ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

ERROR RECOMMENDER REPORT

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. Journal of experimental social psychology, 55, 110-125.

DECISION:

Minor errors that do not affect the core conclusions

Recommendation by

Jamie Cummins, University of Bern

31 July 2025

DECISION & RECOMMENDATION

Based on the reviewer's report and the authors' response, I return the decision that the original article contains **Minor Errors that do not affect the core conclusions**. That is, errors that have the benefit of being detectable thanks to the presence and sharing of research materials, but whose scope and implications are minor. The reviewer reproduced the principal mixed-model results after updating/repairing the SAS code; discrepancies that remain are limited to reporting and rounding issues, effect-size computation differences for reported Cohen's ds, and a denominator degrees-of-freedom typo. A decision of *minor errors* entails the publication of the report, authors response, and recommendation on the ERROR website (error.reviews). The authors are also asked that they recognise the minor errors associated with their manuscript in future discussions of their article. We commend the authors in already taking action to publicly acknowledge and rectify the errors found here, specifically by committing to posting cleaned data and functioning code to OSF. A highly-cited paper which was not computationally reproducible ~1 year ago now has clean, working, openly-available code thanks to the efforts of the reviewer and the author.

RECOMMENDER'S REPORT

Jamie Cummins

Firstly, I would like to thank both Dr. Sørensen (reviewer) and Prof. Cikara (corresponding on behalf of the authorship team) for their cooperation throughout this review. To date, this is the oldest paper which has been audited through ERROR at nearly 11 years old at the time of writing this recommendation. Post-publication peer review can be daunting, particularly when it concerns work conducted more than a decade ago, and I thank Prof. Cikara for her willingness to subject her paper to review. I also thank Dr. Sørensen for his investment of a substantial amount of time and effort into fixing a number of different issues with the original analysis code to get it into a functioning order to facilitate his review.

The substance and style of the reviewer report and authors' response embody what we hope to see more of in academic research: acceptance of the possibility that errors occur; inspection and useful discourse about potential errors that is well-documented and verifiable; and acknowledgement and suitable correction when errors are found.

ERROR recommendations are public documents whose function is to (1) communicate the presence or absence of any errors detected, (2) consider their severity, and (3) provide discussion of how similar errors elsewhere might be prevented or detected. Materials for all error reports can be found at [osf.io/fpw4r](osf.io/fpw4r).

Summary of errors detected & how they could be prevented in future

The reviewer encountered several difficulties reproducing the analyses initially due to issues with the provided SAS code. These scripts required many revisions and repairs to run successfully; in other words, the results of the paper were not computationally reproducible from the code in its original state.

Once corrected, the code fully reproduced the majority of the original manuscript's statistical results, although it is important to note that some discrepancies remained. These discrepancies included small reporting errors, such as an incorrect denominator degrees-of-freedom (df) reported in Table 1 (196 instead of 199), isolated mismatches in p-values, and rounding differences in reported standard errors and effect sizes. Additionally, the calculation of Cohen's d from least-squares means (LSMEANS) contrasts was found to rely on an approximate and somewhat non-standard method. Despite these deviations, the recalculated values did not alter the central conclusions of the paper. In future, such deviations could be prevented by sharing all code and data used to produce the reported results, as well as having a researcher who is not part of the authorship team independently reproduce the results using this code and data before publication. Clear

documentation of any non-standard effect-size calculations used, particularly within mixed-model frameworks, should also be provided. Finally, rounding should be applied in a consistent manner across all studies and analyses reported in the manuscript.

The reviewer also noted certain ancillary descriptive results—specifically, pilot data, participant identification scores, and Cronbach's α reliability coefficients—that were indeterminable due to unavailable data. Although these elements were peripheral to the paper's core findings, such issues could be avoided in the future by ensuring that all data and code are freely shared upon publication of the manuscript.

Discussion of individual issues raised
The reviewer identified several issues related primarily to computational reproducibility, reporting accuracy, and data auditability. Initially, the SAS code provided by the authors was not functional, requiring several hours of work from the reviewer to get it running before it successfully reproduced the manuscript's central statistical results. Following these adjustments, the outputs generally matched those originally reported, albeit with some outstanding minor deviations. The authors have committed to sharing fully operational scripts and cleaned datasets on the Open Science Framework (OSF).

Further reporting inaccuracies were identified, specifically concerning the computation and documentation of effect sizes and statistical parameters. Cohen's d values derived from LSMEANS contrasts were computed using a non-standard approximate method. Nevertheless, the reviewer found that alternative methods yielded similar (but not identical) estimates, meaning substantive conclusions remained unchanged. To enhance clarity and reproducibility in future research, it is recommended that authors clearly document the exact computational methods for effect sizes or alternatively report standardized contrasts or raw LS-mean differences along with their standard errors. Additionally, minor inaccuracies were observed in statistical reporting, including an incorrect denominator degrees-of-freedom listed in Table 1 (reported as 196, but correctly 199), mismatches in reported p-value thresholds (e.g., indicating $p < .001$ instead of $p \approx .0029$), inaccurate summaries of p-values (e.g., reporting all $ps > .35$ where some recalculated values were $< .35$), and rounding discrepancies for standard errors (e.g., reporting .03 instead of .02).

Finally, the reviewer noted limitations in auditability for certain ancillary descriptive analyses—specifically pilot study results, participant identification scores, and Cronbach's α

reliability coefficients—due to missing data. Although these omissions prevented independent verification of these descriptive elements, they are peripheral to the manuscript's core analyses and central claims. The primary findings regarding intergroup empathy bias under competition, its persistence even with reduced competitive threats (Experiments 3a and 3b), and its attenuation under conditions of reduced entitativity (Experiment 4), are generally unaffected by the errors and omissions detailed above.

Unresolved and unexamined aspects

Different reviewers bring different types of expertise, and different areas of focus, into ERROR reviews. It is therefore reasonable to expect that ERROR reviews will leave some aspects of a paper less examined than others. It is useful to acknowledge the potential for such issues so that ERROR recommendations do not artificially convey that they represent the final word on issues of error detection and correction for a given article.

First, as previously noted, pilot data, identification scores, and Cronbach's α reliability coefficients were not made available, leaving some descriptive values indeterminable in their accuracy.

In general, the reviewer paid most attention to the accuracy and reproducibility of the code and results reported by the authors. Less attention was paid to explicitly evaluating the broader methodological suitability or measurement validity underlying the manuscript's inferences. While no immediate issues were flagged, the appropriateness of the chosen measures and procedures to robustly support the reported conclusions was not systematically assessed. Similarly, the accuracy and completeness of citations were not examined in depth, leaving potential inaccuracies or omissions unaddressed. These areas were beyond the scope of the present review and thus remain open for further scrutiny. To be explicit, this is not to say that any of these areas contain errors; it is simply to highlight that they were not examined in this ERROR review.

I sincerely thank both Dr. Sørensen and Prof. Cikara again for their efforts during this process.

Jamie Cummins
Recommender for ERЯOR