RESEARCH ARTICLE

REVISED Punishing the individual or the group for norm violation [version 2; peer review: 2 approved]

Marwa El Zein [1,2], Chloe Seikus[3], Lee De-Wit[3,4], Bahador Bahrami [2,5,6]

[1]Institute of Cognitive Neuroscience, University College London, London, WC1N 3AZ, UK
[2]Centre for Adaptive Rationality, Max Planck Centre for Human Development, Berlin, 14195, Germany
[3]Division of Psychology & Language Sciences, University College London, London, WC1H 0AP, UK
[4]Department of Psychology, Cambridge University, Cambridge, CB2 3EB, UK
[5]Faculty of Psychology and Educational Sciences, Ludwig Maximilian University, Munich, 80802, Germany
[6]Department of Psychology, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK

v2 **First published:** 20 Sep 2019, **4**:139
https://doi.org/10.12688/wellcomeopenres.15474.1

Latest published: 13 Feb 2020, **4**:139
https://doi.org/10.12688/wellcomeopenres.15474.2

Abstract

Background: It has recently been proposed that a key motivation for joining groups is the protection from consequences of negative behaviours, such as norm violations. Here we empirically test this claim by investigating whether cooperative decisions and the punishment of associated fairness-based norm violations are different in individuals vs. collectives in economic games.
Methods: In the ultimatum game, participants made or received offers that they could reject at a cost to their outcome, a form of social punishment. In the dictator game with third-party punishment, participants made offers to a receiver while being observed by a punisher, or could themselves punish unfair offers.
Results: Participants made lower offers when making their decision as part of a group as compared to alone. This difference correlated with participants' overall mean offers: those who were generally less generous were even less so in a group, suggesting that the collective structure was compatible with their intention. Participants were slower when punishing vs not punishing an unfair offer. Importantly here, they were slower when deciding whether to punish or not to punish groups as compared to individuals, only when the offer concerned them directly in second party punishment. Participants thus take more time to punish others, and to make their mind on whether to punish or not when facing a group of proposers.
Conclusions: Together, these results show that people behave differently in a group, both in their willingness to share with others and in their punishment of norm violations. This could be explained by the fact that being in a collective structure allows to share responsibility with others, thereby protecting from negative consequences of norm violations.

Keywords

Social punishment, shared responsibility, group decisions, fairness, norm violations, individual differences

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 13 Feb 2020	✓ report	✓ report
	↑	↑
version 1 20 Sep 2019	✗ report	? report

1 **Justin W. Martin**, Boston College, Chestnut Hill, USA

2 **Yarrow Dunham**, Yale University, New Haven, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Marwa El Zein (marwaelzein@gmail.com)

Author roles: El Zein M: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Seikus C**: Conceptualization, Formal Analysis, Investigation, Writing – Review & Editing; **De-Wit L**: Conceptualization, Writing – Review & Editing; **Bahrami B**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: M.E.Z. is supported by the Wellcome Trust (grant number 204702). B.B. was supported by a starting grant from the European Research Council (NEUROCODEC, 309865), the NOMIS Foundation, and the Humboldt Foundation. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819040 - acronym: rid-O).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: El Zein M, Seikus C, De-Wit L and Bahrami B. **Punishing the individual or the group for norm violation [version 2; peer review: 2 approved]** Wellcome Open Research 2020, **4**:139 https://doi.org/10.12688/wellcomeopenres.15474.2

First published: 20 Sep 2019, **4**:139 https://doi.org/10.12688/wellcomeopenres.15474.1

REVISED Amendments from Version 1

We have modified this version to respond to both our reviewers. We gave details in the responses to their reviews how we addressed every comment. We clarified our hypotheses and stated that our experimental design does not allow to characterize the exact mechanisms underlying the differences observed between individual and group. We explained in more details why we believe the effects may be due to shared responsibility in groups, but also suggested alternative explanations. We clarified that we are investigating an individual decision in a collective context rather than assessing the emergence of a collective decision. We explained why we chose the specific individual differences measures, what were our predictions, and we added multiple comparisons corrections in the analyses about the influence of individual differences. We enriched our discussion to draw conclusions that are adequately supported by our results and discuss all the possible alternatives. We have changed Figure 2 and Figure 5 to address the reviewers' comments. Figure 2 was a representation issue (changing the y axis), Figure 5 corresponds to a change in the analysis as requested by the reviewer.

Any further responses from the reviewers can be found at the end of the article

Introduction

To maintain individual and collective welfare, human society relies on formal and informal institutions of justice that enforce norms and punish norm violations. Punishing an individual for norm violations depends on whether they were the agent of that action and responsible for it (Frith, 2014). To be protected against punishment, individuals delegate decisions to others, deferring responsibility and blame for an unfair behaviour (Bartling & Fischbacher, 2012). An alternative way to shift the blame for an unfair choice is to share, rather than delegate, responsibility by making the decision in a group. Research on collective decisions has primarily focused on the benefits of group decisions in terms of outcome improvement, however, neglecting another facet: for an individual, being in a group could be a good way to reduce responsibility and thereby, the associated punishment for norm violation (El Zein *et al.*, 2019). Performing an action as a group distributes the responsibility among group members and also makes it harder to determine who did what. When the group structure is not sufficiently transparent (Duch *et al.*, 2011; Forsyth *et al.*, 2002; Gerstenberg & Lagnado, 2012), it seems likely that the severity of punishment for the collective as compared to the individual will decrease. Therefore, avoiding punishment may represent a strong motivation to join a group decision (El Zein *et al.*, 2019).

Indirect empirical evidence supports this hypothesis that being in a group could help shift the blame and avoid punishments. People are more likely to display free-riding behaviours in groups (Morgan & Tindale, 2002; Tindale & Kameda, 2017; Wildschut *et al.*, 2003), possibly thinking they might get away easier with their act as a group. Also, a group is judged less responsible (Waytz & Young, 2012) and punished less severely (Newheiser *et al.*, 2012) when perceived as a collection of distinct agents (low-cohesive group) than as a unified agent (high-cohesive group).

Here we aimed to *directly* test the hypothesis that norm violations and their punishments differ indecisions made alone or as a contribution to a group decision. Based on the hypothesis that shared responsibility in groups reduces punishment and blame (El Zein *et al.*, 2019) we developed an experimental paradigm to test two key hypotheses: (1) Participants are more likely to violate norms when they are in a group. (2) For the same level of norm violation, groups are less likely (vs individuals) to receive punishment. To do so, we adapted well-known behavioural economic games, which provide valuable experimental paradigms to measure individual's cooperative behaviours and responses to fairness-based norm violations. These games have repeatedly shown that humans cooperate with unrelated strangers in one-off encounters and bear personal costs to punish others who violate norms (Fehr & Fischbacher, 2004). Previous studies have also identified important in-group biases in cooperative norm-enforcement in such games (for a review, see McAuliffe & Dunham, 2016). While the results are sometimes conflicting, an in-group preference is observed in both adults and children, suggesting that belonging to the same group may protect individual group members from punishment during cooperative interactions. Contrary to this line of research, our aim here is not to investigate how different group members interact with each other. We use a context where no group affiliation exists to explore how people behave if they are making cooperative decisions alone or as a part of a neutral group, and whether attitudes to norm-enforcement changes when cooperative decisions come from a neutral player vs a group of neutral players.

To do so, in our adapted versions of the ultimatum game (UG) and the dictator game with third-party punishment (TP-DG), individuals or groups of three individuals could split their allocated points with receivers. The group condition consisted of an average of offers and did not aim to account for an interactive collective decision. Rather, it accounted for individual behaviour in a context where participants were alone vs a context where individual choices contributed to a group average, making the final offer the responsibility of 3 rather than one person. In the UG, the receiver could reject an unfair offer which results in all players receiving zero points. This rejection is considered as a form of social punishment of the proposer and seems to reflect an emotional reaction (Sanfey *et al.*, 2003) and signal of fairness needs (Camerer & Thaler, 1995). In the TP-DG, a third-party can punish an unfair offer at their own cost. Even though unaffected by the norm violation, third parties display this cooperative behaviour which has been suggested to be driven by fairness needs similarly as in second-party punishment (Fehr & Fischbacher, 2004).

We note that while driven by our shared responsibility in groups hypothesis, our experimental design does not allow to characterize the exact mechanisms underlying differences in behaviour between norm violations and their punishment (and thereby be able to affirm that the effects are due to sharing responsibility solely relying on this experiment). However, to our knowledge, this is the first experimental design addressing whether these differences exist with such a controlled design, and directly comparing second and third-party punishment in repeated one-shot trials and a within-participant design.

In addition to these two adapted games, we re-analysed available data from a previous study (Rand *et al.*, 2009) that involved a public goods game between four players with punishment to test whether the use of punishment changes with the number of people defecting. Similarly to playing in a group, we hypothesize that when several people violate norms, they may get away easier with it under the shield that 'others did it too'. This allows to share responsibility for a punishable act, which thereby may become less prone to punishment.

Applying our key two hypotheses described above to the experimental paradigm, we predicted that 1) an individual in the group will offer less than the same individual alone, and make his/her decision faster because of less hesitation about violating the norms within a group, 2) the group will be punished less than the individual and punishment vs no punishment decisions facing a group will be slower reflecting a more hesitant choice, and 3) inflicting punishment on unfair contributions will decrease with the number of people defecting. In an exploratory analysis, we collected self-reported scales to further investigate individual differences in norm violations and their punishment. Social value orientation was measured to link participants' offering and punishment behaviours to a trait measurement of sharing with others (Murphy *et al.*, 2011). A psychopathy scale (Paulhus & Williams, 2002) was collected to test the idea that higher psychopathy scores may be associated with lower influence of being in a group or punishing a group, as people with higher psychopathy scores may care less about being in a group. Finally, political identification was measured in order to test whether differences would appear in fairness attitudes in the context of playing alone or as a group, because liberals and conservatives have been associated with different considerations of fairness and reciprocity (Graham *et al.*, 2009).

Methods

Participants

A total of 150 healthy participants (79 females, mean age= 23.2±4.2) completed the experiment. The eligibility criteria were: 1) participants aged 18–35 and 2) have no reported history of neurological or psychiatric disorders. This sample size was decided based on a previous economic games study that we re-analysed here (Rand *et al.*, 2009). The punishment treatment in the study included 40 participants. Given our 4 conditions of interest (individual or group proposers in the proposer or the punisher role), we multiplied this number by 4 and tested 150 (instead of 160) participants because of practical issues related to timing and participants recruitment. The study took place in November 2017 (first 80 participants – part of a master's thesis of the second author; recruitment postponed for timing issues) and May 2018 (70 participants) at the Psychology Department testing cubicles (26 Bedford way, University College London (UCL). Participants were recruited through the UCL SONA Psychology Pool. It consists of a platform managed by UCL where the experimenter suggests experiment dates that participants receive by email and register to. Participants provided written consent according to regulations approved by the UCL ethics committee (Project ID Number: 4223/002 and ICN-AH-PWB-3-3-2016c). They were informed that they would receive £7.50 for their participation and could receive a bonus up to £2.5 based on their gains. All participants were accorded the bonus and compensated £10.

Experimental design and procedure

Participants were recruited in groups of 7 to 11 individuals with mixed gender. They briefly met each other before entering separate cubicles to begin the experiment. After they completed practice trials, a message instructed them to wait for the experimenters to launch the experiments so that everyone started together. This setting was used to make participants believe they were playing together. The experiment was adapted from two well-known economic games: the Ultimatum Game (UG) and the Dictator game with third-party punishment (TP-DG) (Figure 1).

Ultimatum game (UG). This game includes 2 roles: the proposer and the receiver. In our rendition of the game, a proposer was given 10 points. S/he then decided how to split the 10 points between themselves and a receiver. The receiver, in turn, could accept or reject the offer. If accepted, each player received the points allocated to them by proposer. If rejected, both players received zero points. Rejection of an offer is a costly choice and is explained as a social punishment of the proposer by the receiver.

Dictator game with third-party punishment (TP-DG). This game includes 3 roles: a proposer, a receiver and a third-party punisher. The proposer was initially given 10 points. S/he then decided how much of 10 points she wanted to give the receiver, and how much to keep. The third-party, who had been allocated 5 points, observed the transaction. S/he had the choice to spend one of her points to reduce the proposer's overall outcome by 30%. The third party did not make any material gain from this choice. Reducing the proposer's gain, therefore is a form of costly social punishment as the third-party loses a point in order to punish a player who acted unfairly.

