# Reply to decision letter reviews #380:
# PCIRR-Stage 1 - McCullough et al. (1997) replication and extension

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response as well as a tally of all the changes that were made in the manuscript. For an easier overview of all the changes made, we also provide a summary of changes.

Please note that the editor's and reviewers' comments are in bold with our reply underneath in normal script.

**A track-changes comparison of the previous submission and the revised submission can be found on: https://draftable.com/compare/pBLzFJDIwGKB**

**A track-changes manuscript is provided with the file:**
**"PCIRR-S1-RNR-McCullough_etal_1997_rep_ext-main-manuscript-track-changes.docx"**

Summary of changes

Below we provide a table with a summary of the main changes to the manuscript and our response to the editor and reviewers:

| Section | Actions taken in the current manuscript |
|---|---|
| General | R2: We clarified the geographical origin of our participants in the abstract and added more information in the PCIRR-Study Design Table.<br>We added a note to better define “empathetic” and “not empathetic” in our survey. |
| Introduction | R2: We expanded and clarified our reasoning for the extensional design, and added some sources to justify our approach of empathy manipulation. |
| Methods | R1&R2: We completed the sentence at the end of the second paragraph.<br>R2: We added a sensitivity test for the post-hoc comparisons and gave justification on whether it is enough to detect our experimental effects.<br>We modified our replication and extension design table and added McDonald’s Omega for our scales. |
| Results | R2: We enlarged our Figures, conducted correlations comparisons for Hypothesis 2c using cocor R package, and completed exploratory analysis with our simulated data.<br>Based on the original paper, we modified our scale of empathy from 8 items to 4 items. |
| Discussion | N.A. |
| Reporting | N.A. |
| Supplementary materials | R2: We explained our approach on dealing with outliers in a more clear manner. |

*Note*. Ed = Editor, R1/R2/R3 = Reviewer 1/2/3

Sidenote: We note that we are not familiar with the titles and ranks of the reviewers, and looking for that information proves tricky. To try and err on the side of caution, we refer to all reviewers with the rank Dr./Prof. . We apologize for any possible misalignments and are happy to amend that in future correspondence.

# Response to Editor: Prof. Chris Chambers

> **I have now received three helpful reviews of your Stage 1 submission. All of the reviews are positive to varying extents. Shuqair is satisfied with the proposal in its current state, while Bartlett and Cao provide a range of constructive suggestions for improvement, with Bartlett highlighting many methodological issues that require further clarification or justification, and Cao questioning the rationale of the replication as well as the validity of key design elements.**
>
> **On the issue of the justification for the replication (and whether replicating McCullough et al 1997 is sufficiently important given more recent movement in the field), the Stage 1 criteria at PCI RR do not address the importance of the research question, only the scientific validity of the question criterion (1A). Overall, I judge that your research question meets the test for overall scientific validity, even if other researchers may judge the replication to be unnecessary. So while I am interested to read your response to this concern from the reviewer, and a thorough response (including revision) is likely to make the final paper more impactful, I don't consider this particular issue is to be roadblock to eventual in-principle acceptance.**
>
> **I hope you find the reviewers helpful and look forward to receiving your revised manuscript and response to the reviewers in due course.**

Thank you for the reviews, feedback, and the opportunity to revise. The feedback was extremely valuable, and we appreciate your and the reviewers’ time and support for our manuscript. We responded to and addressed each of the points made in detail.

# Response to Reviewer #1: Dr./Prof. Wenrui Cao

> **Many thanks for allowing me to read the Stage 1 Registered Report entitled "The impact of Empathy on Forgiveness: Replication and extensions of McCullough et al. (1997)'s Study 1". The authors tried to replicate the classic MuCullough et al. (1997)'s Study and to extend the study by manipulating empathy. I approached it with great personal interest and greatly appreciated the efforts made for reproducibility and replicability in the psychological sciences. Overall, the paper is well written and well structured, but there are still some concerns for this paper.**

Thank you for your time and thoughtful feedback.

