Reply to 2nd round PCIRR decision letter reviews #496:

Norton et al. (2007) replication and extensions

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. 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/vTgCHrZVIKns

A track-changes manuscript is provided with the file: PCIRR-S1-RNR2-Norton-etal-2007-rep-ext-main-manuscript-trackchanges.docx (https://osf.io/v3ze2)

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	R1: PCIRR-Study Design Table- sample column updated to remove reference to potential exclusions (p 5).
Methods	R1: Additional clarification of our stopping rule and management of additional recruited responses now included (p 19). Minimum required sample size and power and sensitivity analyses have been amended (pp 18-19). References to exclusions have been redacted from the main manuscript, and we have clarified that we will not classify outliers or exclude any complete responses from our dataset (p 19, p 26). As studies 1a and 1b have now been separated for randomisation of order in the Qualtrics script, they are now separated in the procedure (pp 22-23).
Supplementary materials	R1: Qualtrics script has now been updated to randomize order of all four studies (1a, 1b, 2 and 4). Tables S1 and S2 now merged and updated.

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

[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. Yuki Yamada

> **Thank you so much for submitting your revised manuscript.**
>
> **Two reviewers have checked it again, and as you can see, one reviewer is satisfied with the revision.**
>
> **Another reviewer is still commenting on the data analysis and power analysis. I hope the authors would address this again.**

Thank you for the reviews obtained, your feedback, and the 2nd invitation to revise and resubmit. We received important feedback and have worked to address all comments.

There is one point that we discuss with Reviewer 1 where we further explain our view and analytical strategy with a request for your consideration.

Reply to Reviewer #1: Dr./Prof. Zoltan Kekecs

> **.1. I don't find most of the arguments by the authors very convincing in why a joint design is needed other than increased sample size. The additional insight in the exploratory research questions is nice, but in a confirmatory RR, these are secondary to being able to adequately address the main effect.**
>
> **However, I am still satisfied with the action the authors took with one small request.**

Thank you for your feedback and suggestions.

> **.2. The authors say that "We therefore pre-register that if we fail to find support for our hypotheses that we rerun exploratory analyses for the failed study by focusing on the participants that completed that study first, and examine order as a moderator. "**
>
> **Maybe the authors misunderstood my comment to mean that I was afraid that the effect would be masked by combining the studies. On the contrary, I am afraid that the effect would only be due to running the studies together. I would like to ask that they re-run the analysis regardless whether the main hypothesis was confirmed or not by focusing on the participants that completed that study first. This way it will be revealed if the effect is only due to combining the studies.**

Response: Thank you for providing further clarification and for your helpful suggestions for addressing potential order effects.

Regarding your request, we please ask for your and the editor's understanding, as we strongly prefer to proceed with our initial plan, and not adjust our data analysis strategy. We will detail our reasons below.

Note to the editor: If you still feel unconvinced by these arguments we will amend given clear editorial guidelines.

We will first reiterate our previous reply, our experience, and our strategy: As noted in our manuscript, this is not the first time we are doing this kind of experimental design combining different studies, also implemented in more than dozen PCIRR replication projects. We have taken many steps to ensure that this would not impact things, such as in the randomization of order, and in our planned analyses. In our many previous replications, order has not been an

issue, and this has been confirmed again with data collections in four PCIRR Stage 1 projects we collected data for this year (one of them with a replication of seven problems reviewed in Read, 1999). For example, some of the most comprehensive demonstrations of that were with our replication of review articles where we ran many studies of the same phenomenon (e.g., 17 effects in our PCIRR replication of Thaler, 1999, in https://osf.io/4ca98, or our replication of nine problems in Kahneman and Tversky, 1972, in https://osf.io/nhqc4/, both concluded as successful), with no indication of order effects. It would be somewhat disappointing for robustness and generalizability of these phenomena if something like order or context would impact these findings. Overall, we (CORE team, 2024) have concluded over 120 replications of heuristics and biases in judgment in decision-making, with dozens using this design, all of them conducted online and with many of them combining many problems into a single unified design, summarizing very high replicability rates, and no indication of order effects. We mentioned some of the completed projects in our methods section: Petrov et al., 2023; Vonasch et al., 2023; Yeung & Feldman, 2022; Zhu & Feldman, 2023; two of those were Registered Reports with Stage 2 endorsed by PCIRR, and there are many more.

Why not run those analyses anyways? Running all these analyses regardless of the results involves at least doubling the number of analyses, decreasing power, increasing decision flexibility, and impacting readability and interpretability, when it is not clear what the added benefits are. It is not about effort/work in running those, these are very straightforward analyses to run, but rather the issue is what to do with those once we run them. Consider the following as some of the examples:

- What to conclude if a study is supported (with a signal) in the higher powered full sample but is not supported when only analyzing it as the first study? What to do in the opposite situation when it is supported as the first study but not with the better powered full sample?
- With 4 studies in random order, there are many ways to analyze such a moderator: is it the positioning in the 4? Is it which of the specific studies came before it?