Key experimental conditions. In both games, we added a variation to the main paradigm to include conditions where groups (proposers) make the offers to the receiver. This condition consisted of a group of 3 individuals making a collective offer. Participants were informed that the group offer was an average of individual offers. They were told that punishment of the group offer would reduce each member's pay-off directly and did not consist of a split of points among group members.

For example, in the UG, if the average group offer was 4, each member of the group kept 6 points if the offer was accepted. If the offer was rejected, everyone received zero points. In the TP-DG, if the average group offer was 4, each member of the group kept 6 points if the third-party did not punish them. If punished by third-party, each member of the proposer group received 4 points (i.e. 6 points reduced by 30% and rounded to nearest integer).

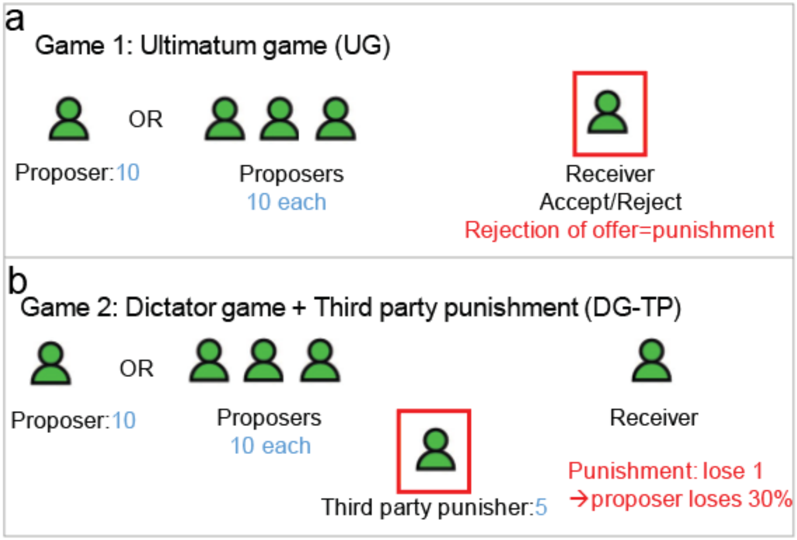

Figure 1. Experimental design. (**a**, **b**) In both games, 1 proposer and a group of 3 proposers had to split points between themselves and a receiver. **a**) In the Ultimatum Game, the receiver could accept or reject the offer in which case no one received any points. **b**) In the dictator game with third party punishment, the receiver could not do anything, but a third-party punisher could punish the proposer(s) by making them lose 30% of their points at their own cost, i.e. losing 1 from their allocated 5 points.

Design. All participants completed 60 one-shot interaction trials in total and played all roles in both games. The trials were played anonymously assuring that participants could not build reputations that might influence their decisions. While participants were told they were playing online together, we computed all the interactions, and everyone did the same experiment (with randomized order of rounds for each participant).

Conditions of interest consisted of 48 trials, where participants played either receiver in UG (24 trials) deciding whether to accept or punish offers or the third-party punisher in TP-DG (24 trials). In these trials, offers were perceived to have been made by three individuals (group condition) on half of the times (12 trials) and by one individual (individual condition) in the other half. The participants did not know that these offers were algorithmically generated so that they ranged from 0 to 5 (each repeated twice within each individual and group condition) and therefore primarily consisting of unfair offers.

Participants also completed 12 trials in which they played the other roles. This included playing the proposer in the UG and TP-DG where they selected a number (out of ten, on the computer keyboard) to offer to the receiver. They played twice as an individual proposer and twice as a group of proposers in both games. They also played the receiver in the TP-DG in which they received an offer but could not respond (twice receiving the offer from an individual proposer and twice from a group of proposers). For these conditions, the other players' choices were computed as follows: The proposers offers were randomly generated numbers between 0 and 6. The decisions to reject offers in UG or punish in TP-DG were based on the participant's offer (or the mean offer with the other two simulated offers): if the offer was between 0 and 4, then there was a 50% chance it will get rejected/punished. If the offer was 5 or more then it was accepted/not punished.

Before starting the experiment, participants completed a practice with one round in each of the condition (five possible roles played in the group and individual condition – ten practice trials total).

Trial structure. At each round, participants first saw which game they were paying for 5 seconds: the image depicted all the possible roles with the role they were assigned to on that round framed with a black rectangle. The points each player had was also reminded at each round. If they were in a group condition, three proposers appeared on the screen.

If they were playing the proposer role:

They were asked: *How much would you like to offer?* They could press a number on the keyboard to make their offer within 4 seconds. A spinner then appeared on the screen for ~5 seconds and it was written: *You offered* (or *you and the 2 other players offered* in the group condition) [amount offered], *the receiver* (UG) or *the punisher* (TP-DG) *is making a choice*. Then they saw what the receiver or punisher decided: 'The receiver accepted' or 'rejected'/ 'The proposer(s) was/were punished'

If they were playing the receiver role in UG or the third-party punisher:

They first saw a spinner for about ~5 sec and it was written: *The proposer is (or the 3 proposers are) making an offer.*

Second, the proposed offer was written: The proposer offered [amount offered].

Third, they were asked:

Would you like to accept the offer? (if receiver in UG) or *Would you like to punish the proposer?* (if punisher in TP-DG) They could press 'Y' for Yes or 'N' for No on the keyboard to give their answer. They had 4 seconds to make their choice.

If they were the receiver in the TP-DG, they observed what was happening, with spinners while proposer(s) made an offer and when the punisher was decide whether to punish or not, and the outcomes of each stage.

At the end of each round, participants were shown the outcomes for each player below the image depicting the player for 5 seconds (for example: The proposers each keep 6 – The receiver gets 4 – The punisher keeps 5)

The exact timeline of each round can be observed by following the link to the online experiment:

https://www.ucl.ac.uk/icn-crowd-cognition/Marwa/gamesexp/rungames.html

Incentives. Participants were told that they would have the chance to win a bonus and receive up to an additional £2.5 on the basis of their outcome in a randomly selected trial at the end of the experiment (with 1 point=0.25pounds). This made sure that every trial counted for towards the participant's earning and helped to make sure that they keep focused in all 60 trials.

Questionnaires. Online questionnaires (using www.qualtrics.com/) were sent to the participants via email and filled out before the day of the experiment. Participants had to respond to these questionnaires in order to be eligible to participate in the experiment; however, they were not selected based on these scales in order to fit different groups. The questionnaires measured social value orientation (SVO) (Murphy *et al.*, 2011), self-reported political identification (POI) (from extreme left to extreme right) and psychopathy traits extracted from The Dark Triad Scale (Paulhus & Williams, 2002). We checked whether these three different scales co-varied with the four dependent variables: mean offers proposed OFF, mean punishment PUN, difference in offer between group and individual OFFDIFF, and difference in punishment given to a group vs individual PUNDIFF, and the associated reaction times (RTs). We also checked the relation between the scales and these variables separately in the UG and TP-DG.

Statistical analyses. Analyses were performed using MATLAB (R2016b). Non-parametric analyses were performed as all data (offers made as proposers, proportion punishment and reaction times) were not normally distributed (Kolmogorov-Smirnov and Shapiro-Wilk tests rejecting the null hypothesis that the data come from a normal distribution). These analyses include the Wilcoxon signed-rank test, Friedman test, Spearman correlations and generalized linear mixed-effects models. Effect size (r) for Wilcoxon tests are reported, calculated as: r = Z/sqrt(N) with N = number of observations.

Re-analysis of data from the public goods game. To investigate whether punishment use decreases with the number of defectors (3rd prediction in the introduction) in the public goods game, we reanalyzed available data from a previous study (Rand *et al.*, 2009) that involved a PGG between four players with punishment (Figure 5a). We tested our third prediction that inflicting punishment on unfair contributions will decrease with the number of people defecting by examining the use of punishment at each played round as a function of the number of people defecting (rather than the number of people giving an offer as we did in our experimental setting). We considered as defectors the players who gave less than half of the maximum amount of contribution at each round (less than 10, maximum amount=20).

We performed a mixed model to test the hypothesis that punishment option (1 if any punishment is used, i.e., punishing 1 or more players, 0 if no punishment) was predicted by number of defectors. The number of defectors at the round, the player's contribution and the group's payoff were entered as fixed-effect predictors of punishment use, and participants were entered as random-effects (40 participants).

Results

Proposer role

Two independent variables could influence the offers made by proposers and reaction times to make the offers: the game (Ultimatum Game UG or Dictator Game with third party punishment DG-TP) and the group condition (Individual proposer IND or group proposer GRO).

Proposer offers. Offers made in the UG correlated with those made in the DG-TP (ρ =0.60, p<0.001) confirming that people who are generous in one game were also generous in the other. Moreover, higher offers were made in UG as compared to DG-TP (Z = 3.86, p < 0.001, r = 0.22).

To test our first hypothesis, that an individual in the group will offer less than the same individual alone, we turn to the effect of group condition on offers. Confirming our hypothesis, a main effect of group condition was observed with higher offers made by participants as individual proposers (IND) as compared to being part of a group of proposers (GRO) (Z = 2.23, p = 0.025, r = 0.12) (Figure 2a). This was also true when considering only the first trial where people made an offer individually and the first trial where they made the offer as a group average (Z=3.24, p=0.001, r=0.18). The difference between IND and GRO did not significantly differ between games (Z = 0.67, p = 0.49, r = 0.03). Interestingly, the offer difference between IND vs GRO correlated negatively with the mean offer made by each participant in all conditions (Spearman correlation

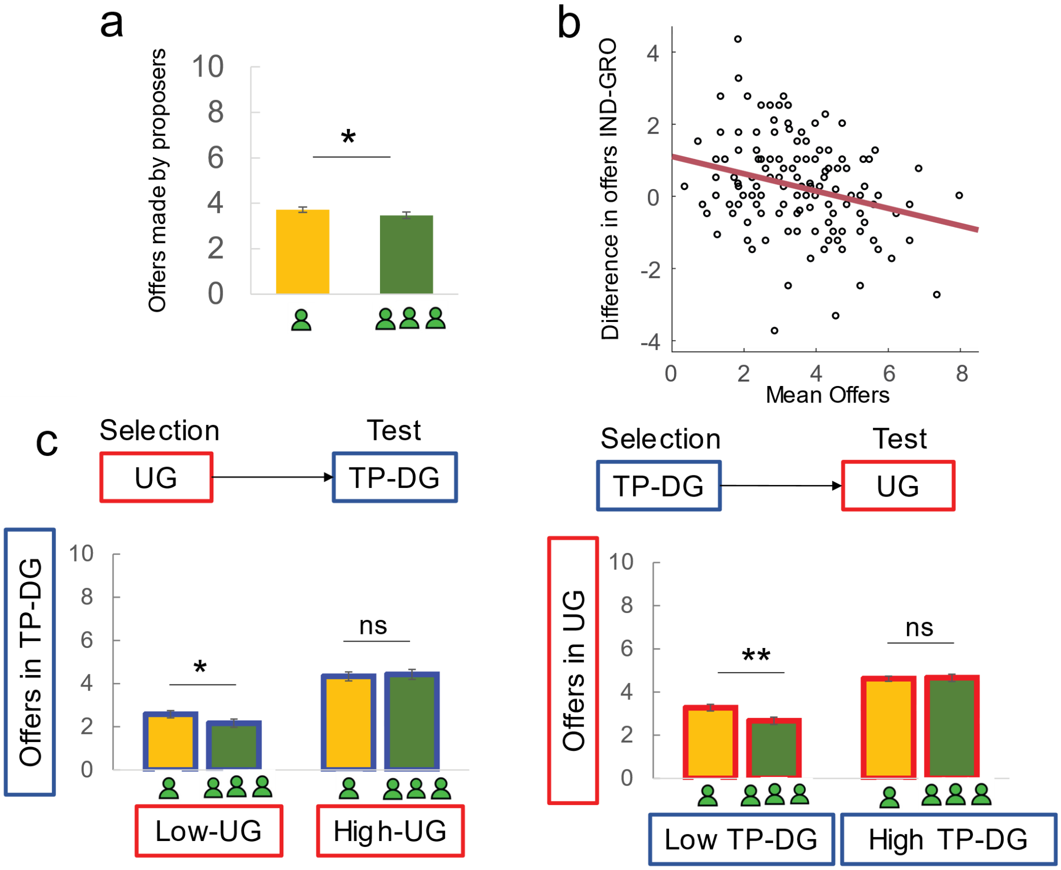

Figure 2. Offers made in the proposer role. (**a**) Mean offers over both games made individually (yellow) or as a group (green). (**b**) Difference in offers made individually or as a group as a function of mean offers over both games. (**c**) Offers in each game as a function of a selection made in the other game: Left panel, offers in the third-party dictator game (TP-DG) separated by those who gave low or high offers in the Ultimatum Game (UG). Right panel, offers in the UG separated by those who gave low or high offers in the TP-DG. ** p<0.01, *p<0.05; ns, non-significant.

