc B-biol Sci* 274(1610), pp. 605–610.

Shayo, Moses and Asaf Zussman (2010). Judicial ingroup bias in the shadow of terrorism. *Q J Econ*.

Singelis, Theodore M., Harry C. Triandis, Dharm P. S. Bhawuk, and Michele J. Gelfand (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-Cult Res* 29(3), pp. 240–275.

Stanca, Luca, Luigino Bruni, and Luca Corazzini (2009). Testing theories of reciprocity: Do motivations matter? *J Econ Behav Organ* 71(2), pp. 233–245.

Stenstrom, Douglas M., Brian Lickel, Thomas F. Denson, and Norman Miller (2008). The Roles of Ingroup Identification and Outgroup Entitativity in Intergroup Retribution. en. *Pers Soc Psychol B*.

Tajfel, Henri and John C. Turner (1979). "An integrative theory of intergroup conflict". In: *The Social Psychology of Intergroup Relations*. Ed. by William G.

Austin and Stephen Worchel. Monterey, CA: Brookes/Coole. Chap. 3, pp. 33–47.

Tsvetkova, Milena and Michael W Macy (2014). The social contagion of generosity. *PloS one* 9(2), e87275.

— (2015). The Social Contagion of Antisocial Behavior. *Sociological Science* 2, pp. 36–49.

Van Lange, Paul AM (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *J Pers Soc Psychol* 77(2), p. 337.

Weisel, Ori and Robert Böhm (2015). "Ingroup love" and "outgroup hate" in intergroup conflict between natural groups. *J Exp Soc Psychol* 60, pp. 110–120.

Wilson, Bart J. and Vernon L Smith (2017). Sentiments, Conduct, and Trust in the Laboratory. *Social Philosophy and Policy* 34(1), pp. 25–55.

Wilson, Michael L and Richard W Wrangham (2003). Intergroup relations in chimpanzees. *Annu Rev Anthropol* 32(1), pp. 363–392.

World Bank (2011). *World Development Report 2011: Conflict, Security, and Development*. World Bank.

Wrangham, Richard W and Luke Glowacki (2012). Intergroup aggression in chimpanzees and war in nomadic hunter-gatherers. *Hum Nature* 23(1), pp. 5–29.

Wrangham, Richard W, Michael L Wilson, and Martin N Muller (2006). Comparative rates of violence in chimpanzees and humans. *Primates* 47(1), pp. 14–26.

Yamagishi, Toshio and Toko Kiyonari (2000). The Group as the Container of Generalized Reciprocity. *Soc Psychol Quart* 63(2). contains references to literature on in-group favoritism in 2 person PDs, pp. 116–132.

SUPPLEMENTARY MATERIALS

Appendix A: Complete matching scheme

Period	Group							
	1	2	3	4	5	6	7	8
1	Blue 2 (GR)	Blue 1 (GR)	Green 4 (GR)	Blue 3 (B)	Red 2 (DR)	Blue 4 (B)	Green 1 (IG)	Red 4 (B)
	Red 1 (B)	Yellow 2 (GR)	Brown 4 (B)	Green 3 (GR)	Brown 2 (DR)	Red 3 (DR)	Green 2 (IG)	Yellow 3 (IG)
	Yellow 1 (GR)	Purple 2 (B)	Purple 3 (GR)	Purple 4 (GR)	Purple 1 (B)	Brown 3 (DR)	Brown 1 (B)	Yellow 4 (IG)
2	Green 3 (GR)	Red 3 (B)	Blue 4 (GR)	Blue 2 (GR)	Blue 3 (DR)	Green 2 (DR)	Blue 1 (B)	Red 2 (IG)
	Yellow 1 (B)	Green 1 (GR)	Green 4 (B)	Yellow 4 (GR)	Red 1 (B)	Brown 4 (B)	Brown 1 (IG)	Red 4 (IG)
	Purple 1 (GR)	Purple 3 (GR)	Yellow 2 (GR)	Brown 2 (B)	Yellow 3 (DR)	Purple 2 (DR)	Brown 3 (IG)	Purple 4 (B)
3	Red 1 (GR)	Red 4 (GR)	Blue 3 (B)	Red 3 (GR)	Green 4 (DR)	Blue 2 (DR)	Blue 1 (IG)	Yellow 3 (B)
	Brown 4 (GR)	Yellow 4 (B)	Red 2 (GR)	Green 2 (B)	Yellow 1 (B)	Green 3 (B)	Blue 4 (IG)	Purple 2 (IG)
	Purple 1 (B)	Brown 1 (GR)	Brown 3 (GR)	Brown 2 (GR)	Purple 4 (DR)	Yellow 2 (DR)	Green 1 (B)	Purple 3 (IG)
4	Blue 4 (GR)	Blue 3 (GR)	Green 2 (GR)	Blue 1 (B)	Red 4 (DR)	Blue 2 (B)	Green 3 (IG)	Red 2 (B)
	Red 3 (B)	Yellow 4 (GR)	Brown 2 (B)	Green 1 (GR)	Brown 4 (DR)	Red 1 (DR)	Green 4 (IG)	Yellow 1 (IG)
	Yellow 3 (GR)	Purple 4 (B)	Purple 1 (GR)	Purple 2 (GR)	Purple 3 (B)	Brown 1 (DR)	Brown 3 (B)	Yellow 2 (IG)
5	Green 4 (GR)	Red 4 (B)	Blue 3 (GR)	Blue 1 (GR)	Blue 4 (DR)	Green 1 (DR)	Blue 2 (B)	Red 1 (IG)
	Yellow 2 (B)	Green 2 (GR)	Green 3 (B)	Yellow 3 (GR)	Red 2 (B)	Brown 3 (B)	Brown 2 (IG)	Red 3 (IG)
	Purple 2 (GR)	Purple 4 (GR)	Yellow 1 (GR)	Brown 1 (B)	Yellow 4 (DR)	Purple 1 (DR)	Brown 4 (IG)	Purple 3 (B)
6	Red 2 (GR)	Red 3 (GR)	Blue 4 (B)	Red 4 (GR)	Green 3 (DR)	Blue 1 (DR)	Blue 2 (IG)	Yellow 4 (B)
	Brown 3 (GR)	Yellow 3 (B)	Red 1 (GR)	Green 1 (B)	Yellow 2 (B)	Green 4 (B)	Blue 3 (IG)	Purple 1 (IG)
	Purple 2 (B)	Brown 2 (GR)	Brown 4 (GR)	Brown 1 (GR)	Purple 3 (DR)	Yellow 1 (DR)	Green 2 (B)	Purple 4 (IG)