> **While I strongly agree with the importance of McCullough's paper and the significant foundation it provides for the study and development of forgiveness. I doubt the contribution of replicating the study of McCullough et al. (1997).**
>
> **First, correlations or causal relationships between apology, empathy, and forgiveness have been tested by forgiveness researchers individually or together, and even meta-analyzed (e.g., Carlisle et al., 2012; Fehr et al., 2010; Konstam et al., 2001; Paleari et al., 2005). These efforts have to some extent replicated the work of McCullough et al. (1997), making a simply close replication does not seem necessary. I would rather suggest that the authors extend alternative meaningful models while partially replicating the study.**

Thank you for sharing your views.

The question of the value of this specific replication is tied to the broader question of the value of replications overall, especially given its impact and lack of direct well-powered pre-registered replications. Pre-registered direct replications are core to the scientific process and help update our knowledge regarding the target phenomenon, in terms of generalizability, effect size estimates, etc. A single study in a specific context should be considered as a first step in establishing a phenomenon. In our view, replications are not related to whether or not there are doubts regarding the effect, by now there are many examples of highly impactful phenomena that were considered beyond doubt that we have repeatedly failed to replicate in several large scale Replication Registered Report collaborations: "social" priming and ego depletion are to name a couple, but the list is very long. These had vast literatures with hundreds of articles summarized in comprehensive meta-analyses concluding support for the effects, only to later realize were biased literatures. Independent well-powered pre-registered/Registered Report replication efforts serve an important function that goes beyond conceptual replications and meta-analyses.

There are many challenges with the existing past literature, especially given that our literature suffers from publication bias towards positive and novel findings, and is based on underpowered studies that were not pre-registered and with no materials, data, and code shared to allow for error checking and reproducibility. There are some literatures where many of the reported effects come from the same group, with no independent replications. Beyond that, that a study has shown an effect on a small student sample in the 1990s does little to tell us how the effect would hold over time and for other samples.

Unfortunately, meta-analyses, especially those conducted before and covering literature before 2015 suffer from a host of fatal flaws, such as not addressing the issues of publication bias (e.g., ego deplication meta-analysis by Hagger et al., 2010), including studies that report impossible effects or fraudulent studies (e.g., hostile priming meta-anlaysis by DeCoster & Claypool, 2004, which includes studies by fraudster Diedrik Stapel, and by Srull and Wyer, 1977 who reported an impossible effect size of *d* = 5.7).

Therefore, we are in urgent need of independent well-powered (anti-publication bias) direct Replication Registered Reports to revisit our classics, to ensure that they reproduce and replicate, and to obtain a better estimate of the effect size, updating our knowledge of the phenomenon.

All that said, we did not restrict our investigation to a replication, and added several important extensions that would allow us to gain important insights regarding the phenomenon. From integrating additional measures such as revenge, to examining the possibility of causality with manipulation of empathy.

Replications and extensions allow us to both revisit what has been done before while testing new directions. Whatever we find, we gain new insights. If the replication succeeded, but the extension failed, we will know that our foundations are solid, but we need to rethink our extensions. If both the replication and the extension fail, we then realize that we need to update our priors regarding the overall phenomenon.

> **Second, as the field of forgiveness continues to evolve (based on McCullough and colleagues), the classification has become clearer, i.e., trait forgiveness (or so-called forgivingness; Brown, 2003) and state forgiveness (as in your case), and the term "forgiving" is mentioned less and less. I do understand that you use "forgiving" and "forgiveness" interchangeably through the manuscript, but this is not rigorous to the scientific development process. In addition, operationalizations of forgiveness measures have advanced, such as the Transgression-Related Interpersonal Motivation (TRIM), and the Enright Forgiveness Inventory (EFI) has been validated to have good internal consistency (Card, 2018). Indeed, the original source of the TRIM was Susan Wade Brown's doctoral dissertation at Fuller Seminary (Wade, 1990). It may not be appropriate to replicate the study simply using measures from two decades ago.**

In a replication study, we aim to repeat and replicate the classic research based on their original approach, including hypotheses, methods, stimuli, and analyses (Lebel et al., 2018). Replication studies are not meant to update or resolve any issues regarding the methodology of the target article, but rather focus on reproducing what the target did and assessing whether repeating the same methodology will lead to a similar outcome with a similar interpretation. This is a tricky point for replicators, because if we do not conduct a close replication using the target's and the replication fails, then critics can claim that our adjustments are what led to the failure. On the other hand, if we repeat the same methods exactly, and they fail, some may argue that it is the lack of updating that might be the cause for the failure. We chose to begin with a close replication, and, if that fails, we can update our priors and proceed to investigate the updates in the literature, if those still seem relevant.