ρ = -0.251, p = 0.002) (Figure 2b). This correlation persisted within each game: In the UG, the difference between IND and GRO co-varied with the mean offer in UG (ρ = -0.31, p < 0.001) and the mean offer in TP-DG (ρ = -0.18, p = 0.02). In the TP-DG, the difference between IND and GRO co-varied with the mean offer in TP-DG (ρ = -0.19, p = 0.01) and the mean offer in UG (ρ = -0.16, p = 0.04). In other words, individuals who tended to make overall lower offers (regardless of the game played) diminished their offers even further when in a group, suggesting that the group condition was compatible with the individual's intention to make less generous offers.

To further understand this finding, we categorized people as low proposers and high proposers using a median split in each of the games separately. Using the split based on the UG, we checked whether there was a difference between IND and GRO in low vs high proposers in the TP-DG: A significant effect appeared only for low proposers (Z = =2.04, p = 0.04, r =0.16), but not high proposers (Z =-0.61, p = 0.53, r =0.04, difference between the two types of proposers – low vs high proposer Z=1.9, p=0.05, r=0.15) (Figure 2c). Similarly, using the split based on the TP-DG, we checked whether there was a difference between IND and GRO in low vs high proposers in the UG: A significant effect appeared only for low proposers (Z =3.16, p = 0.001, r =0.25), but not high proposers (Z =-0.23, p = 0.81, r =0.01, difference between the two types of proposers Z=2.83, p=0.004, r=0.23) (Figure 2c).

Reaction times to make offers. To test the second part of our first hypothesis that an individual in the group will make his/her decision faster as compared to the same individual alone, we turn to the differences in reaction times between the conditions. No main effect of game (p=0.18) or group (p=0.45) was observed. But there was an interaction between the two factors (Z = 2.72, p = 0.006, r=0.16): Reaction times were faster for decisions within a group as compared to individually only in the TP-DG (Z = -2.65, p = 0.008, r=0.15) and not in the UG (Z = -1.47, p = 0.14, r=0.08). The second part of our first hypothesis was confirmed, but only when third-party and not second party punishment was involved.

To conclude on the proposer role, participants gave lower offers as a group vs alone, and were faster to do so in the dictator game. This may suggest that participants were expecting less punishment when playing in a group as compared alone, which leads to our next question: Are groups punished less than individuals for the same norm violation and is the decision to punish or not punish a group as compared to one individual less intuitive?

Punishment

Three independent variables could influence punishment: the amount of offers proposed (0 to 5), the game (UG or DG-TP) and the group condition (individual proposer IND or group proposer GRO).

Proportion punishment. Proportion punishment in the UG correlated with proportion punishment in the DG-TP (ρ = 0.64, p < 0.001), suggesting that participants show a consistent pattern of punishment in different contexts, here second-party and third-party punishment. Proportion punishment also correlated with the amount of offers in the proposer role (ρ = 0.44, p < 0.001), showing that those who were more generous as proposers were also more prone to punish smaller offers, thereby caring about fairness both in their offers and in their punishment behavior (Figure 3b). Similarly as for the proposed offers, there was more overall punishment in the UG than in the TP-DG (Z = 2.46, p = 0.014, r=0.14), consistent with a previous experiment directly comparing second and third-party punishment (Fehr & Fischbacher, 2004).

A main effect of offers was observed, with punishment increasing as the offers decreased (Friedman test χ^2 =517.18, p < 0.001) (Figure 3a). Contrary to our second hypothesis however, there was no main effect of group condition on proportion punishment (Z = 0.37, p = 0.7, r = 0.02) and no interaction between group and game.

Reaction times for punishment decision. Confirming the second part of our second hypothesis, participants were slowed down to make their punishment vs no punishment decisions when facing a group of proposers as compared to an individual proposer (Z = 3.33, p=0.001, r=0.19). They were also slower to respond in the TP-DG as compared to the UG (Z = -9.33, p < 0.001, r=0.53). An interaction was observed between these group condition and games: the difference between reaction times for individuals vs groups was more important in the UG as compared to the TP-DG (Z = 1.96, p=0.04, r=0.11), with a significant difference between decision time for GRO vs IND only in UG (Z = 4.19, p < 0.001, r=0.24) and not in

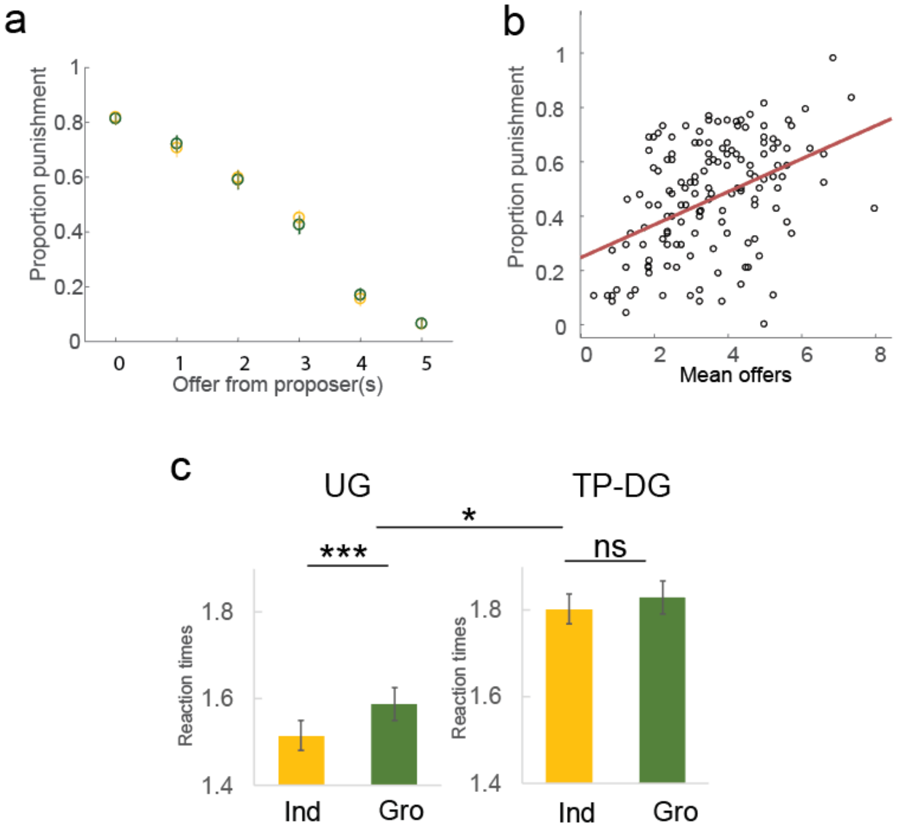

Figure 3. Punishment decisions. (**a**) Proportion punishment as a function of the amount of offers proposed, green= for punishment of group, yellow= for punishment of individual. (**b**) Proportion punishment as a function of mean offers. (**c**) Reaction times for punishment vs no punishment decisions separated for the individual (yellow) and group (green) condition in the Ultimatum game (left panel, UG) and the third party punishment dictator game (right panel, TP-DG). *** p<0.001, * p<0.05; ns, non-significant.

TP-DG (Z = 1.22, p = 0.22, r=0.07) (Figure 3c). This shows that participants slowed down to make a decision when receiving an offer from a group vs an individual, only when they were directly receiving the offer. The slowing down to make a decision when facing a group vs individual proposer was true even for very low offers, i.e., 0 and 1 (Z = 1.98, p = 0.04, r=0.11), excluding the interpretation that slowing down is due to avoiding the punishment of fair participants trapped in a group with unfair partners as for offers of 0 and 1, every member of the group surely offered low amounts. Moreover, the difference in reaction times for IND vs GRO was only significant for punishment vs no punishment decisions of unfair offers (30% or less (Sanfey *et al.*, 2003), so 0 to 3 here) and not fair offers (fair Z=1.1, p = 0.27, r=0.06; unfair Z= 3.43, p<0.001, r=0.2, difference Z=1.39, p=0.16, r=0.08). Participants were thus slowed down when it comes to punishing groups vs individuals who violated fairness norms.

Reaction times as a function of punishment or no punishment decision. Participants were slower to punish as compared to not punish in both UG and TP-DG (Z = 4.02, p <0.001, r=0.23). This did not interact with the main effect of group on reaction times (Z=0.46, p=0.64, r=0.02). An interaction between the choice to punish or not to punish and the amount of offers proposed was observed: When the decision was to 'not punish', reaction times were slower for low (0, 1, 2) as compared to high (3, 4, 5) offers (Z = 4.74, p<0.001, r=0.27). When participants chose to 'punish', the reverse was observed as choices were faster for low offers vs high offers (Z = -3.07, p = 0.002, r=0.17).

Individual differences

We accounted for the effects of all three scales on the different variables by entering them as predictors (SVO, psychopathy and political identification) of these variables in a generalized mixed model. This involved running 16 GLMs (8 variables OFF, PUN, OFFDIFF and PUNDIFF and associated RTs) X 2 games (UG and TP-DG). Although the 3 predictors were simultaneously entered in all GLMs, we present the results separately for each predictor for clarity. We provide the p values corrected for multiple comparison by multiplying the p values by the number of performed GLMs (16).

SVO. SVO separates individuals in competitive, individualistic, prosocial, and altruistic profiles. On the total of 150 participants, 44 as individualistic, 105 scored as prosocial, and 1 as altruistic (Individualistic, prosocial and altruistic profiles were entered as 1, 2, 3 respectively in the GLMs). The amount of offers OFF was predicted by SVO (UG z = 4.38, p<0.001, corrected p<.01; TP-DG z=4.93, p<0.001, corrected p<.01). Punishment PUN was also predicted by SVO (UG z = 1.98 p=0.04, corrected p=0.64; TP-DG z = 2.76, p = 0.006, corrected p=0.096). Indeed, prosocials, compared to individualistics, gave higher offers as proposers (over both games Z=4.42, p<0.001, r=0.36; UG Z=3.66, p<0.001, r=0.29; TP-DG Z=4.32, p<0.001, r=0.35) and punished more as second and third-party punishers (over both games Z=2.08, p=0.03, r=0.16; UG Z=1.39, p=0.16, r=0.11; TP-DG Z=2.2, p=0.02, r=0.17) (Figure 4a).

To conclude on SVO, people who have a generous trait gave higher offers in cooperative games. They also punished more unfair offers, however this last result was not robust enough to survive multiple comparisons.

Psychopathy. PUN was predicted by psychopathy scores in the UG only (z = 2.37 p = 0.01, corrected p=0.16). Indeed, when participants were split into 3 (based on second and third quantile): high, moderate and low psychopathy, high psychopathy participants punished significantly more than Low psychopathy participants in the UG (Z=2.77, p=0.005, r=0.27 Figure 4b) and not in the TP-DG (Z = 0.41, p = 0.68, r=0.04).

Also, in the UG, the difference in punishment between groups and individuals PUNDIFF was predicted by psychopathy (z= -2.15, p = 0.03, corrected p=0.48):PUNDIFF significantly differed between Low and High psychopathy participants in the UG only (over both games Z=2.59, p = 0.009, r=0.25; UG Z = 2.78, p = 0.005, r=0.27; TP-DG Z=0.19, p=0.84, r=0.01) (Figure 4b): In Low psychopathy, there was a higher proportion punishment of individuals as compared to groups (Z = 2.07, p = 0.03, r=0.20, UG Z=1.69, p=0.08, r=0.16; TP-DG Z=0.48, p=0.62, r=0.04). In High Psychopathy, there was no difference overall (Z=-1.27, p=0.2, r=0.12), but when only the UG was considered, it seemed like individuals were actually punished even less than groups (Z=-1.75, p=0.07, r=0.17).

The results thus show that high psychopathy participants rejected more offers overall and tend to do so more from groups than individuals. On the contrary, low psychopathy participants seem to reject more offers coming from individual as compared to group proposers. Again, these results are interesting but are to be taken with caution as the regressions results do not survive multiple comparison correction.

