Reply to 2nd RNR decision letter reviews #609: Kahneman and Tversky (1973) replication and extension

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response to each item. We also provide a summary table 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/HYCoBubBEkwm

A track-changes manuscript is provided with the file: PCIRR-RNR2-Kahneman-Tversky-1973-replication-main-manuscript-track-changes.docx (https://osf.io/nge9z)

[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.]

Reply to Editor: Dr./Prof. Rima-Maria Rahal

> **I have now received three re-reviews. There are few remaining issues to integrate:**
>
> **Ensure that the target sample size is clear**
> **Consider expanding the discussion on the role that replicating this particular set of studies plays for the existing literature in the area**
>
> **Please consider this point in a final revision and response, and we will then be ready to move forward with Stage 1 IPA.**

Thank you for feedback obtained, and the invitation to revise and resubmit.

Reply to Reviewer #1: Dr./Prof. Peter White

Thank you for the positive and constructive feedback.

> **The authors' replies go through my comments one by one so here are corresponding replies to that.**
>
> **.1. I am happy with the revisions made in response to this.**
>
> **.2 I perhaps didn't express myself very well there. I wasn't trying to argue that there wasn't any need or value in replicating the studies, only that the eventual publication should have a sound justification for it, which means setting it in the context of subsequent developments in that research area. If there is real doubt as to whether the findings would replicate or not, then certainly the replication attempt is justified, and the subsequent literature does seem to justify feeling some doubt about it. I think I just wanted the authors to have a clear idea about what it would mean if the findings were replicated, and also what it would mean if they were not replicated.**

Understood. Thank you for clarifying, we agree.

We will be discussing implications of the findings and directions for future research in Stage 2. We do note that we will generally try to be careful and humble regarding the contribution of a single replication to say anything definite about this literature and the ongoing debates, but we hope that with full reproducibility of the materials with accompanying data and analysis code many other will follow to retest and expand on these so as a community we can together come to a better understanding of these findings.

For now, in Stage 1, we focus on the empirical aspects of reproducing what was done and testing replicability.

> **.3 The authors' reply is satisfactory - I just wanted to be sure that readers of the eventual publication would get a clear understanding from the paper of why the replication matters.**
>
> **.4 O.K.**
>
> **.5 I don't have any idea of a measure of confidence that would be trustworthy but I stand by my original comment that explicit judgments of confidence are prone to response biases. Replicability is not a guide to trustworthiness because it might mean only that the same response biases were operating in both the original and the replication. I'm not saying the**

> **authors shouldn't obtain confidence judgments, just that they should perhaps include some nuanced discussion of the results when they get them. What they have said in their reply to my comment is the right sort of thing, in my view.**

Understood, thank you. We agree. We added the following planned discussion for Stage 2:

> [Planned for Stage 2: Following Dr./Prof. Peter Anthony White's comment regarding the self-reported measure of confidence, we will discuss challenges, our findings, and future directions. In our peer review exchange we wrote the following as an initial base: "We conducted two large scale pre-registered replications, on MTurk (using CloudResearch) and Prolific, and both concluded very similar findings 45 years later (pre-registrations, materials, data, code, and reports available on https://doi.org/10.17605/OSF.IO/C3YVK). [...] To combat the possible explanation of this being a self-presentation bias, there are studies that show under-confidence in some studies with a similar methodology. The literature is nicely summarized in a recent book by Don Moore (2020) "Perfectly Confident: How to Calibrate Your Decisions Wisely".]

> **.6 O.K.**

> **.7 In the original submission it reads "scores supposedly either representing academic achievement, mental concentration, and sense of humour". The two choices are (i) remove "either" and (ii) change "and" to "or". The authors can go for whichever of those they prefer. Apologies for being a bit pedantic about this.**

We appreciate this feedback, thank you. Changed to:

> In Study 5, participants were given input scores supposedly representing academic achievement, mental concentration, or sense of humor of ten students (between-subjects), then they gave predictions about their GPA.

> **.8 O.K.**

> **.9 O.K., that is very useful. The research I do pretty much has to be done face-to-face so I have never explored online alternatives. The use of Prolific is probably more common in some areas of psychology than others.**

> **.10 It is the "If things fail..." that concerns me. A simple way to deal with the problem would be to analyse separately for each experiment the data from the participants for whom it was the first one they saw - at that point their judgments could not be affected by the other studies because they**

> **haven't done them yet. If the results for that sub-sample resemble those for the full sample, then no problem.**

We adjusted accordingly.

Our aim was to address the main concerns regarding decision flexibility and interpretability. In your specific suggestion here, we note that in running only the sample in which each study was displayed we lose a lot of power, and we therefore see these as exploratory analyses.

We therefore reiterate that our interpretation of the replication success is only based on the full higher-powered sample. We consider the order effect "exploratory", and we addressed multiple analyses by adjusting the alpha. We appreciate the view that these analyses are potentially of interest regardless of the outcomes, even if there are issues of complexity and interpretability, and so we adjusted accordingly. We also added an explanation of what "moderator analyses" would look like based on your suggestion:

> We, therefore, pre-registered that we would examine order as an exploratory moderator, meaning that we will run the analyses first with the study displayed first and then with the study not displayed first, and report the differences between the two, and examine whether the confidence intervals of the effect overlap. To compensate for multiple comparisons and the increased likelihood of capitalizing on chance, we set the alpha for the additional analyses to a stricter .001.

and under "Design: Replication and Extension":

> [Note: We will test for order effects, with each study when it is displayed first. See "data analysis strategy" section.]