Concerning your comments about the use of the words "forgiving" and "forgiveness", we recognize there may be different dimensions to study this concept, e.g., disposition, emotion, and in our initial submission we followed the use in the target article, which referred to the two terms interchangeably. To address this point, in our revision we standardized these to "forgiveness" as nowadays it seems to be the more commonly used term.

> **Third, as you wrote in the manuscript on Page 11, McCullough et al. (1997) conceptualized interpersonal forgiving as "the set motivational changes whereby one becomes (a) less motivated to retaliate against an offending relationship partner, (b) more motivated to maintain estrangement from the offender, and (c) more motivated towards conciliation and goodwill for the offender, despite the offender's hurtful actions", such that the increased conciliation motivation and reduced avoidance and revenge motivation is defined in the nature of forgiveness. So there is little meaning to examining the associations between forgiveness and these motivations (hypothesis 2). Instead, it may be more interesting to investigate real behaviors/actions.**

We understand the critic, though this is a critic of the target article and relevant literature. We closely followed the methods in the target article and are focusing on the empirical demonstration, even if there are some gaps in the theory and definitions, and updated in the literature. The core of our investigation is the replication, to ensure the findings are solid and hold, and to try and gain an updated and more precise estimate of the effects. Real behavior is indeed valuable, but far exceeds the scope of our aim at a replication, and is something for future studies to look into, once we successfully revisited the classics and increased our confidence regarding the replicability of the phenomenon.

> **I am also concerned about testing the mediation model with a manipulated mediator (i.e., empathy). Given that mediation analysis is inherently causal, that is IV predicts Mediator and DV, M predicts DV. The original model of McCullough et al. (1997) was based on a correlation design and their limitations in causal inference are unquestionable. However, I don't think a manipulated Mediator helps to establish the causality of the impact of perceived apology on empathy, let alone the mediation model in your Figure 10.**

Thank you for the note. This has helped us realize that we have not explained our planned analysis clearly enough.

When we referred to testing the mediation model, we were planning to conduct the analyses only on the control group, in which empathy is not manipulated. This was reflected in our accompanying code and the degrees of freedom as reported for the analyses. To be clearer about this, we now explicitly state "in the control condition" in the subsection "Exploratory analysis: Mediation analyses", and "Exploratory mediation analyses in the control condition." in the mediation figure.

We also noted in the manuscript that our planned mediation model is meant to be exploratory, and it is not meant to establish the causality of the impact of perceived apology on empathy. In our extension of manipulating recalled empathy, we aim to establish causality atleast for the link between empathy and forgiveness.

> **Other issues:**
> **The authors manipulated empathy with only one sentence "you were highly/not empathetic toward the person who had hurt you". I would like to know if this manipulation is valid in previous studies, if so, please cite their work, if not, a pilot study to justify this manipulation in necessary.**

Thank you for the feedback.

We aimed to use the baseline recall task in the target article, and to only manipulate the recall to be about a scenario involving a set level of empathy. We did not find a similar empathy manipulation of a recall task, and therefore constructed our own manipulation. Given that the focus of our investigation is the replication, and the extension is predefined as exploratory, we can consider this data collection as the pre-test for this suggested manipulation to inspire and inform future research. How a pretest would differ from what we are already doing here is not clear to us, given that any pre-test would also need all the same elements of a well-powered sample to detect the impact of the manipulation, a pre-registered plan to ensure confirmatory analyses, and given all that there is not reason to not go ahead a add all the dependent variables that we are testing here for little to no additional cost. Doing this as part of the Registered Report ensures adhering to open-science best practices, so that regardless of our findings, whether this works or not, future research can track all we did here and build on it to make additional attempts if deemed valuable.