Political identification. Political identification was measured on a scale from 1 to 7 from Strongly Liberal to Strongly Conservative (4=Neutral, entered this way in GLMs). In total, 80 participants identified as liberals, 19 as conservative and 51 as moderate. Difference in RT for punishing groups vs individuals was predicted by POI (z = 2.81, p = 0.005, corrected p=0.08).

Indeed, the observed slowing down for punishing groups was more important in liberals than in conservatives (UG Z=2.9, p = 0.003, r=0.29 TP-DG Z=-0.76, p=0.44, r=0.07). The difference in reaction times between punishing groups and individuals was the strongest in liberal participants (Z =-4.61, p<0.001, r=0.36; moderate participants Z=-1.84, p=0.06, r=0.18, conservatives Z=1.0, p=0.3, r=0.16) (Figure 4c).

Reanalysis of a public good game: Punishment as a function of the number of defectors

The results of our study show that only when participants are directly concerned by an offer, the number of people giving that offer influenced punishment behaviour: there is a consistent slowing down to make the decision of whether or not to punish three individuals as compared to one individual. We did not however find an effect on proportion punishment.

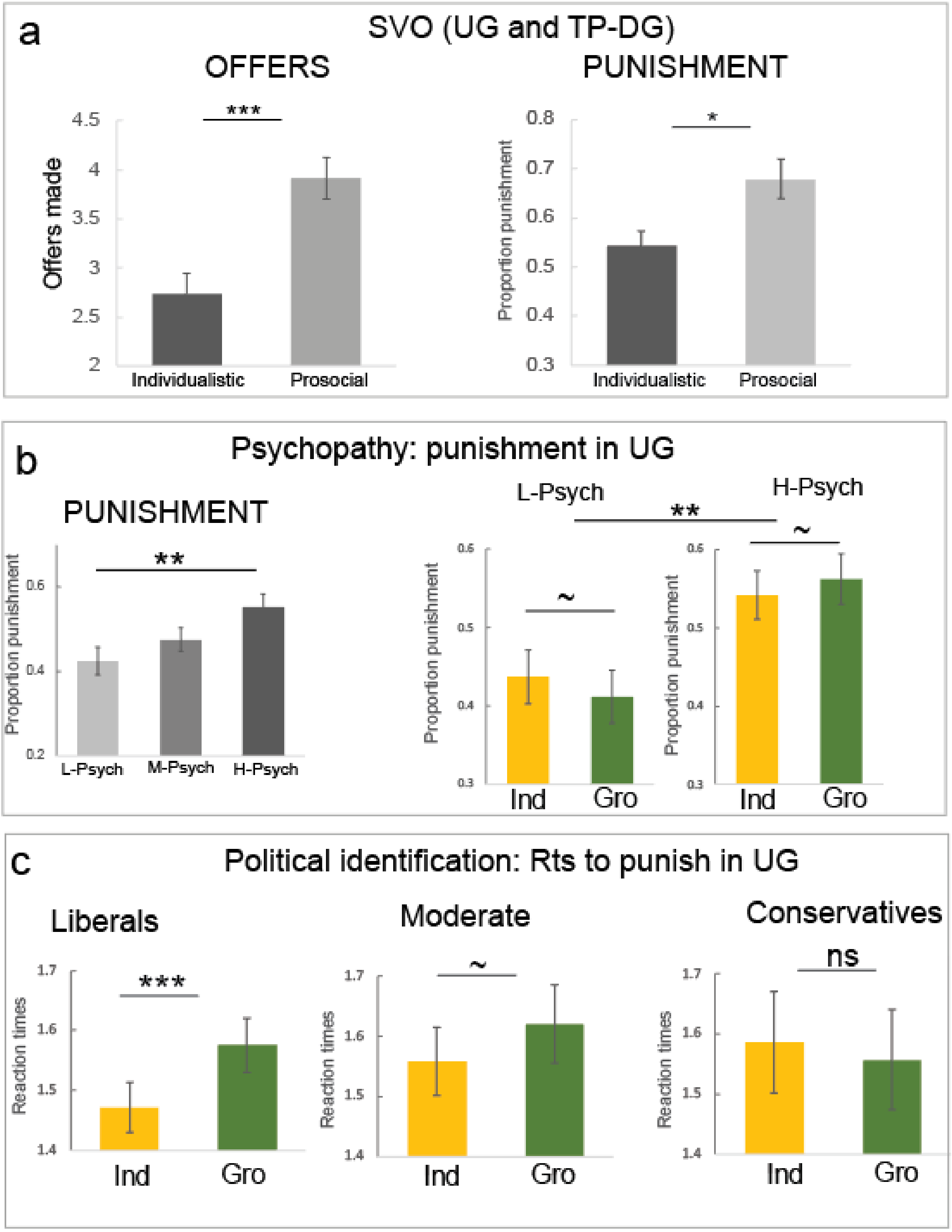

Figure 4. Individual differences. a) Social value orientation influence on offers (Left panel) and proportion punishment (Right panel). **b**) Psychopathy influence on proportion punishment in the ultimatum game. L-Psych: Low Psychopathy, M-Psych: Moderate psychopathy, H-Psych: High Psychopathy. An interaction was observed between Low and High psychopathy and the difference between individual and group punishment: For L-Psych, individuals are punished more than groups while for H-Psych, groups are punished more than individuals **c**) Political identification influence on reaction times for the punishment decision in the ultimatum game. Reaction times when faced with an individual or a group in liberals, moderate and conservatives., ~ p<0.06, *** p<0.001; ns, non-significant.

The UG involved punishment by rejecting an offer, the TP-DG involved a second step punishment that did not concern the third-party directly. A game that combines these 2 types punishment is the public goods game (PGG) with punishment, in which people can punish those who defect to a common good. In that case, people are directly concerned as they receive money from the common good (like in the UG) and they can decide to make a costly punishment at a second stage (like in the TP-DG).

To investigate what happens in such a context, we reanalyzed available data from a previous study (Rand *et al.*, 2009) that involved a PGG between 4 players with punishment (Figure 5a). We tested our third prediction that inflicting punishment on unfair contributions will decrease with the number of people defecting. The reasoning here is that the more defectors on a given round, i.e., the more people who violate the norm, the more their behaviour can be justified (they are not the only ones!) and can therefore benefit from reduced punishment. The number of defectors (from 1 to 4) decreased the probability of using punishment (Estimate=-0.504±0.22, Z=-2.20 p=0.02, no=516), even when accounting for the group's payoff and the players' contribution (Figure 5b).

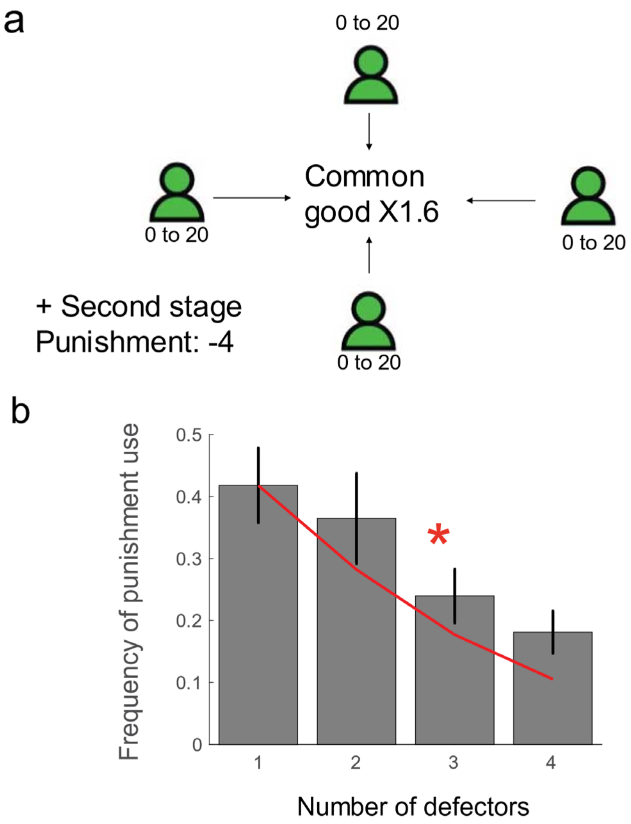

Figure 5. Re-analysis of a public good game with punishment. (**a**) Structure of the public good games in Rand *et al.*, 2009: players can contribute to a common good from 0 to 20. The common good is multiplied by 1.6 and redistributed to all players. In a second stage, participants can punish others for their contributions by -12 at their own cost of losing 4. (**b**) The frequency of using punishment as a function of the number of defectors. *p<0.05 significant decrease in the frequency of punishment use with the number of defectors in the mixed model.

Discussion

In this paper, we investigated whether norm violations and their punishments differ when made alone or as a group. We predicted that being in a group can shift the blame and punishment away from the individual because of shared responsibility for norm violations in a group. Our results confirmed our prediction in three ways: 1) Participants gave less generous offers (violated more the norm) when playing alone vs in a group of three. They were also faster to do so in the TP-DG. 2) Punishing a group vs an individual for norm violations required more time as participants were slowed down to make the punishment vs non-punishment decision. This was the case only in the UG with second-party punishment, when offers directly concerned the punisher. 3) Participants were less inclined to punish others for norm violations when the number of people committing these norms violations was high.

Less generous offers in the group

Our current finding that people are less generous in a group corroborates the idea that people in groups violate the norms more than when alone given that (1) it replicates previous studies showing that individuals in groups display free-riding behaviours (Morgan & Tindale, 2002; Tindale & Kameda, 2017; Wildschut *et al.*, 2003). These previous studies compared groups facing groups to individuals facing individuals and showed that groups are more competitive (Wildschut *et al.*, 2003), defect more in a prisoner dilemma game (Morgan & Tindale, 2002) and offer less in a joint decision in an ultimatum game (Bornstein & Yaniv, 1998). Our results complement these studies by showing that even when facing one individual, people are less generous if they are part of a group vs alone. (2) Interestingly, here, we show for the first time that this decreased generosity in group correlates with people's overall generosity. Indeed, only those who gave low offers displayed a difference between playing in a group or alone. This shows that the group was compatible with the intention of those who were less sensitive to the norms and violated them more. (3) It has been suggested that increased defection in groups may relate to reduced 'identifiability' as a group, supporting the idea that people feel less 'accountable' when making selfish choices (Kugler *et al.*, 2012). Backing this, the fact that groups of two members that could be easily identified playing a dictator game gave higher offers as a group (Cason & Mui, 1997) was linked to a reduced anonymity in such a context (Luhan *et al.*, 2009). On the contrary, when anonymity is preserved in computer-based rather than face-to-face interactions, group members defect more than individuals. (4) Finally, groups that are procedurally interdependent (Wildschut *et al.*, 2001) are less cooperative. This may be due to the fact that individual choices could not have been traced back, again allowing for anonymity and hiding behind the group (Kugler *et al.*, 2012). Despite substantial evidence that the effects observed in our study may be due to increased norm violation in groups because of shared responsibility, our experimental design does not allow us to disentangle the exact mechanism(s) underlying the observed reduction of offers in groups, and other explanations may as well be possible. For example, people may become more rational (Bornstein & Yaniv, 1998), and less biased as a group. They also may feel supported enough to act in a selfish, i.e. more greedy way (Schopler *et al.*, 1995). In addition, one may argue that in our study, participants may have been basing their decision on the averaging procedure, by adjusting their offers to the expectations that others in the group will do. For example, a person who wants to give a low offer may lower their offer to offset the other group members offers. However, this last explanation is not consistent with the fact that only low offerors showed a difference between playing alone or in group as one would also expect those who want to give high offers to then adjust their behaviour by giving higher offers in groups (while they don't). Finally, another possible explanation is that participants lowered their offers because of their expectation that other group members will also give low offers, given that predominantly unfair offers were presented in the context of the experiment (when participants were playing the receiver in the UG and the punisher in the TP-DG). However, the offer decrease for groups was already significant in the first round played in group as compared to the first round played individually, when they had less time to learn about the context they were playing in.