To try and give some perspective to this issue, it might be helpful to compare those to other deviations in our replication from the original. One could argue that the participants' age, education, or time (time of day, day of the week, month of the year, or season, etc.) are influencing factors, and we should therefore always test age, education, and time as moderators, and also rerun the main analyses focusing only on the same time (whichever definition), education background, or age group, as in the original. These may or may not be influencing factors-we do not know, and have no reason to suspect they would be, as these factors are not related to the core generalizable theory. Each possible analysis for each one of these factors adds

another forking path of complexity and impacts our ability to interpret the results. Especially given our experience with order, we see no reason to suspect in advance that these would matter.

We therefore thought - given the strengths of the unified design - how to best reassure reviewers while minimizing flexibility and maximizing utility? The added benefits of running these analyses for replications are especially valuable when the moderator may potentially cause the effect to go from signal to no-signal. If the effects replicate well, then there is little benefit in the additional analyses, these are not the core of the theory nor are these the main goals for the replication, just in the same way that the replication does not focus on age, education, or time. If someone were to later want to look at those factors, we make everything available and they are welcome to run additional analyses. However, when effects do not replicate well, we can conduct a series of exploratory suggestive analyses, acknowledging limitations of the analyses (such as lower power).

[We lastly note that in this revision we made an adjustment to our Qualtrics' survey flow to also fully randomize the order of Study 1a and Study 1b. Previously we presented Study 1b only after presenting Study 1a, and after having considered this feedback we now see the value in ensuring a fully randomized order that would ensure order is controlled for across all studies, and can be analyzed later, if needed.]

.3. Reply to Response #7:

This reviewer note was about target sample size. The authors say that they intend to analyze all valid cases, and say that they "see no reason to worry about or suspect optional stopping". Nobody is "worried about optional stopping" before they started collecting data for their own study. Everyone is the hero in their own life's story. Nevertheless, having clear stopping rules and pre-specified analyzed sample size targets still make sense to prevent conscious or unconscious biases in research. The study is already well powered, with considerable slack. I still suggest that the authors only analyze the data from the first 800 valid responses in their confirmatory analyses. In the exploratory, anything goes.

Response: We realized that there might still be a misunderstanding regarding the process here, so we will first make some clarifications and then try to further strengthen our plan.

First, there is no exploration, there is no flexibility, and there is no room for conscious or unconscious bias. The criteria is very clear and strict, and we are transparent about every step of the process. We set the target sample size in Prolific, and whatever comes from that single data

collection - we analyze. This is only one point at which we analyze data, and at no point in time do we analyze the data before the final data collection.

The stopping rule is set within Prolific to 800 participants, in accordance with our criteria. However, and this perhaps the point of misunderstanding - from personal experience we are aware that Prolific occasionally recruits additional participants beyond the specified sample size, due to incorrectly classifying completed responses as incomplete (either 'timed out' or 'returned', after which they message us and we mark as successfully completed). Nonetheless, Prolific will stop recruitment automatically, and is not influenced in any way by us, the researchers. We owe it to the participants and our stakeholders to include all valid and paid-for participants' responses in our data analysis. Therefore, to align with our pre-registered data analysis procedures, we will not exclude any complete valid responses that are obtained from Prolific.

Action: We added a clarification of stopping rule is provided on manuscript page 19 (footnote 4), as follows:

> "In some instances, Prolific recruits participants beyond the specified sample size. This is due to the platform sometimes incorrectly classifying valid completed responses as 'timed out' or 'returned'. We will not exclude any complete valid responses from our dataset, and will include any additional completed responses obtained from Prolific."

> **.4. Relatedly, I would also like to ask the authors to specify the exclusion criteria from the analysis. I could not find now in the manuscript what are the planned exclusion criteria, although 10% exclusion is accounted for in the sample size rationale.**

Response: Thank you, you are correct, mentioning exclusions was an oversight. We amended the manuscript to clarify that we will not classify outliers.

Action: We updated the PCIRR-Study Design Table to remove references to potential exclusions.

References to exclusions redacted from 'Power and sensitivity analyses' section on page 18.

Following statement added under 'Data analysis strategy' section:

> "We did not classify outliers in this study. All data from participants who successfully completed the survey were included."

.5. Reply to Response #8:

This note was related to the power to detect all effects, if you have multiple tests and plan for 90% power to detect each effect separately. The authors reply that this is not common practice, and that the community on X was also divided.

Most of the detailed responses you got on X seemed to agree with the point. (Others seemed to misinterpret the question and responded about alpha adjustment, which is not really an issue here).

The important thing is that this is a mathematical necessity. You can calculate this on a napkin, or in R. Simply run a simulation of a study having two effects (with the same effect size and being independent of each other to simplify things), and a sample size powered to be 90% powerful to detect any one of these effects. When you look at how many times you were able to detect both effects, you will find that the probability is 81%. As some posters on X point out, this 81% is a "worst case scenario", because if the effects do correlate, you will have a correspondence of when you are able to find them, so you power to detect all effect will be closer to the individually calculated powers.