Appendix B: Allocations in the DR condition

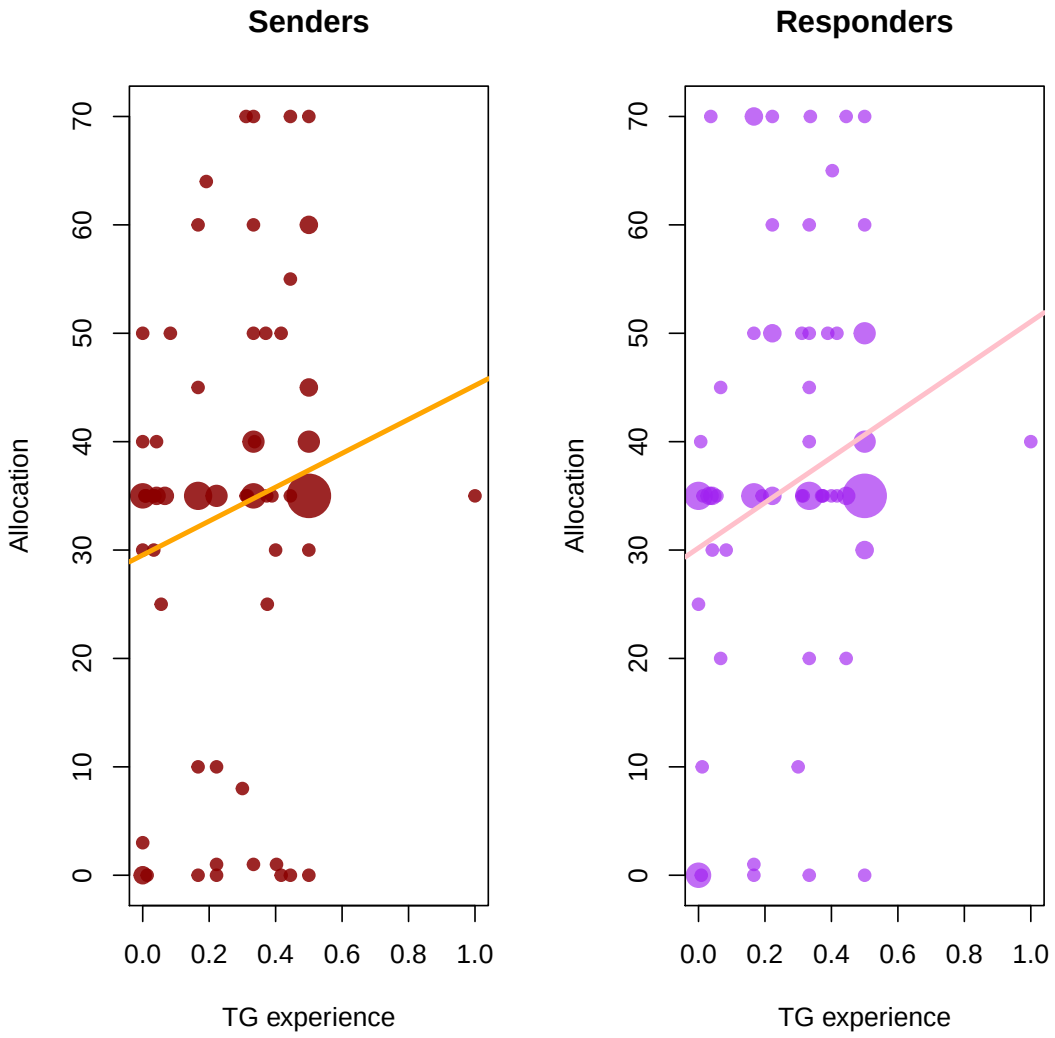


Figure B.1: Allocations in the Direct Reciprocity condition versus the TG experience. Circles show individual data points (circle size proportional to number of observations). Lines show linear regressions.

Appendix C: Experimental instructions

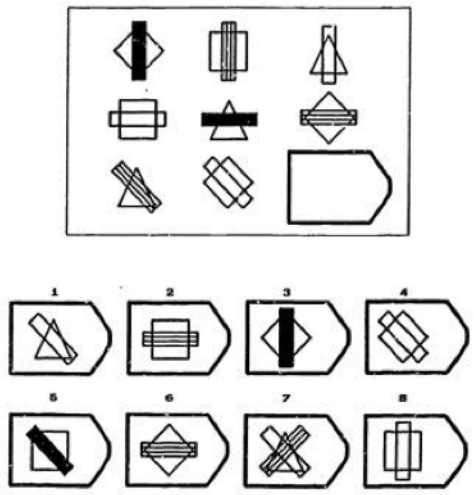
Instructions for the experiment

<Presented as a pdf document and available throughout the experiment>

These instructions are identical to all the participants.
The experiment is composed of five separate and different phases. At the beginning of the experiment, all participants will be allocated into **teams of four**. Each team has a unique **colour**. These teams will remain fixed throughout the experiment.
Before each part, we will distribute and read the relevant instructions for that part. In each part the participants will be reallocated into groups. The number of participants in a group can change from part to part. The payments in the part will be determined