We also note the following:

1. Our extension approach of manipulating elements of a recalled past event is aligned with the replication and follows commonly used methods in social psychology that study evaluations of emotion ladened situations. We previously implemented similar manipulations in recall tasks in various judgment and decision making replication projects (e.g., Chen et al., 2023; Yeung & Feldman, 2022), based on classic articles in the literature that have previously employed a manipulation of factors in the recalled scenario (e.g., Carter & Gilovich, 2012; Gilovich & Medvec, 1994).
2. We took the following measures to increase the chances of a successful manipulation:
    a. We enlarged, bolded, and underlined the caption to make it more apparent and used italics to draw attention to the relevant parts of the manipulation.
    b. We added a definition of empathy/no-empathy to the description, to make that clearer
3. We added comprehension questions after the replication dependent variables to allow us to check that participants understood the instructions of the manipulation and reacted accordingly.

References:

- Chen, E. Y., Chee, M. X., & Feldman, G. (2023). Revisiting the Differential Centrality of Experiential and Material Purchases to the Self: Replication and Extension of Carter and Gilovich (2012). *Collabra: Psychology*, 9(1). https://doi.org/10.1525/collabra.57785
- Yeung, S. K., & Feldman, G. (2022). Revisiting the Temporal Pattern of Regret in Action Versus Inaction: Replication of Gilovich and Medvec (1994) With Extensions Examining Responsibility. *Collabra: Psychology*, 8(1). https://doi.org/10.1525/collabra.37122

**The second line on Page 15 is unfinished.**

Thank you for alerting us to this oversight. The whole sentence has been revised and now reads:

> “A sensitivity analysis shows this sample is enough to detect correlations of *r* = 0.21, which is weaker than the lower bound of the weakest effect in the target article (apology vs empathy: *r* = 0.36, 95% Cl [0.24, 0.47]).”

## Response to Reviewer #2: Dr./Prof. James Bartlett

> **The authors present an interesting replication with detailed supplementary material. I think the rationale is convincing and potentially impactful, and the methods are replicable. I like the approach of using the Qualtrics function to generate responses to demonstrate how you will analyse the data. I was not aware of this before, so this will be useful to demonstrate code on how the eventual data will look.**
> **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.**

Thank you very much for the encouraging and positive opening note and the time invested and detailed feedback in helping us improve. It means a lot to us.

> **Manuscript**
>
> **1. (to address) abstract / methods – In the abstract, you refer to the sample as British Prolific, but American elsewhere. Please clarify/edit which is the best description.**

Thanks for catching that. That was an oversight, apologies.

We clarified and standardized it as "a US American Prolific".

> **2. (to address) PCIRR Study Design table – Under rationale for deciding the sensitivity of the test, this entry is quite vague. In the guidance, it focuses on the inferential criteria for supporting the hypotheses or not, so it would be worth outlining your choices for alpha, power etc.**

Thank you for urging us to do better.

We added more details to our power analysis subsection of the method section, with a sensitivity test for the replication and the extension, including specifying clearly the choices for alpha and power. We also updated the design table to be in-line with that section.

The summary of that section for the inferential criteria basically comes down to: "Alpha of 5% followed the target's, and high power of 95% is on par and higher than typical replications in PCIRR.".

**3. (to address) Introduction page 13 - In general I find the introduction informative and focused, but one section that would benefit from elaboration is the causal link extension. Is there precedent on how effective this manipulation can be on empathy?**

Thank you for the suggestion. We welcome the opportunity to elaborate further.

We added the following to the manuscript in the extensions section:

> Our main focus was the replication, with the extension added as an exploratory direction. Therefore, the extension is using the same recall method about the elicited past experience, and builds on top of that. Our aim with the extension was to manipulate the elicitation of recalled situations in which empathy has been experienced so that the person can reflect and evaluate other factors in that situation. Therefore, the manipulation is of the recalled past experience and not the empathy that the participant is experiencing while taking part in the experiment. This is different from some of the research that tried to manipulate empathy through a perspective-taking approach for emotions experienced during the experiment., in which participants were asked to remain objective (vs emotionally-attached) to the main character when reading a scenario (Berenguer, 2007, 2010).
>
> Our extension approach of manipulating elements of a recalled past event is therefore aligned with the replication and follows commonly used methods in social psychology that study evaluations of emotion ladened situations. We previously implemented similar manipulations in recall tasks in various judgment and decision making replication projects (e.g., Chen et al., 2023; Yeung & Feldman, 2022), both based on classic articles in the literature that have previously employed a manipulation of factors in the recalled scenario (e.g., Carter & Gilovich, 2012; Gilovich & Medvec, 1994).