Here is a simple simulation showing the issue: we are simulating 5 effects independent from each other, with a sample size enough to detect each effect 90% of the times. However, in any study, there is only about 59% chance for all of the 5 effects to be significant: https://github.com/kekecsz/power_to_detect_all/blob/main/power_to_detect_all.R

All I am saying is, that in a study where power is set to 90% to detect each effect, the study will have a lower chance to detect all effects in the study.

"Note: We would be happy to revise given clear editorial guidelines and instructions on what to amend. If the reviewer or editor feel that an adjustment in sample target is needed - then we ask that you please provide us with relevant citations and an example or two of other Registered Reports (preferably PCIRR, preferably replications) that has done something similar, and taking into consideration cost/benefit of going beyond the already large planned sample of 800."

– I find this request unnecessary. This mathematical fact does not require citation in my view, since it is easy to demonstrate (see code above)(although the responses on X did contain some useful works if you are interested).

I suggest the authors add a paragraph in the power analysis section, that says something like this:
"It is worth noting that even though the power for this study to detect each hypothesized effect is at least 90%, the power of this study to detect all of these effects simultaneously is unknown."

(If you don't like "unknown", here you can give the worst-case scenario estimate as I mentioned above, or, if you have reliavble pilot data, calculate the tue power based on the dependency of the effects from there. For all the effects in this study with various effect sizes this might b a complicated calculation, maybe easiest to do with simulation).

Response: Thank you for further clarification of your concerns, and excellent suggestions on how we may address this. We have amended the main manuscript to include a statement regarding this matter. We have updated the statistical power to 80%, for reasons provided in our response to comment 6.

Action: Statement added under 'Power and sensitivity analyses' (page 18, footnote 3):

> "Although the power for this study to detect each hypothesized effect is at least 80%, the power of this study to detect all of these effects simultaneously may be lower."

> **.6. Relatedly, the authors say in this sentence: "We conducted a series of a priori power analyses based on these effect sizes and we found that 234 participants would be enough to detect the effect sizes with 90% statistical power at alpha = .05 (see supplementary materials and analysis code for more details)." I don't understand why the authors say 234. In the PCIRR Study Design Table they say "Based on the reported correlations between knowledge, similarity, and liking (Study 3 in Norton et al., 2007), we conducted a power analysis. It revealed that N = 310 and 400 would achieve statistical power of 80% and 90% respectively to detect the interaction effect." Shouldn't the authors have used 400 instead of 234?**

Response: Thank you for catching that. It has helped us identify a broad oversight in our power analysis. We asked for additional external feedback and revamped our power analysis altogether, addressing additional issues. None of those affects the calculated sample size, as the target sample size we initially planned is above and beyond what was needed, and so these changes are mostly about the underlying calculations and the justifications.

About this specific analysis. In hindsight, we should have approached this differently. This specific analysis was an exploratory extension of a conceptual replication adopted in Study 3 from the target's Study 2 using a different sample and approach that we did not aim to replicate. There are also other issues with aiming to conduct a mediation analysis using a correlational design, and so all along we considered this analysis as exploratory and suggestive. In addition, this analysis has little if anything to do with the small telescope approach we previously applied broadly to all analyses by multiplying the end result.

We combined the two previous tables in the supplementary separately documenting the effects and power into a single table and more clearly labeled which of the effects are included in the power analysis, raising the target power to 95%. Please see revised combined power analysis Table S1.

Based on that, we revised our power and sensitivity analysis section:

> "We first computed target effect sizes for direct replication (summarized in Table 1). Effect size and confidence intervals were calculated with R (Version: 4.1.2; R Core Team, 2020) with the help of a guide by Jané et al. (2024), and power analyses were then conducted with a combination of R and GPower (Version 3.1.9.6; Faul et al., 2007) for the factors that the authors found support for in the target article (flagged as significant results). We conducted a series of a priori power analyses based on these effect sizes and we found that we require 289 participants to detect the effects reported in the target

article with 95% statistical power at alpha = .05 (see supplementary materials Table S1 and analysis code for more details).

Given the likelihood that the original effects are overestimated, we used the suggested Simonsohn (2015) small telescopes approach with the generalized rule of thumb of multiplying the largest required sample size among all target studies (208) by 2.5 to 723, rounding up to 800 participants. A sensitivity analysis indicated that a sample of 800 would allow the detection of d = 0.23 for independent t-test contrasts and r = .12 (both 95% power, alpha = .05, one-tail), typically considered weak to medium effects in social psychology research (Jané et al., 2024), and half or less than the effects reported in the target article."

.7. I found all other responses by the authors adequate and have no other issues about the registered report.

Thank you very much for the very detailed review and all the helpful constructive suggestions. We are very grateful.

Reply to Reviewer #2: Dr./Prof. Philipp Schoenegger

> **The authors have responded to all my comments, either directly changing their manuscript properly in response or explaining why they did not follow my recommendations. Why I do not persoanlly agree with all their reasoning in cases where they chose not to follow my recommendations, I can see their point of view, with the remaining disagreements not being scientifically important.**
>
> **I am thus happy to recommend the Stage 1 for acceptance and am looking forward to seeing the results!**

Thank you for your support and feedback throughout.