We also added the following to the planned Stage 2 discussion:

> [Planned for Stage 2: Following Dr./Prof. Peter Anthony White we will discuss the potential weaknesses and strengths of the unified design with our collecting all studies with the same sample in a with-in design, and our exploratory order analyses.]

> **.11 O.K.**
>
> **.12 O.K.**
>
> **.13 I sympathise with the authors and I think their discussion of the issue is intelligent and appropriate. I agree with the decision to set alpha at .001 for exploratory analyses. For the analyses where .005 is used, I would suggest that, if they get results significant at .01 but not at .005, they could discuss these or at least list them, so that readers could get a feel for whether there**

> **is any likelihood of type 2 errors, but I'm happy for the authors to go with their own judgment on this.**

Thank you, great suggestion. We added the following:

> [Planned for Stage 2: Following Dr./Prof. Peter Anthony White's comment: "I would suggest that, if they get results significant at .01 but not at .005, they could discuss these or at least list them, so that readers could get a feel for whether there is any likelihood of type 2 errors". We will discuss replications of problems in which the findings fell in between .05 and our set alpha (.005/.001)]

> **.14 O.K.**

> **Overall. I would like the authors to bear my comment on .4 in mind when writing up the results. The grammatical error commented on in .7 should be corrected. The authors should consider the suggestion for further analysis in .10 but I will leave it to them to decide whether to do it or not. They should also consider the suggestion made in .13 but again it's up to the whether they do it or not. I have no further requests for changes.**

Thank you for the feedback, much appreciated!

Reply to Reviewer #2: Dr./Prof. Regis Kakinohana

> **The authors addressed all my comments with detailed responses and adjustments to the manuscript. Therefore, I have no further questions or suggestions regarding Stage 1.**

Thank you for all the positive and constructive feedback.

Reply to Reviewer #3: Dr./Prof. Naseem Dillman-Hasso

> **Thank you again for the opportunity to review this RR. I greatly appreciate the authors' detailed consideration of feedback and critiques from all reviewers, with the goal of making this project stronger. I have almost no comments for this round. I believe this project is ready for data collection.**

Thank you for the positive and constructive feedback.

> **Follow ups to first round comments:**
>
> **.1. Regarding the design of the replication (i.e., all studies run by each participant), I appreciate the overview of the work your team and others has conducted. I stand corrected in terms of what the previous research shows, but I still confess that I do not like the design, but I struggle with a reasonable objection at this point besides "it's just not what I like." Given that, and the overwhelming evidence, I accept the current design.**

We understand, and appreciate that. Once data is collected, we aim to do a series of exploratory analyses to examine possible implications.

> **.2.**
>
> - **I appreciate the restructuring of the manuscript, I believe it reads much better now.**
>
> - **Regarding using 99 as a missing data code, I still stand by my point: I would use NA for missing values, as opposed to 99. It may seem implausible for you, but given the commitment to sharing datasets publicly for reuse, it is important to consider that others may not see 99 as implausible, or may not fully read through data dictionaries. All missing values in my opinion should be replaced with "NA," I offered 999 as an alternative option if the authors would prefer to use a specific value. It's not our job to determine what is a reasonable age is, as you hinted in the response. Additionally, using NA for missing values makes many descriptive and analytic functions in R easier to implement for others who wish to review your datasets (i.e., mean(df$age, na.rm = TRUE) requires no additional coding of missing values).**
>
> - **Regarding .sav/.csv files, I understand the difficulties with CSV files and that .sav files solve many of those issues. I would still prefer a .csv file, given that individuals without SPSS or PSPP would be unable to open a .sav file**

on their computer easily without importing it through another software, but given the other benefits of .sav files, I am fine with this.

We understand. These are rather minor technical preferences. To address any such gaps, we make all data and reproducible Rmarkdown code with outputs openly available.

Second minor round comments:

.3. Clarify the "minimum sample of 800 participants": under what conditions will it be different? If you collect more than 800 participants, will extra data be discarded, or still used?

Still used. We set the target of 800 participants on Prolific, but will analyze whatever participants have completed the survey through that data collection. There are various issues such as some participants timing out, after which we reward them for their participation even though they were not counted towards the 800 Prolific set out. We owe it to them and to our stakeholders whose funding was used to pay them to include them. To make it clear and reiterate - we only look at the data once data collection has been completed, and include all completed surveys in our analyses.

.4. I would recommend updating to R version 4.4.0 and using that, given a recent major vulnerability reported: https://nvd.nist.gov/vuln/detail/CVE-2024-27322; https://hiddenlayer.com/research/r-bitrary-code-execution/

Yes, thank you.

This is a Stage 1 Registered Report, and we will update everything we do and the tools we used in Stage 2 after data collection in the manuscript.