**4. (to address) – Method page 15 – the first sentence ends suddenly when outlining the r = .21 effect. Make sure this is explained fully. As an aside, r = .21 is also close to the lower bound (r = .24) of the confidence interval for the replication r = .36 effect in the OSF files. It might be worth mentioning this as additional justification for the sample size and smallest effect size of interest.**

Yes, great point and suggestion. Much appreciated.

We added a comparison between the result of our sensitivity test and the lower bound of our 95% confidence interval of the weakest significant effect in the target article. The sensitivity analysis suggested that the sample was sufficient to detect correlations of *r* = 0.21, which is weaker than the lower bound of weakest effect in the target article (apology vs empathy: *r* = 0.36, 95% Cl [0.24, 0.47]).

**5. (suggestion) – Method page 15 – The extension is ultimately exploratory, but it would be interesting to see a comment on the smallest effect size of interest. For the pairwise comparisons, you would essentially have 95% power for effects of d = 0.33, so you could comment on whether this is useful or not in relation to past research on these kinds of manipulations.**

Thank you for the suggestion.

We added more details about the sensitivity of our pairwise comparisons in our main manuscript. The post hoc comparisons would be enough to detect the effect of *d* = 0.33 (95% power, alpha = 5%, two-tail), which corresponds to a medium effect in social psychology research (Lovakov & Agadullina, 2021). Based on our previous experience of differences in recall tasks in our other replication projects generally revealed medium to very strong effects (Chen et al., 2023; Feldman et al., 2016; Yeung & Feldman, 2022). Thus, we believe aiming for medium effects in our recall task project is justifiable. We added those details to the power analysis subsection in the methods section.

**6. (suggestion) – Table 3 page 16 – For someone unfamiliar with McCullough et al., it would be worth adding the geographical origin of their participants to complement specifying your sample will be US American students.**

We suspect McCullough et al. (1997) were using a US American sample since all authors were working in the US at that time, but it was not explicitly mentioned in the target article, so we added that as a note in the table. We added it as a note:

> "Origin was not explicitly mentioned in the target article, though we suspect it was US American, given the authors' affiliation at the time."

**7. (to address) – Table 4 page 17 and measures throughout – The source of the scales is provided for some entries, but not all. For completeness, can you add the source for each entry?**

Thank you for the suggestion. We did not include those given that these were our target articles, yet we agree that spelling that out would make things clearer.

We added the source (i.e., McCullough, 1997/8) to each of our scales.

**8. (to address) – Measures throughout page 20 onwards – The authors report alpha for the scales which has some limitations as it rests on stringent assumptions (Dunn et al., 2014). Can you either add 95%**

> **confidence intervals to the alpha values or consider reporting Omega as Dunn et al. suggest?**

Thank you. To both align with the replication’s methods and address your request, we now include both the Cronbach's alpha and McDonald's Omega for all our measures, in both the analysis code and our reporting in the main manuscript.

> **9. (to address) – Data analysis page 27 – In the manuscript and supplementary appendix, you explain you will not check for outliers but plan on using Pearson’s r as a parametric test. Parametric tests are generally robust but outliers can be one of the biggest problems (Knief & Forstmeier, 2021), so how do you plan on analysing the data if there are problems with outliers, like switching to a non-parametric or robust test?**

Thank you for raising this concern. This is a good point.

We initially were aiming to closely follow what the target article reported doing, and they did not report any such analyses, yet we see the value of adding additional exploratory analyses in case we failed to find support without outlier exclusions.