Amount of punishment for groups vs individuals

Previous studies using economic games investigating punishment behaviours in groups have looked at how a group vs an individual punishes norm violations. They showed that when acting as third-party punishers in groups or alone, groups punish less severely in response to norm violations because of the diffusion of responsibility (Feng *et al.*, 2016). In the present study, we examined how a group vs an individual is punished for norm violations rather than how the group punishes others. Contrary to our prediction that shared responsibility will also decrease the punishment of a group, we did not find any difference between the punishments of norm violations made by an individual vs a group. However, and in line with our prediction in our re-analysis of the public goods game, we did find evidence for decreased punishment with the numbers of defectors violating norms. Previous work has shown that a group is judged less responsible (Waytz & Young, 2012) and punished less severely (Newheiser *et al.*, 2012) when perceived as a collection of distinct agents (low-cohesive group) than as a unified agent (high-cohesive group). An explanation to the discrepancy in results could therefore be that in the public goods game, other players were perceived as a collection of individuals. On the contrary, in the current adapted version of the UG and TP-DG the group was possibly perceived as an entity as participants always saw the three group members when faced with the group and told that they can punish 'the group' rather than an individual in the group. Another possible reason why we did not observe decreased punishment for groups vs individuals in the context of our experimental design is linked to the effectiveness of the individual vs group punishment. Indeed, in the group condition, punishing the group imposed three times the total cost imposed when punishing in the individual condition. This implies that punishing a group was more effective than punishing an individual and one could then expect that the group should be punished more than the individual. This could have then overridden a decreased punishment of the group because of shared responsibility for low offers, possibly explaining why we did not observe any difference in the punishment of groups vs individuals.

Decision time for punishing vs not punishing

We found that people were slowed down to punish as compared to not punish others for their norm violations. This suggests that punishing is less intuitive than not punishing. It relates to a series of discussions on whether the selfish (here not punishing) or the cooperative option (here punishing) is less of the default option for people. While some studies suggest that as observed here, it is less intuitive to choose the cooperative vs the selfish option (Krajbich *et al.*, 2015), others suggest the opposite (Rand *et al.*, 2012). These discussions were related to amounts of contributions in economic games (cooperative as high contributions and selfish as low contributions). Here we extend the discussion to punishment decisions, and show that in the context of a UG and TP-DG, people are slower to choose the punishment (and more cooperative) option. We importantly found that the punishment vs no punishment decisions were slower when punishers were faced with groups vs individuals, suggesting that it is also less intuitive to choose whether to punish or not to punish a group. This is in line with previous findings in an ultimatum game showing that participants spent less time considering whether to punish or not offers from opposite race as compared to same race (Kubota *et al.*, 2013). Possibly, being faced with an individual vs group also made decisions faster because of a lower group affiliation when facing an individual vs a group. It is important to note that this effect was only present in second-party and not third-party punishment, suggesting that it applies only if unfairness is directed toward the self. Participants generally showed more punishment in second-party vs third-party punishment, reflecting a higher emotional response when being directly involved which may entail stronger inequity aversion and a higher need for fairness signalling (Fehr & Fischbacher, 2004; Nowak *et al.*, 2000). This higher emotional involvement could also explain why the sensitivity to the group was higher in second-party vs third-party punishment.

Social and antisocial punishment

The amount of punishment was predicted by both social value orientation and psychopathy scale in the ultimatum game (although should be taken with caution as the regression results did not survive multiple comparison corrections, but post-hoc analyses showed significantly higher punishment in both prosocials and high psychopathy participants). This could at first glance seem contradictory. Punishment consists of a cooperative option as it incurs a cost on the punisher, which explains why prosocial, compared to individualistic participants (as assessed in the social value orientation test), showed higher punishment rates in both second and third-party punishment. Interestingly, only in second-party punishment, proportion punishment also increased with the psychopathy scale. In the ultimatum game, punishment decisions have been associated with emotional reactions associated with anger (Pillutla & Murnighan, 1996). Higher punishment in higher psychopathy participants could thus be associated to increased emotional reaction and an antisocial rather than prosocial reaction. This is line with the suggestion that in second-party, not third-party punishment, the decision to punish need not to reflect only cooperative behaviours but can also be associated with anti-social spiteful motives (Jensen, 2010). Accordingly, our results also show that higher psychopathy is associated with higher punishment of the group vs the individual, while on the contrary the group benefited from lower punishment by low psychopathy participants, as initially predicted by our shared responsibility hypothesis (El Zein *et al.*, 2019).

To conclude, using cooperation economic games, we show that people's attitudes related to norm violations are influenced by whether they were made by an individual or a group. People are less generous as a group, use less punishment when more people defect the norms, and take more time to punish a group vs an individual who behaves unfairly to them. Together, these results support the idea that being part of a group may protect one from punishments and norm violations, possibly because of shared responsibility among group members for the same acts that can reduce blame and punishments (El Zein & Bahrami, 2019; El Zein *et al.*, 2019).

Data availability

Open Science Framework: Supplemental materials for preprint: Punishing the individual or the group for norm violation. https://doi.org/10.17605/OSF.IO/HPVBG (El Zein, 2019).

This project contains the following underlying data:

- Data_punishment.csv (data for each task performed by each participant; a data dictionary is available in the Description).

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

References

Bartling B, Fischbacher U: **Shifting the Blame: On Delegation and Responsibility.** *Rev Econ Stud.* 2012; **79**(1): 67–87.
Publisher Full Text

Bornstein G, Yaniv I: **Individual and Group Behavior in the Ultimatum Game: Are Groups More "Rational" Players?** *Exp Econ.* 1998; **1**(1): 101–108.
Publisher Full Text

Camerer C, Thaler R: **Anomalies: Ultimatums, Dictators and Manners.** *J Econ Perspect.* 1995; **9**(2): 209–219.
Publisher Full Text

Cason T, Mui V: **A Laboratory Study of Group Polarisation in the Team Dictator Game.** *Econ J.* Royal Economic Society, 1997; **107**(444): 1465–1483.
Publisher Full Text

Duch R, Stevenson R, Przepiorka W: **Responsibility Attribution for Collective Decision Makers.** *Am J Polit Sci.* 2011; **59**(2): 372–389.
Publisher Full Text

El Zein M: **Supplemental materials for preprint: Punishing the individual or the group for norm violation**. 2019.
http://www.doi.org/10.17605/OSF.IO/HPVBG

El Zein M, Bahrami B: **Collective decisions divert regret and responsibility away from the individual**. 2019.
Publisher Full Text

El Zein M, Bahrami B, Hertwig R: **Shared responsibility in collective decisions.** *Nat Hum Behav.* 2019; **3**(6): 554–559.
PubMed Abstract | Publisher Full Text | Free Full Text

Fehr E, Fischbacher U: **Third-party punishment and social norms.** *Evol Hum Behav.* 2004; **25**(2): 63–87.
Publisher Full Text

Feng C, Deshpande G, Liu C, *et al.*: **Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study.** *Hum Brain Mapp.* 2016; **37**(2): 663–677.
PubMed Abstract | Publisher Full Text

Forsyth DR, Zyzniewski LE, Giammanco CA: **Responsibility Diffusion in Cooperative Collectives.** *Pers Soc Psychol Bull.* 2002; **28**(1): 54–65.
Publisher Full Text

Frith CD: **Action, agency and responsibility.** *Neuropsychologia.* 2014; **55**: 137–142.
PubMed Abstract | Publisher Full Text

Gerstenberg T, Lagnado DA: **When contributions make a difference: explaining order effects in responsibility attribution.** *Psychon Bull Rev.* 2012; **19**(4): 729–736.
PubMed Abstract | Publisher Full Text

Graham J, Haidt J, Nosek BA: **Liberals and conservatives rely on different sets of moral foundations.** *J Pers Soc Psychol.* 2009; **96**(5): 1029–1046.
PubMed Abstract | Publisher Full Text

Jensen K: **Punishment and spite, the dark side of cooperation.** *Philos Trans R Soc Lond B Biol Sci.* 2010; **365**(1553): 2635–2650.
PubMed Abstract | Publisher Full Text | Free Full Text

Krajbich I, Bartling B, Hare T, *et al.*: **Rethinking fast and slow based on a critique of reaction-time reverse inference.** *Nat Commun.* 2015; **6**: 7455.
PubMed Abstract | Publisher Full Text | Free Full Text

Kubota JT, Li J, Bar-David E, *et al.*: **The price of racial bias: intergroup negotiations in the ultimatum game.** *Psychol Sci.* 2013; **24**(12): 2498–2504.
PubMed Abstract | Publisher Full Text | Free Full Text

Kugler T, Kausel EE, Kocher MG: **Are groups more rational than individuals? A review of interactive decision making in groups.** *Wiley Interdiscip Rev Cogn Sci.* 2012; **3**(4): 471–482.
PubMed Abstract | Publisher Full Text

Luhan WJ, Kocher MG, Sutter M: **Group polarization in the team dictator game reconsidered.** *Exp Econ.* 2009; **12**(1): 26–41.
Publisher Full Text

McAuliffe K, Dunham Y: **Group bias in cooperative norm enforcement.** *Philos Trans R Soc Lond B Biol Sci.* 2016; **371**(1686): 20150073.
PubMed Abstract | Publisher Full Text | Free Full Text

Morgan PM, Tindale RS: **Group vs Individual Performance in Mixed-Motive Situations: Exploring an Inconsistency.** *Organ Behav Hum Decis Process.* 2002; **87**(1): 44–65.
Publisher Full Text

Murphy RO, Ackermann KA, Handgraaf M: **Measuring Social Value Orientation.** *Judgment and Decision Making.* (SSRN Scholarly Paper No. ID 1804189), 2011; **6**(8): 771–781.
Publisher Full Text

Newheiser AK, Sawaoka T, Dovidio JF: **Why do we punish groups? High entitativity promotes moral suspicion.** *J Exp Soc Psychol.* 2012; **48**(4): 931–936.
Publisher Full Text

Nowak MA, Page KM, Sigmund K: **Fairness versus reason in the ultimatum game.** *Science.* 2000; **289**(5485): 1773–1775.
PubMed Abstract | Publisher Full Text

Paulhus DL, Williams KM: **The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy.** *J Res Pers.* 2002; **36**(6): 556–563.
Publisher Full Text

Pillutla MM, Murnighan JK: **Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers.** *Organ Behav Hum Decis Process.* 1996; **68**(3): 208–224.
Publisher Full Text

Rand DG, Dreber A, Ellingsen T, *et al.*: **Positive interactions promote public cooperation.** *Science.* 2009; **325**(5945): 1272–1275.
PubMed Abstract | Publisher Full Text | Free Full Text

Rand DG, Greene JD, Nowak MA: **Spontaneous giving and calculated greed.**

Nature. 2012; **489**(7416): 427–430.
PubMed Abstract | **Publisher Full Text**

Sanfey AG, Rilling JK, Aronson JA, *et al.*: **The neural basis of economic decision-making in the Ultimatum Game.** *Science.* 2003; **300**(5626): 1755–1758.
PubMed Abstract | **Publisher Full Text**

Schopler J, Insko CA, Drigotas SM, *et al.*: **The role of identifiability in the reduction of interindividual-intergroup discontinuity.** *J Exp Soc Psychol.* 1995; **31**(6): 553–574.
Publisher Full Text

Tindale RS, Kameda T: **Group decision-making from an evolutionary/adaptationist perspective.** *Group Process Intergroup Relat.* 2017; **20**(5): 669–680.
Publisher Full Text

Waytz A, Young L: **The group-member mind trade-off: attributing mind to groups versus group members.** *Psychol Sci.* 2012; **23**(1): 77–85.
PubMed Abstract | **Publisher Full Text**

Wildschut T, Lodewijkx HFM, Insko CA: **Toward a reconciliation of diverging perspectives on interindividual-intergroup discontinuity: The role of procedural interdependence.** *J Exp Soc Psychol.* 2001; **37**(4): 273–285.
Publisher Full Text

Wildschut T, Pinter B, Vevea JL, *et al.*: **Beyond the group mind: a quantitative review of the interindividual-intergroup discontinuity effect.** *Psychol Bull.* 2003; **129**(5): 698–722.
PubMed Abstract | **Publisher Full Text**

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 06 March 2020

https://doi.org/10.21956/wellcomeopenres.17253.r37911

✓

Justin W. Martin
Department of Psychology, Boston College, Chestnut Hill, MA, USA

I thank the authors for their detailed responses to the concerns/issues I raised about the first version of this manuscript. At this point, I have no further reservations about this work.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Punishment, cooperation, moral judgment.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 20 February 2020

https://doi.org/10.21956/wellcomeopenres.17253.r37910

✓

Yarrow Dunham
Department of Psychology, Yale University, New Haven, CT, USA

The authors have satisfactorily responded to my comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Social cognition.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 26 November 2019

https://doi.org/10.21956/wellcomeopenres.16925.r37132

?

