Review report

Short summary

The Effect of Individual and Group Punishment on Individual and Group-Based Dishonesty is a stage 1 RR proposing to investigate the effect of punishment (structure) and payoff structure on dishonest behavior.

The authors provided a clear introduction of the existing literature and then detailed the hypothesis generation process, followed by sample size planning and a clear illustration of experimental procedures, with the main task being a tax evasion game. The main IVs are payoff structure (individual vs. group) and punishment (no vs. individual vs. group). The authors also planned to measure a few variables including feeling of commitment, guilt and anger, as well as humility-honesty.

The main hypotheses are:

1) Group payoff increases dishonesty (i.e., non-compliance in the tax evasion game) compared to the individual payoff (baseline) for the no punishment treatment;

2) A main effect of risk of punishment. Risk of punishment (vs. no punishment) reduces dishonesty (i.e., non-compliance);

3) A main effect of type of punishment. Group punishment will show stronger effects in reducing dishonesty compared to individual punishment; and

4) $H4_0$: No interaction effect of type of punishment and payoff structure. $H4_1$: An interaction effect of risk and type of punishment and payoff structure. Group payoff increases dishonesty compared to individual payoff, but only for no punishment or individual punishment. For group punishment, individual payoff increases dishonesty compared to group payoff.

The analysis plan is consistent with the hypotheses overall. I also thank the authors providing the details of power analyses and experimental materials in the supplementary.

However, I do have a few comments that I feel the authors should consider before the IPA. Below please find my specific comments. I hope that they are helpful and apologize in advance if I have misunderstood parts of the study.

Comments

Abstract

1. ‘High-powered’ is not a precise term. Whether or not a sample offers high power also depends on the specific tests and the expected effect sizes. Perhaps

this term can be removed (Minor).

Introduction

2. In the abstract and the introduction, the authors mentioned the severity of a test quite often. From my perspective, the proposed study examines the presence of punishment (yes. Vs. no) but not the severity, unless here we consider that the group punishment condition is more severe than the individual punishment condition. However, these two conditions differ not only in terms of severity.
3. In Zickfeld et al. 2023, the effect is reported in terms of Hedge's g, not Cohen's d. The conversion between the two could be made clearer in the manuscript (Minor).

Method

4. The authors stated that 'In the unlikely case that we must exclude more than 20% of participants based on the registered exclusion criteria, we will collect another round of participants to fill up the original sample size.'. Does that mean that the authors will only do another round of collection when the exclusion is 20% or higher? Then it could mean that for some tests, you don't have sufficient power. Why not recruit participants until your valid sample size reaches the planned sample size?
5. For the simulated power analyses, if the total sample is 630. Then the sample for H1 would be 210 (2 no punishment conditions). Therefore, the power analysis should report the results with a sample of 210 instead of 630 (as reported in the design Table). The same principle applies to other hypotheses. Could the authors confirm that the specific sample sizes for different hypotheses are all sufficiently powered?
6. In the procedures, the authors propose to measure feeling of commitment at the end of the tax task as manipulation check. The manipulation check might be influenced by the task outcomes. what about measuring it before the formal task but after the practice?
7. On the analytical side, what would count as successful manipulation? It would also be nice to include the manipulation check test in the design table.
8. Stage 1 report should not include exploratory analysis. Therefore, I suggest to only keep the planned analyses for manipulation check and the testing of H1 to H4. The authors can of course still run the additional analysis in Stage 2 but report them as exploratory.
9. Exploratory analyses on actual punishment, shouldn't it be that actual punishment from the previous round influence behavior in this round?
10. Exploratory analysis on moral emotions: perhaps it would also be interesting to add a measure of shame, given the theoretical relevance of shame vs. guilt? (e.g., https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6143989/)
11. For the mediation analysis, if the moral emotions are not measured after each round of the game but at the end of the tax task. I wonder if it is appropriate to run such a mediation analysis. It's likely the case that the task outcomes

influenced the emotions rather than the other way around.

12. Design table: it seems that for rejecting H0, you are using NHST, but for not rejecting H0, you are using minimal effect testing against SESOI. However, it could be case that the test is statistically significant but smaller than the SESOI. See here for recommendation on testing under a unified framework: https://psycnet.apa.org/fulltext/2024-76241-001.pdf https://royalsocietypublishing.org/doi/full/10.1098/rsos.221328
13. In the supplementary materials, 'For the experimental treatments we expected a small effect in the individual payoff/individual punishment treatment. Given previous findings in the dishonesty literature (Gerlach et al., 2019; Leib et al., 2021; Zickfeld et al., 2023), we set our smallest effect size of interest (SESOI) for these main effects at d = +/- .15. Employing this effect for the group payoff/individual punishment treatment would suggest an increase in compliance of 5% (d ~ .16). We expected a somewhat stronger effect for the individual payoff/individual punishment treatment and set this at an increase of 10% (d = .33) or a compliance of 75%.' The effect of individual payoff/individual punishment treatment appeared twice with different effect size expectations. Could the authors check?

Yikang Zhang