We added further details in our supplementary under the section “Handling outliers: Strategy”:

> “In this study, all the data would be included for analyses, so we would not identify outliers.
>
> Yet, we recognized parametric tests are highly sensitive to extreme value. Thus, we would compute a composite outlier score (see the 'check_outliers' function in the 'performance' R package; Lüdecke et al., 2021) via the joint application of multiple outliers detection algorithms (Z-scores, Iglewicz, 1993; Interquartile range (IQR); Mahalanobis distance, Cabana, 2019; Robust Mahalanobis distance, Gnanadesikan and Kettenring, 1972; Minimum Covariance Determinant, Leys et al., 2018; Invariant Coordinate Selection, Archimbaud et al., 2018; OPTICS, Ankerst et al., 1999; Isolation Forest, Liu et al. 2008; and Local Outlier Factor, Breunig et al., 2000). This composite outlier score would only classify data as outliers by at least half of the methods used.
>
> If we fail to find any support for our hypotheses, we will run statistical analyses after excluding the “outliers” with compensation for alpha (i.e. alpha = .01) to account for multiple tests. But this part of our analyses would be purely exploratory since it would deviate from our initial study protocol. “

**10. (to address) – Results page 29 and 34 – The correlations include 16 tests and the ANOVAs includes two DVs and pairwise comparisons. Do you plan on correcting for multiple comparisons and if not, why?**

In a replication study, we tend to strictly follow the data analysis approach in the target article, McCullough et al (1997) did not mention that they made any correction for their tests, so we tended to follow that. Given previous experience with comparing targets to replications, we would rather not get into the sensitivities of what it would mean to uphold the replication to a stricter criteria than the target article, though we agree that weaknesses in the target need to be discussed. Also, our replication evaluation mostly depends on the comparison of effect sizes and CIs of the target and replication.

We understand and agree that not adjusting alpha for multiple comparisons is suboptimal. We therefore added a planned discussion of this point in our Stage 2 discussion as a limitation.

**11. (suggestion) – Results Figures 2-9 – In general, I like the ggstatsplots to display the distribution of variables, but the text becomes very small when you have multiple plots like Figures 4-6. They also have information you have not presented or justified in the manuscript like Bayes factors for the correlations, so it might be worth breaking them down and presenting them to the reader where they can easily view them.**

Thank you, we were planning on making all those formatting adjustments in Stage 2 after the data collection, but we agree that it is better to already implement everything to make things clearer and easier to follow and understand.

We enlarged all figure fonts, enlarged their display, and placed those in separate pages.

**12. (to address) – Results page 30 – To address hypothesis 2c, it sounds like you want to compare the magnitude of two correlations. Its not clear how you plan on doing this despite saying they don’t differ significantly for the mock data, so it would be useful to explain your planned statistical methods here like Fisher’s z test.**

Thank you for your suggestion.

We agree that it would be best to address the H2c analysis already at Stage 1, and therefore decided to employ the “cocor” R package to compare the correlations. This package is powerful and comprehensive that it makes comparisons for overlapping correlations from dependent groups using up to 10 approaches (i.e., Dunn & Clark, 1969; Hendrickson et al.,  1959; Hittner et al., 2003; Hotelling, 1940; Meng, 1992;  Olkin, 1967; Pearson & Filon, 1898; Steiger, 1980; Williams, 1959; Zou, 2007).

We included some more details under the section “data analysis strategy".

**13. (suggestion) – Table 9 page 32 – For comparing the replication to the original findings, it would be helpful to include the replication effect and 95% CI as an additional column. Otherwise, readers would need to scroll back to the table where you reproduce the results.**

We tried to address your concern by incorporating the results from the original findings into Table 9. We also made adjustments to Table 8 to bring it close to Table 2 reporting the target's correlations, to allow for an easier comparison.

**14. (to address) – Exploratory analysis page 36 – the final sentence is left as […] whereas the other results are presented with the mock data. It would be worth including the text you plan on including when you have the real data like you did for other sections.**

Thank you for your suggestion.

We expanded and added back a simulated data analysis in the section of exploratory analysis with our simulated data.

**Supplementary appendix**

**1. (to address) Exclusion criteria page 4 – You say you will focus on the full sample but its not clear how this applies to your general criteria here. Will you analyse all the data first, including those who complete the study multiple times or do not provide consent?**

Thank you for raising that. We agree that we did not do a good enough job explaining this, and could have been much clearer.

