Overview

The authors submitted a complete stage 2 manuscript which is entirely consistent with their plans at stage 1. There are no substantive changes between the stage 1 and stage 2 manuscript apart from entering the new values. Its refreshing to see replication results that are incredibly like the target study. It makes for a short discussion, but this is why there should be more replication efforts across areas of psychology. I like the feature of commenting on the prediction poll in the discussion and encouraging replication as a positive initiative.

I’ve split my comments into separate sections below on the manuscript, supplementary material, and reproducibility of OSF files. I have also labelled my comments as to address or suggestions. To address are key points to respond to while suggestions are more stylistic that you can ignore.

Manuscript

(suggestion) – Abstract – For the extension results towards the end, if there is space, it might be worth commenting on the key findings rather than the overall ANOVA. For instance, how forgiveness and apology was highest in the high empathy group.

(suggestion) – page 22 – I recommended reporting omega for internal consistency but one thing I did not notice in the stage 1 revision was just using the symbols. Given alpha and omega might not be obvious to all readers, it would be worth specifying in the first instance it represents Cronbach’s alpha and McDonald’s omega, then shortening to just the symbols beyond.

(to address) – Results and page 48 - The issue of following the target article’s analysis techniques regardless of appropriateness came up in the stage 1 report but its something I still think is worth reporting. In the discussion on page 48, the authors conclude there was no impact on the results as the findings were similar to the target article, but this is the wrong comparison as both could be equally wrong.

The good news is the results do seem to be robust to potential problems. I touch on the OSF materials below, but thanks to the authors sharing their data and code, I could quickly explore an example. For example, the relationship between apology and empathy looks problematic in figure 2 since there are many low values in both variables. However, using the performance package’s check\_model, it only looks like normality could be an issue (screenshot below). If you switch the correlation method to Spearman, the results are robust, with the correlation changing from .45 to .43.

Long story short, this is the comparison which should reassure the reader that the results are robust. It does not need to change the whole results section, just a note on how the results are robust to potential assumption problems while also honouring the target article’s analysis approach, and the authors can provide further details in the appendix given its more of an exploratory check.

A picture containing screenshot, text, diagram, line

Description automatically generated

(to address) – page 39 – In the manipulation check, the difference between low empathy and control is described as weaker but not specifically significant or non-significant. It might not be obvious to readers without seeing the output produced in R, but the p-value is literally .05. Its only a manipulation check, but it would be worth specifically commenting on whether you consider it significant or non-significant, since its technically statistically significant precisely on .05.

(suggestion) – page 40 – once you start reporting the manipulations, the reporting style changes to not include Cohen’s d, whereas it was reported for the manipulation check. Personally, I think it would be useful for readers to see the Cohen’s d reported too, but there is at least the mean difference already.

(suggestion) – page 41 – for the post-hoc tests, exact p-values are reported everywhere (bar p < .001) apart from this paragraph on perceived apology. I would recommend reporting exact p-values here too as a suggestion.

Supplementary appendix

(to address) – pages 3-5 – In the appendix, the handling outliers and exclusion criteria are still presented as future tense without any comments on the final data. Its presented as “if the replication is unsuccessful”, but it would be at least worth commenting that it was not performed as it was not deemed necessary.

OSF

I downloaded all the files from the OSF and I could again reproduce all the .Rmd files without edits, so great work here. Its been refreshing to review a paper where the code and data have been reproducible at all steps. The materials are also complete with PDF, Word, and Qualtrics file versions available.

In the interests of transparency, I sign my reviews.

Dr James Bartlett, University of Glasgow, UK.