Yarrow Dunham
Department of Psychology, Yale University, New Haven, CT, USA

An interesting take on the role of individual decisions made in the context of individuals or the contexts of a collective. I had several comments that I hope will help the authors as they go forward in this work.

I don’t think the decision that individuals make in the collective condition is really a “collective decision” in that the individual decisions are just summed; something like a voting procedure to determine the decision would feel collective; this feels more like an individual decision in a collective context. I think making this distinction clearer would be useful.

The introduction might want to do an earlier and clearer job distinguishing the present work from the literature on punishment of individuals who are identified as members of groups, i.e. on ingroup bias in punishment decisions.

One interesting point that distinguishes the two conditions in these studies is that punishment or rejection is more effective in the group contexts because it imposes 3x the total cost; what impact, if any, do the authors think this could have?

Th offers that were responded to ranged for 0 to 5 of 10, each repeated twice, in each game. As the authors note this means they were mostly unfair offers. Could this have affected responses? That is, over the course of the task responders and punishers are learning that the offers tend to be low, meaning they are learning something about the context of cooperation (or lack thereof). Could the authors do any analyses of trial effects to see if punishment and/or rejection increases in frequency over time?

I also wonder about the potential role of expectation. If I am in a collective context and I think that others in my collective are likely to give low offers, what effect, if any, might this have on my offers? Is it possible I would lower my own offer as a result of the expectation that my group will make generally low offers? If so, this isn’t diffusion of responsibility or free-riding in the classic sense.

Is Fig 3b correct? Why does punishment positively correlate with mean offer? Shouldn’t it be the other way around?

The authors should justify why they chose the particular individual difference measures that they included. This was not well stated or established in the paper. Why these and not any of the many other measures related to group psychology? Also these measures involved analyses; were any of these predictions outlined in advance and so confirmatory tests? If not perhaps some sort of p-value correction would be needed to avoid spurious findings?

In terms of the Rand re-analysis, what do the authors make of the fact that the difference did not appear between 1 and 2 defectors? More generally the logic of including the Rand re-analysis in this paper wasn't as clear as it could be; more discussion of what it adds would be helpful.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Social cognition.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 Feb 2020

Marwa El Zein, University College London, London, UK

An interesting take on the role of individual decisions made in the context of individuals or the contexts of a collective.

We thank the reviewer for noting the interest of our work.

I had several comments that I hope will help the authors as they go forward in this work. I don't think the decision that individuals make in the collective condition is really a "collective decision" in that the individual decisions are just summed; something like a voting procedure to determine the decision would feel collective; this feels more like an individual decision in a collective context. I think making this distinction clearer would be useful.

We agree with the reviewer and have changed our vocabulary throughout the text, to clarify that we are investigating an individual decision in a collective context rather than assessing the emergence of a collective decision. We avoided any use of 'collective decision', which we replaced by 'making decision as part of a group'. We clarify since the introduction that it is a group average that we use in our design:

'Here we aimed to *directly* test the hypothesis that norm violations and their punishments differ in decisions made alone or as a contribution to a group decision.'

'In our adapted versions of the ultimatum game (UG) and the dictator game with third-party punishment (TP-DG), individuals or groups of three individuals could split their allocated points with receivers. The group condition consisted of an average of offers and did not aim to account for an interactive collective decision. Rather, it accounted for individual behaviour in a context where participants were alone vs a context where individual choices contributed to a group average, making the final offer the responsibility of 3 rather than one person.'

The introduction might want to do an earlier and clearer job distinguishing the present work from the literature on punishment of individuals who are identified as members of groups, i.e. on ingroup bias in punishment decisions.

We added this paragraph in the introduction to make the requested distinction:

'These games have repeatedly shown that humans cooperate with unrelated strangers in one-off encounters and bear personal costs to punish others who violate norms (Fehr & Fischbacher, 2004). Previous studies have also identified important in-group biases in cooperative norm-enforcement in such games (for a review, see McAuliffe & Dunham, 2016). While the results are sometimes conflicting, an in-group preference is observed in both adults and children, suggesting that belonging to the same group may protect individual group members from punishment during cooperative interactions. Contrary to this line of research, our aim here is not to investigate how different group members interact with each other. We use a context where no group affiliation exists to explore how people behave if they are making cooperative decisions alone or as a part of a neutral group, and whether attitudes to norm-enforcement changes when cooperative decisions come from a neutral player vs a group of neutral players.'

One interesting point that distinguishes the two conditions in these studies is that punishment or rejection is more effective in the group contexts because it imposes 3x the total cost; what impact, if any, do the authors think this could have?

We thank the reviewers for noting this important point that should be discussed. The point of using such a payoff was to keep the outcome similar from the perspective of the individual. This was at the expense that indeed punishing a group vs an individual imposes an overall 3X total cost, making the punishment more effective in the group condition. From this perspective, one could expect increased punishment in the group condition (because it is more effective). However, we did not observe any differences in the proportion punishment. One possibility is that:

1. There is higher punishment in the group condition because it more effective.
2. But the fact that it is group sharing responsibility for low offers, allows the group to benefit from reduced punishment (our hypothesis).

- These two points can cancel each other out so that no difference is observed between the individual and group condition. Future experiments would allow to test this explanation, for example, by using a similar design as ours, but precising that the group decision is an average of 2 previous participant's offers (that will not incur any cost in the present experiment) and a current participant that would incur the cost of punishment. This would allow to judge a decision coming from a group of three, but impose a similar cost as in the individual condition. If in this scenario, the proportion punishment would decrease, then it would support the explanation offered here.

We added this discussion point to the manuscript:
'Another possible reason why we did not observe decreased punishment for groups vs individuals in the context of our experimental design is linked to the effectiveness of the individual vs group punishment. Indeed, in the group condition, punishing the group imposed three times the total cost imposed when punishing in the individual condition. This implies that punishing a group was more effective than punishing an individual and one could then expect that the group should be punished more than the individual. This could have then overridden a decreased punishment of the group because of shared responsibility for low offers, possibly explaining why we did not observe any difference in the punishment of groups vs individuals.'

The offers that were responded to ranged for 0 to 5 of 10, each repeated twice, in each game. As the authors note this means they were mostly unfair offers. Could this have affected responses? That is, over the course of the task responders and punishers are learning that the offers tend to be low, meaning they are learning something about the context of cooperation (or lack thereof). Could the authors do any analyses of trial effects to see if punishment and/or rejection increases in frequency over time?

To answer the reviewer, we observed punishment behaviours over the course of time by splitting the rounds into 3:

Graph 1 represents the proportion punishment (or rejection, pooled or both games), pooled over individual and group conditions. A significant decrease in proportion punishment was observed in the third part of the experiment as compared to the second part (Z=2.69, p=0.007). No other changes were significant.

When individual and group conditions were separated, it appeared that no changes in punishment behaviour were observed through time in the individual condition(Graph 2). On the contrary, there was a significant increase in group punishment in the second vs first part of the experiment (Z=-2.01, p=0.0439) and then a reversal with a significant decrease in punishment from part 2 to part 3 (Z=3.61, p<.001)(Graph 3).

When checking whether differences between individual and group punishment appeared in each of the parts separately, still no significant change was observed. Nevertheless it is worth noting that in part 2, there was a tendency for higher punishment of the group vs individuals (Z=-1.67, p=0.0935) while the contrary tendency was observed in part 3 with higher punishment of individual vs group (Z=1.60, p=0.108). This somehow fits with the 2 opposite predictions from the effectiveness vs shared responsibility in group discussed above, and suggests that maybe people may use different strategies in how they punish groups vs individuals in different timepoints of the experiment. However, this would need further investigation and is just a discussion given the current presented results.

To go back to the reviewer’s point that participants may have learned that this is a context with low cooperation and increase their punishment over time, we seem to rather observe a decrease in punishment in the last part of the experiment. And this decrease is specific to the group punishment. It is possible to imagine that people increase the group punishment after observing in the first part of the experiment that offers are low (and the increase observed only in group because it is more effective to punish the group), but then give up on their norm-enforcement behaviour in the last part of the experiment. If this last decrease was observed in both group and individual conditions, it could have been solely explained by the motivation to maximize gains at this point of the experiment. However, the fact that the decrease is only in the group condition may suggest that they end up punishing less the group specifically because the group shares responsibility for the low offers that everyone seems to be giving in this context (decrease their norm-enforcement of a group specifically in a context where cooperation is low).

I also wonder about the potential role of expectation. If I am in a collective context and I think that others in my collective are likely to give low offers, what effect, if any, might this have on my offers? Is it possible I would lower my own offer as a result of the expectation that my group will make generally low offers? If so, this isn’t diffusion of responsibility or free-riding in the classic sense.

This is a very interesting and plausible explanation to the observed decrease in the group offers in the group condition. We attempt to test for whether it explains our results: We reasoned that the collective context of low cooperation should be first learned, before participants would lower their offers because everyone else is also giving low offers (so I expect others to give low offers and prefer to have a similar behaviour as everyone else). Therefore, if the participants give lower offers as group vs individual players since the first decision as offerors, then there is much less chance that they would have already learned the context of (lack of) cooperation. We compared the first offer as an individual player to the first offer as a group and found a significant decrease in the offer in the group vs individual condition (Z=3.24, p=0.0012), We added this analysis in the results: ‘This was also true when considering only the first trial where people made an offer individually and the first trial where they made the offer as a group average (Z=3.24, p=0.001, r=0.18).’

And in the discussion: ‘Finally, another explanation is that participants lowered their offers because of their expectation that other group members will also give low offers, given that mostly unfair offers were presented in the context of the experiment (when participants were playing the receiver in the UG and the punisher in the TP-DG). However, the offer decrease for groups was already significant in the first round played in group as compared to the first round played individually, when they had less time to learn about the context they were playing in.’

Is Fig 3b correct? Why does punishment positively correlate with mean offer? Shouldn’t it be the other way around?

Yes Fig 3b is correct indeed: punishment positively correlated with mean offer showing that people who are more generous, also care more about norm-enforcement, possibly to preserve a cooperative context. We agree with the reviewer that the opposite could have also been true so that people who are more generous don’t want to punish other. We believe this to relate to our discussion on social and antisocial punishment : one can expect punishment to be social because it has the good intention of establishing cooperative norms, but also punishment can be antisocial in the sense that is driven by an angry reaction toward an unfair offer and involves incurring a cost on others. Interesting we found that both prosocials and those who score high on psychopathy seem to show an increased punishment behaviour corroborating this idea of the co-existence of social and antisocial punishment.

The authors should justify why they chose the particular individual difference measures that they included. This was not well stated or established in the paper. Why these and not any of the many other measures related to group psychology? Also these measures involved analyses; were any of these predictions outlined in advance and so confirmatory tests? If not perhaps some sort of p-value correction would be needed to avoid spurious findings?

We now explain in the introduction why we decided to include these specific questionnaires:

'In an exploratory analysis, we collected self-reported scales to further investigate individual differences in norm violations and their punishment. Social value orientation was measured to link participants' offering and punishment behaviours to a trait measurement of sharing with others (Murphy et al., 2011). A psychopathy scale (Paulhus & Williams, 2002) was collected to test the idea that higher psychopathy scores may be associated with lower influence of being in a group or punishing a group, as people with higher psychopathy scores may care less about being in a group. Finally, political identification was measured in order to test whether differences would appear in fairness attitudes in the context of playing alone or as a group, because liberals and conservatives have been associated with different considerations of fairness and reciprocity (Graham et al. 2009).'

We performed GLMs to assess which questionnaire score significantly predicted our behavioural variables. Indeed, as the reviewer points out, we still had to perform 16 GLMs (8 variables OFF,PUN, OFFDIFF and PUNDIFF and associated RTs X 2 games UG and TP-DG), and to properly correct for multiple comparisons we should multiply our p-values by 16. We now state this in the results, keep the analyses only for the GLMs with the corrected p-values. We remove the correlation analyses that suffer even more from the multiple comparison problem (16X3 scales=48scales). We only keep post-hoc analyses that illustrate the findings of the GLMs (for figure 4).

Even though many of the mixed models results become close to significance/ not significant at all, we still report all the effects that were significant in the GLM, but now showing the corrected p-value and giving more cautious conclusions and interpretations. We believe that the questionnaire results, even though exploratory and not robust, are interesting and worse considering for future replications.

In terms of the Rand re-analysis, what do the authors make of the fact that the difference did not appear between 1 and 2 defectors? More generally the logic of including the Rand re-analysis in this paper wasn't as clear as it could be; more discussion of what it adds would be helpful.