according to the decisions of the participants in the team. It is possible, but not necessary, that another participant will be in the same group as you in two different parts. In each part of the experiment you will be able to know which team each of the participants in your group belongs to.
Your final payment in the experiment will be the total of your gain in all of the parts.
At the end of the experiment, you will be presented with the payments in each part and your total payment, in points and in shekels. Please remain seated until the experimenter calls you for payment.
If you have any questions, please raise your hand now and the experimenter will come to you.

Experiments for the first part

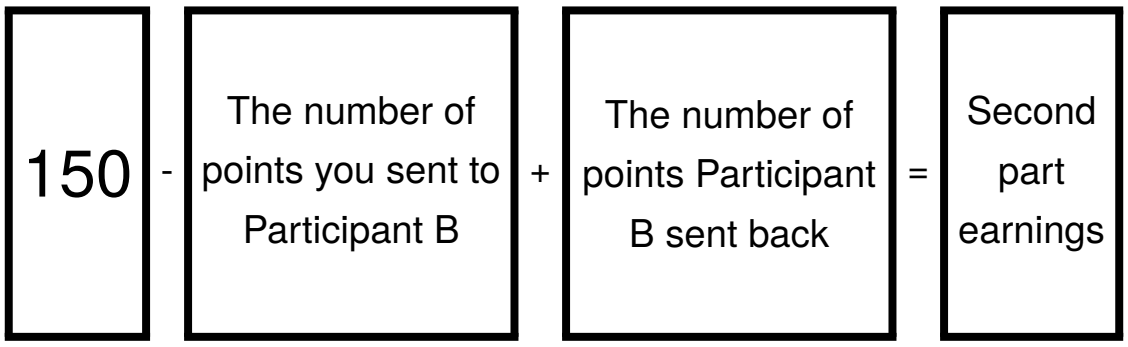
In this part, you and the members of your team perform a pattern completion task. The computer will present you with five questions. Each question is comprised of eight pictures, and the team members wil be asked to choose a ninth picture out of eight possible pictures to complete the pattern. For example:



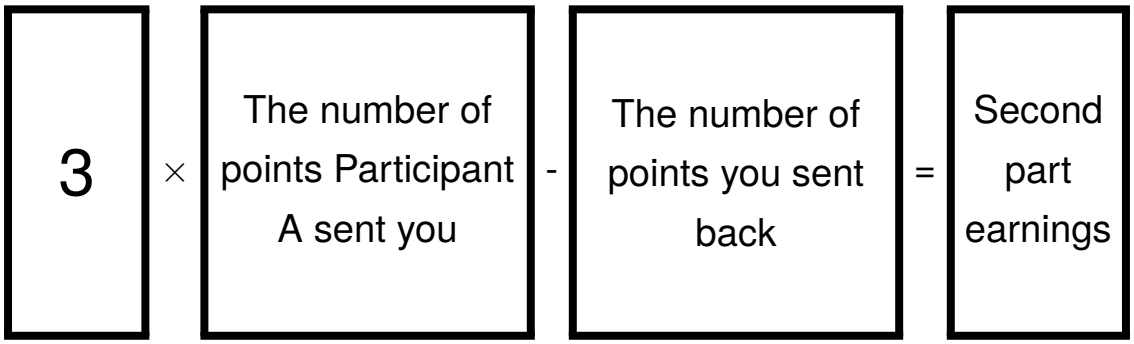
Each team member must answer all of the questions. For each correct answer, the team member will receive **10 points**. Additionally, if all of the team members answer correctly, the whole team will receive a **team bonus of 20 points, to be equally divided among the team members**.
Each question will be allocated two minutes. During this time, the team members can **consult each other** using electronic chat. Enter your answer and click Confirm. You can change your answer and click Confirm again at any point during the two minutes. The last answer to be entered is the final answer.
Attention: Do not reveal any identifying information. If any participant in the session identifies themselves, we will stop the experiment and release all participants with only the showup fee.
If you have any questions, please raise your hand now and the experimenter will come to you.

Instructions for the second part

In this part participants will be matched in **pairs**. In each pair, one participant will be in role A and the other participant in role B. Participant A receives an allocation of **150 points** and decides how many of the 150 points to **send to Participant B**. The amount is **tripled**. Next, Participant B will decide how many points out of the points received to **send back to to Participant A**. These points will not be multiplied.
If you are allocated to role A, your payment in this part will be:



If you are allocated to role B, your payment in this part will be:



Before making your decision, you will be able to test the payment calculation in a **practice phase**, in which you will be able to make decisions as both **Participant A** and as **Participant B**. In this stage, you will enter decisions in both roles, and see the final payments. The practice will repeat five times.
If you have any questions, please raise your hand now and the experimenter will come to you.

Instructions for the third part

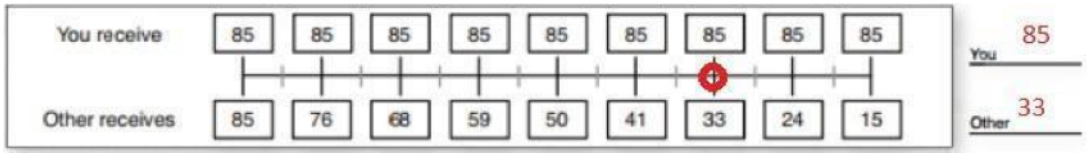
In the third part, all participants will be matched in **groups of three**. Each of the three participants in the group will choose how to **divide 100 points** between the three group members, such that he himself receives **30 points**, and **freely allocates** the remaining **70 points** between the other two group members. This stage has **6 rounds**, and you will be **rematched in a new group**.

Payment calculation in the part

At the end of the experiment, the computer will randomly choose one of the six rounds, and one participant in each group. The payment for this part will be determined according to the decision of the randomly chosen participant in the randomly chosen round.
If you have any questions, please raise your hand now and the experimenter will come to you.

Instructions for the fourth part

In this part, participant will be matched in **pairs**.
Each participant will be presented with **6 rulers** that include nine possible allocations of money to the two participants. The amount you chose to **keep for yourself** is indicated above each ruler, and the amount you choose to **give to the other participant** is indicated below the ruler. You are to choose your preferred allocation of the nine possible allocations. For example,



You can choose any point on the ruler. For example, assume you chose the point marked in red. You will receive 85 points and the other participant will receive 33 points.
At the end of the part, the computer will randomly choose on of the two participants in the pair and one of the nine rulers. your payment in this part will be determined by the decision of the randomly chosen participant for the randomly chosen ruler.
If you have any questions, please raise your hand now and the experimenter will come to you.

Instructions for the fifth part

In this part you will be asked to answer several questions. The questions have to do with the way one sees himself and his surroundings in different situations. Your task is to indicate how much you agree or disagree with each statement, using the following scale:

1. Strongly disagree.
2. Disagree.
3. Neither agree nor disagree.
4. Agree.
5. Strongly agree.

Note: there are no right and wrong answers. Please indicate the answer that best reflects your character with respect to the statement. Take your time and think about your answer.