We use two main methods for preventing duplicate responses: 1) CloudResearch platform, 2) Qualtrics measures. In both cases, duplicates are blocked, and not allowed to proceed to the survey. Those can be seen that in the survey flow in the Qualtrics survey, employing the Qualtrics anti spam measures that we mentioned in the "Participants" subsection ("fraud and spam prevention measures: reCAPTCHA, prevent multiple submissions, prevent ballotstuffing, bot detection, security scan monitor, and relevantID").

The same holds for those who do not consent or fail the initial verifications, they are not allowed to proceed to take the survey. You can see that in the survey flow, or test that in the Qualtrics preview we provided (https://hku.au1.qualtrics.com/jfe/preview/previewId/d52a8690-3476-4219-ae46-f7076b96a388/SV_0VeUxDaZu96QUfk?Q_CHL=preview&Q_SurveyVersionID=current ).

The one tricky bit that we did not explain in the first submission refers to what happens when participants do not complete the survey and drop out in the middle. In this case we specified that we will not include those results, even if they might hold some data. We will only include those who completed the study.

Therefore, in our study the sample refers to all the participants who started and finished the whole survey once, based on the currently implemented measures in CloudResearch and Qualtrics.

We updated that in the supplementary material's "Exclusion criteria" subsection:

> We will run our analyses on the full sample of all those who completed the study successfully and answered all questions. Those who dropped out will not be included.
>
> In the case of a failed replication, as a supplementary analysis and to examine any potential issues, we will also determine further findings reports with exclusions. In any case, we will report exclusions in detail with results for the full sample and results following exclusions (in either the manuscript or the supplementary. Criteria:
>
> 1. Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale)
> 2. Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale).
> 3. Participants who indicated issues or having seen these materials before in the funneling section (manually coded).

**2. (to address) Additional analyses page 6 – This is also relevant to the results and your data analysis plan. You mention equivalence testing here but nowhere in the manuscript, and this section is currently blank. I can see how you plan on interpreting the correlation results in the absence or presence of an effect with Table 9 and LeBel et al. graphs in the supplementary, but for the ANOVAs and mediation results, its not clear whether you are just interested in significant vs non-significant effects, or whether you have a smallest effect size of interest to incorporate into equivalence testing.**

Thank you for your question. The section you pointed to was a template placeholder in case in Stage 2 we added exploratory analyses that include equivalence testing. We removed it in this section.

We noted the issues with the mediation analyses, including that as to get closer to what the target article did, and classified those as exploratory. We will be comparing the target to the replication on the main effect correlations using the LeBel et al. (2019) criteria, but did not plan any additional analyses that pertain to the mediation analysis.

**<u>OSF</u>**

**I downloaded all the files from the OSF and I could reproduce all the .Rmd files without edits, so great work here. The materials are also complete with**

**PDF, Word, and Qualtrics file versions available.**

We are very grateful that you attempted that, that is still a rare practice in peer review, and so we are very grateful. Thank you for your time and valuable detailed feedback.

## Response to Reviewer #3: Dr./Prof. Saleh Shuqair

> **Thank you for inviting me to review this pre-registration (replication study on the link between empathy and forgiveness in McCullough et al. (1997)'s Study 1.)**
> **Overall, the process is explained thoroughly.**
> **· I agree with the authors that to date, there are currently no published independent direct replications of this article. McCullough et al. (1998), thus, i believe that extending their model by adding other variables such as commitment, the impact of the offense, and rumination, into predicting forgiveness would provide a decent contribution to the literature.**
> **· The authors provide a sufficient justification for replicating the study, the research design is logical and addresses the coherence and credibility of the hypotheses**
> **· The introduction provides a good rationale and justification for the study,**
> **· The sample size in the current study aimed to recruit 800, which is a well-powered sample size to predict this effect.**
> **· The study procedures and analyses, and the experimental conditions are appropriate the scales also are appropriate.**
> **Best of luck with your research.**

Thank you very much for the positive supportive feedback. It is very encouraging and very much appreciated!