We hope to have clarified the logic of having this analysis. The idea is that the higher the number of participants violating norms by contributing with low amounts, the less this behaviour will be punished. So, being backed up by others in defection allows to protect from punishment. We added in the introduction:
'In addition to these two adapted games, we re-analysed available data from a previous study (Rand *et al.*, 2009) that involved a public goods game between four players with punishment to test whether the use of punishment changes with the number of people defecting. Similarly to playing in a group, we hypothesize that when several people violate norms, they may get away easier with it under the shield that 'others did it too'. This allows to share responsibility for a punishable act,

which thereby may become less prone to punishment.'

And in the results:

'We tested our third prediction that inflicting punishment on unfair contributions will decrease with the number of people defecting. The reasoning here is that the more defectors on a given round, i.e., the more people who violate the norm, the more their behaviour can be justified (they are not the only ones!) and therefore the more they can benefit from reduced punishment.'

We have also changed the way we choose the defectors in this re-analysis following the other reviewer's comment, and the new results show a clearer linear decrease from 1 to 4 defectors (new figure 5). With our initial analyses, the results suggested that the decrease becomes significant only when there are enough defectors, i.e. 3, but not 2.

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 October 2019

https://doi.org/10.21956/wellcomeopenres.16925.r36622

❌ **Justin W. Martin**
Department of Psychology, Boston College, Chestnut Hill, MA, USA

In this article, the authors investigate the influence of being in a group (vs. being an individual) on cooperation and punishment. In particular, they vary whether unfairness is a result of individual vs. group decision-making in the context of second-party punishment (indexed using the Ultimatum Game) and third-party punishment (indexed using the Third-Party Punishment variant of the Dictator Game). They find cooperation is lower when decisions are made by a group relative to when made by an individual. They find no overall punishment difference between group vs. individual decision-making, and find greater punishment as a second-party vs. a third-party. Finally, they find that reaction times are slower when deciding whether or not to punish a group (relative to punishing an individual), though they do not find that the decision to punish itself is slower when punishing a group.

There are some things to like about this article. In particular, the topic is interesting and the authors have made their task and data publicly available.

However, this article also has some major limitations.

First, it is not clear that the authors' main manipulation allows them to get at the principle mechanism they articulate. Specifically, the authors suggest that individuals may be less cooperative in the context of a group and punish groups more because group membership reduces responsibility (p. 2 first and third paragraphs). However, the authors do not assess perceptions of responsibility or include any additional manipulation that would directly target perceptions of responsibility. In the absence of any such mechanistic evidence, any effect of their manipulation could be due to other factors. In particular, one plausible alternative explanation focuses on practical considerations. That is, because the group's offer was an average of the offers of the constituent members, any influence of being in a group (vs. being an individual) could simply be a function of the averaging that occurs to the offers. If I want to make a low offer as part of a group, I need to make my own offer especially low, so as to offset whatever offers the others may make. More generally, my point is that the authors have wished to interpret the lowered offers they find in the group condition as indicative of a willingness to violate norms in the context of group decision-making, but they do not provide any direct evidence in favor of that possibility.

Second, the authors should do more to either qualify their interpretations or better justify them. For instance, they explain lower offers made as part of a group with the following (p. 7): "This suggests that participants were expecting less punishment when playing in a group as compared alone." Do the authors have any data to suggest that participants' expectations were the driving factor behind lower offers? In the absence of such evidence, these kinds of statements need to be qualified. A number of mechanisms could account for this difference (e.g. a different norm for offers in group vs. individual contexts; the strategic account I offered above; the responsibility focused account the authors proposed) and the authors need more evidence if they want to suggest that one is the primary factor.

Third, the authors make predictions in the introduction about the relationship between reaction time and offers but do not explain or justify these predictions. More should be included here. In the absence of any explanation, these hypotheses feel post-hoc and unjustified. This is especially important given that these hypotheses were not pre-registered, there is no internal replication and many of these p-values hover around $p = 0.05$. Similarly, the authors should include more discussion of the scales they chose to include, as well as what their predictions for the relationship between these scales and their other data were. This is again especially important because these analyses were not pre-registered, were not replicated internally and sometimes hover around $p = 0.05$. And, the authors perform many analyses and comparisons using these scales, which would necessarily yield some significant results. In the absence of such justification, it would be better for the authors to describe these analyses as exploratory and caution their interpretation.

Fourth, the authors should be more cautious in their discussion of their results. For instance, they write (p. 9): "there is a consistent slowing down to punish three individuals as compared to one individual". This implies that punishment was slower in the group condition than the individual condition. However, this is not true: The authors found no relationship between the group manipulation and whether or not punishment was administered. They did find that responses (ignoring whether that response was to punish or not) were slower in the group condition. The way the authors describe this result, however, seems to indicate an effect that they did not find. Similarly, in the discussion they write (p. 11): "We importantly found that the punishment decisions were slower when punishers were faced with groups vs individuals, suggesting that it is also more time costly and less intuitive to choose whether to punish a group". This follows the authors' discussion of their overall finding of slower punishment vs. non-punishment. Again, readers will likely interpret this sentence as suggesting an effect that the authors do not find. When discussing these results, greater care needs to be taken in describing exactly which results were significant and which were not. Relatedly, the authors describe their reaction time results as suggesting that punishers (p. 11) "are more reluctant to punish a group vs. an individual". Unless the authors have other evidence specifically implicit "reluctance", they should be more cautious in interpreting their data. Slower reaction time could result from a number of factors (e.g. punishing a group may involve thinking about all 3 individuals impacted, slowing punishment relative to a condition in which only 1 person is impacted).

Fifth, the authors note that they arrived at their sample size by selecting 40 participants per treatment and then multiplying by 4, yet they arrive at 150 participants (not 160). Was this perhaps a typo? If not, they should explain why they recruited 10 fewer participants than would make sense given their logic.

I also have a number of more minor concerns.

First, the authors interpret the slowed reaction time to punish vs. not punish as indicating that punishment is (p. 10) “more costly than not punishing”. Given the context of material costs paid and punishment itself (which is often discussed as being materially costly or non-costly), the authors should be clear earlier that they mean costly in terms of time. This is a somewhat non-standard meaning of costly and so clarifying it would be useful.

Second, there are a decent number of typos/grammar issues (errors bolded; p. 3 : “To be protected against punishment, individuals delegate **decision** to others,...”; p. 3: “for an individual, being in a group could be a good way to reduce responsibility and thereby, the associated punishments for norm violations; p. 7: “To conclude on the proposer role, participants gave lower **offer** as a group vs alone“). I would encourage the authors to go over the paper carefully to correct these issues.

Third, comparing directly to the data in Rand *et al*. (2009)[1] struck me as odd. There are likely many differences between that study and this one, and so it is not clear that the comparison is warranted. Furthermore, the analysis seems odd. When the authors say “We considered as defectors the players who gave less than the mean amount of contribution at each round”, what mean are they referring to? The mean for that round? This would then indicate that there were generally 2 defectors on every round (unless everyone contributed the same amount, e.g.). This seems like an arbitrary way of establishing what is considered defection. What about using half of the amount possible as a standard? And if the authors used a grand mean as the threshold, that seems problematic as well. For instance, if there was a group that was contributing below the grand mean, and yet all members were contributing the same, is it reasonable to call this defection? More generally, the relationship between responsibility and the number of defectors was not made totally clear.

Fourth, in general I am not a fan of y-axis that do not include at least 1 anchor (ideally both should be included). In Figure 2A, the authors hone in on just the range from 3-4, only 10% of the entire scale. A more accurate scale would at least include the lowest value possible (0) but ideally would include the full range of the scale. Otherwise, the authors could be accused of restricting their y-axis to make their effect look larger.

Fifth, in addition, in Figure 2A, the authors use two asterisks to mark the significance of the effect, which their caption describes as indicating p < 0.01. However, in-text they describe this difference as p = 0.025. This error should be corrected.

Sixth, I would encourage the authors to pre-register their sample size, hypotheses and analysis plans in the future.

References

1. Rand DG, Dreber A, Ellingsen T, Fudenberg D, et al.: Positive interactions promote public cooperation. *Science*. 2009; **325** (5945): 1272-5 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Punishment, cooperation, moral judgment.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 07 Feb 2020

Marwa El Zein, University College London, London, UK

In this article, the authors investigate the influence of being in a group (vs. being an individual) on cooperation and punishment. In particular, they vary whether unfairness is a result of individual vs. group decision-making in the context of second-party punishment (indexed using the Ultimatum Game) and third-party punishment (indexed using the Third-Party Punishment variant of the Dictator Game). They find cooperation is lower when decisions are made by a group relative to when made by an individual. They find no overall punishment difference between group vs. individual decision-making, and find greater punishment as a second-party vs. a third-party. Finally, they find that reaction times are slower when deciding whether or not to punish a group (relative to punishing an individual), though they do not find that the decision to punish itself is slower when punishing a group.
There are some things to like about this article. In particular, the topic is interesting and the authors have made their task and data publicly available.
However, this article also has some major limitations.

We thank the reviewer for acknowledging the positive aspects of the paper, and for pointing out to major issues that we hope to have addressed.

First, it is not clear that the authors' main manipulation allows them to get at the principle mechanism they articulate. Specifically, the authors suggest that individuals may be less cooperative in the context of a group and punish groups more because group membership reduces responsibility (p. 2 first and third paragraphs). However, the authors do not assess perceptions of responsibility or include any additional manipulation that would directly target perceptions of *responsibility. In the absence of any such mechanistic evidence, any effect of their manipulation could be due to other factors. In particular, one plausible alternative explanation focuses on practical considerations. That is, because the group's offer was an average of the offers of the constituent members, any influence of being in a group (vs. being an individual) could simply be a function of the averaging that occurs to the offers. If I want to make a low offer as part of a group, I need to make my own offer especially low, so as to offset whatever offers the others may make. More generally, my point is that the authors have wished to interpret the lowered offers they find in the group condition as indicative of a willingness to violate norms in the context of group decision-making, but they do not provide any direct evidence in favor of that possibility.*

We agree with the reviewer that our experiment alone does not allow to exclusively explain our findings by the fact that being in a group reduces individual responsibility. Nevertheless, our design was driven by this hypothesis, and we now clarify this:

'We note that while driven by our shared responsibility in groups hypothesis, our experimental design does not allow to characterize the exact mechanisms underlying differences in behaviour between norm violations and their punishment (and thereby be able to affirm that the effects are due to sharing responsibility solely relying on this experiment). However, to our knowledge, this is the first experimental design addressing whether these differences exist with such a controlled design, and directly comparing second and third-party punishment in repeated one-shot trials and a within-participant design.'

Moreover, we strongly believe that previous literature supports shared responsibility as a possible underlying mechanism for the behaviours we observe and now provide a more elaborate explanation why in the discussion section:

'Our current finding that people are less generous in a group corroborates the idea that people in groups violate the norms more than alone given that (1) it replicates previous studies showing that individuals in groups display free-riding behaviours (Morgan & Tindale, 2002; Tindale & Kameda, 2017; Wildschut et al., 2003). These previous studies compared groups facing groups to individuals facing individuals and showed that groups are more competitive (Wildschut et al., 2003), defect more in a prisoner dilemma game (Morgan & Tindale, 2002) and offer less in a joint decision in an ultimatum game (Bornstein & Yaniv, 1998). Our results complement these studies by showing that even when facing one individual, people are less generous if they are part of a group vs alone. (2) Interestingly, here, we show for the first time that this decreased generosity in group correlates with people's overall generosity. Indeed, only those who gave low offers displayed a difference between playing in a group or alone. This shows that the group was compatible with the intention of those who were less sensitive to the norms and violated them more. (3) It has been suggested that increased defection in groups may relate to reduced 'identifiability' as a group, supporting the idea that people feel less 'accountable' when making selfish choices (Kugler et al. 2012). Backing this, the fact that groups of two members that could be easily identified playing a dictator game gave higher offers as a group (Cason and Mui, 1997) was linked to a reduced anonymity in such a context (Luhan et al. 2009). On the contrary, when anonymity is preserved in computer-based rather than face-to-face interactions, group members defect more than individuals. (4) Finally, groups that are procedurally interdependent (Wildschut et al. 2003) are less cooperative. This may be due to the fact that individual choices could not have been traced back, again allowing for anonymity and hiding behind the group (Kugler et al. 2012).'

Finally, we discuss other possibilities as the reviewer urged us to do, including the one given by the reviewer:

‘Despite substantial evidence that the effects observed in our study may be due to increased norm violation in groups because of shared responsibility, our experimental design does not allow us to disentangle the exact mechanism(s) underlying the observed reduction of offers in groups, and other explanations may as well be possible. For example, people may become more rational (Bornstein & Yaniv, 1998), and less biased as a group. They also may feel supported enough to act in a selfish, i.e. more greedy way (Schopler et al. 1995). In addition, one may argue that in our study, participants may have been basing their decision on the averaging procedure, by adjusting their offers to the expectations that others in the group will do. For example, a person who wants to give a low offer may lower their offer to offset the other group members offers. However, this last explanation is not consistent with the fact that only low offerors showed a difference between playing alone or in group as one would also expect those who want to give high offers to then adjust their behaviour by giving higher offers in groups (while they don’t). Finally, another possible explanation is that participants lowered their offers because of their expectation that other group members will also give low offers, given that predominantly unfair offers were presented in the context of the experiment (when participants were playing the receiver in the UG and the punisher in the TP-DG). However, the offer decrease for groups was already significant in the first round played in group as compared to the first round played individually, when they had less time to learn about the context they were playing in.’

Second, the authors should do more to either qualify their interpretations or better justify them. For instance, they explain lower offers made as part of a group with the following (p. 7): “This suggests that participants were expecting less punishment when playing in a group as compared alone.” Do the authors have any data to suggest that participants’ expectations were the driving factor behind lower offers? In the absence of such evidence, these kinds of statements need to be qualified. A number of mechanisms could account for this difference (e.g. a different norm for offers in group vs. individual contexts; the strategic account I offered above; the responsibility focused account the authors proposed) and the authors need more evidence if they want to suggest that one is the primary factor.

We agree with the reviewer and now suggest different possible interpretations of the results in the discussion. We added to the specific sentence mentioned here: ‘This *may* suggest that participants were expecting less punishment…’ as this sentence allowed us to do the transition to the punishment section, but do not elaborate further in the results section, in order to discuss different explanations in the discussion.

Third, the authors make predictions in the introduction about the relationship between reaction time and offers but do not explain or justify these predictions. More should be included here. In the absence of any explanation, these hypotheses feel post-hoc and unjustified. This is especially important given that these hypotheses were not pre-registered, there is no internal replication and many of these p-values hover around p = 0.05. Similarly, the authors should include more discussion of the scales they chose to include, as well as what their predictions for the relationship between these scales and their other data were. This is again especially important because these analyses were not pre-registered, were not replicated internally and sometimes hover around p = 0.05. And, the authors perform many analyses and comparisons using these scales, which would necessarily yield some significant results. In the absence of such justification, it would be better for the authors to describe these analyses as exploratory and caution their interpretation.

We apologize for not being clear about some of our predictions, we now better explain the motivations behind the hypotheses we have in the introduction:

‘Applying our key two hypotheses described above to the experimental paradigm, we predicted that 1) an individual in the group will offer less than the same individual alone, and make his/her decision faster because of less hesitation about violating the norms within a group, 2) the group will be punished less than the individual and punishment vs no punishment decisions facing a group will be slower reflecting a more hesitant choice, and 3) Inflicting punishment on unfair contributions will decrease with the number of people defecting. In an exploratory analysis, we collected self-reported scales to further investigate individual differences in norm violations and their punishment. Social value orientation was measured to link participants’ offering and punishment behaviours to a trait measurement of sharing with others (Murphy et al., 2011). A psychopathy scale (Paulhus & Williams, 2002) was collected to test the idea that higher psychopathy scores may be associated with lower influence of being in a group or punishing a group, as people with higher psychopathy scores may care less about being in a group. Finally, political identification was measured in order to test whether differences would appear in fairness attitudes in the context of playing alone or as a group, because liberals and conservatives have been associated with different considerations of fairness and reciprocity (Graham et al. 2009).’

We performed GLMs to assess which questionnaire score significantly predicted our behavioural variables. Indeed, as the reviewer points out, we still had to perform 16 GLMs (8 variables OFF,PUN, OFFDIFF and PUNDIFF and associated RTs X 2 games UG and TP-DG), and to properly correct for multiple comparisons we should multiply our p-values by 16. We now state this in the results, keep the analyses only for the GLMs with the corrected p-values. We remove the correlation analyses that suffer even more from the multiple comparison problem (16X3 scales=48scales). We only keep post-hoc analyses that illustrate the findings of the GLMs (for figure 4).
Even though many of the mixed models results become close to significance/ not significant at all, we still report all the effects that were significant in the GLM, but now showing the corrected p-value and giving more cautious conclusions and interpretations. We believe that the questionnaire results, even though exploratory and not robust, are interesting and worse considering for future replications.

Fourth, the authors should be more cautious in their discussion of their results. For instance, they write (p.9): “there is a consistent slowing down to punish three individuals as compared to one individual”. This implies that punishment was slower in the group condition than the individual condition. However, this is not true: The authors found no relationship between the group manipulation and whether or not punishment was administered. They did find that responses (ignoring whether that response was to punish or not) were slower in the group condition. The way the authors describe this result, however, seems to indicate an effect that they did not find. Similarly, in the discussion they write (p. 11): “We importantly found that the punishment decisions were slower when punishers were faced with groups vs individuals, suggesting that it is also more time costly and less intuitive to choose whether to punish a group”. This follows the authors’ discussion of their overall finding of slower punishment vs.non-punishment. Again, readers will likely interpret this sentence as suggesting an effect that the authors do not find. When discussing these results, greater care needs to be taken in describing exactly which results were significant and which were not. Relatedly, the authors describe their reaction time results as suggesting that punishers (p. 11) “are more reluctant to punish a group vs. an individual”. Unless the authors have other evidence specifically implicit “reluctance”, they should be more cautious in interpreting their *data. Slower reaction time could result from a number of factors (e.g. punishing a group may involve thinking about all 3 individuals impacted, slowing punishment relative to a condition in which only 1 person is impacted).*

As suggested here, we changed the vocabulary used to 1) clarify when the decision is the decision to punish or not to punish, and not only the 'punishment' decision as might have been implied. We changed the sentence to '…there is a consistent slowing down to make the decision of whether to punish or not three individuals as compared to one individual.'. We added in every place where it was needed: 'punishment or no punishment decision' and 'punish or not punish'.
2) avoid incorrect interpretations of the slower RTs when facing groups by using 'slowed down' or 'slower' instead of 'reluctance'.
We agree that the slowing down when facing a group vs individuals may results from different factors. We had thought about the interpretation that the it might be due to thinking about who gave what and avoiding the punishment of fair participants trapped in group with unfair partners. We excluded that interpretation by showing that the slowing down was true even for very low offers, where the individual members offers were not ambiguous. We then suggest that given that the difference in reaction times is only present for low offers, participants seem to be slowed for groups vs individuals only when norms were violated (and only in the ultimatum game). We believe the explanation suggested by the reviewer that punishing a group may involve thinking about 3 individuals impacted (vs 1 in the individual condition) to be complementary to our conclusion: indeed, what the group offers is that the person punishing will take into account several people into account instead of one, and thereby be more hesitant to decide whether to punish or not the group. The fact that this only happens in second-party punishment and for low offers makes the effect specific to norm violations toward one self, rather than a general thinking about impacting 3 vs 1 person (or it should have been observed for all offers and in third-party punishment as well).

Fifth, the authors note that they arrived at their sample size by selecting 40 participants per treatment and then multiplying by 4, yet they arrive at 150 participants (not 160). Was this perhaps a typo? If not, they should explain why they recruited 10 fewer participants than would make sense given their logic.

We recruited 10 fewer participants (1 experimental session) because of practical issues related to timing and participants recruitment. As our sample size was based on a previous study using a different design, and our experiment design was novel, we do not believe having a few less participants than planned is highly problematic here. Indeed, most of our results were similar when the analyses were done on the first recruited group of 80 participants. We clarify in the methods: 'Given our 4 conditions of interest (individual or group proposers in the proposer or the punisher role), we multiplied this number by 4 and tested 150 (instead of 160) participants because of practical issues related to timing and participants recruitment.'

I also have a number of more minor concerns.

First, the authors interpret the slowed reaction time to punish vs. not punish as indicating that punishment is (p. 10) "more costly than not punishing". Given the context of material costs paid and punishment itself (which is often discussed as being materially costly or non-costly), the authors should be clear earlier that they mean costly in terms of time. This is a somewhat non-standard meaning of costly and so clarifying it would be useful.

We thank the reviewer for noting that using 'costly' for time in this context can be confusing. We decided not to use it and describe the slowing down effects as ‘less intuitive’ as more commonly described (Rand et al., 2012; Fehr & Fischbacher, 2004).

Second, there are a decent number of typos/grammar issues (errors bolded; p. 3 : “To be protected against punishment, individuals delegate ***decision*** *to others,...”; p. 3: “for an individual, being in a group could be a good way to reduce responsibility and thereby, the associated punishments for norm violations; p. 7: “To conclude on the proposer role, participants gave lower* ***offer*** *as a group vs alone“). I would encourage the authors to go over the paper carefully to correct these issues.*

We sincerely apologize for the errors and hope to have fixed them all.

Third, comparing directly to the data in Rand et al. (2009) struck me as odd. There are likely many differences between that study and this one, and so it is not clear that the comparison is warranted. Furthermore, the analysis seems odd. When the authors say “We considered as defectors the players who gave less than the mean amount of contribution at each round”, what mean are they referring to? The mean for that round? This would then indicate that there were generally 2 defectors on every round (unless everyone contributed the same amount, e.g.). This seems like an arbitrary way of establishing what is considered defection. What about using half of the amount possible as a standard? And if the authors used a grand mean as the threshold, that seems problematic as well. For instance, if there was a group that was contributing below the grand mean, and yet all members were contributing the same, is it reasonable to call this defection? More generally, the relationship between responsibility and the number of defectors was not made totally clear.

We are aware that there are differences between Rand et al. 2009 study and ours, but still believe that the re-analysis of the public good game with a punishment treatment is very relevant to our hypothesis that shared responsibility reduce punishment. We apologize for not making our reasoning clear enough and hope to have done so in our revised version of the manuscript. The idea is that the higher the number of participants violating norms by contributing with low amounts, the less this behaviour will be punished. So, being backed up by others in defection allows to protect from punishment. We added in the introduction:
‘In addition to these two adapted games, we re-analysed available data from a previous study (Rand *et al.*, 2009) that involved a public goods game between four players with punishment to test whether the use of punishment changes with the number of people defecting. Similarly to playing in a group, we hypothesize that when several people violate norms, they may get away easier with it under the shield that ‘others did it too’. This allows to share responsibility for a punishable act, which thereby may become less prone to punishment.’

And in the results:

‘We tested our third prediction that inflicting punishment on unfair contributions will decrease with the number of people defecting. The reasoning here is that the more defectors on a given round, i.e., the more people who violate the norm, the more their behaviour can be justified (they are not the only ones!) and therefore the more they can benefit from reduced punishment.’

<u>On the how we established who defected:</u>
We used the mean contribution per round to have a baseline for each group – because participants decide whether to punish after they see the contribution of each player. Our reasoning was that

they may re-scale who they would want to punish based on what was the mean contribution at a given round. However, we understand that this may seem less intuitive (and possibly more problematic) than just sticking to the standard way, which is taking half the contribution as suggested by the reviewer.

We redid the analyses now considering the defectors as those who contributed less than 10, and the results similarly show the more defectors, the less use of punishment. With this way of calculating defectors, there are trials with 1 to 4 defectors (i.e. rounds where everyone is considered as defector).

Fourth, in general I am not a fan of y-axis that do not include at least 1 anchor (ideally both should be included). In Figure 2A, the authors hone in on just the range from 3-4, only 10% of the entire scale. A more accurate scale would at least include the lowest value possible (0) but ideally would include the full range of the scale. Otherwise, the authors could be accused of restricting their y-axis to make their effect look larger.

We have changed the figure as suggested including the full range of the scale.

Fifth, in addition, in Figure 2A, the authors use two asterisks to mark the significance of the effect, which their caption describes as indicating p < 0.01. However, in-text they describe this difference as p = 0.025. This error should be corrected.

We apologize for this error, it is now corrected.

Sixth, I would encourage the authors to pre-register their sample size, hypotheses and analysis plans in the future.

We agree with the reviewer that this should be done whenever possible and we will attempt to do that in the future.

Competing Interests: No competing interests were disclosed.