Reply to PCIRR 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/DDHWZqiEixgI

A track-changes manuscript is provided with the file: "PCIRR-RNR-Kahneman-Tversky-1973-replication-main-manuscript-track-changes.docx " (https://osf.io/h7qbp)

(Please note: Given the reviewers' request to change the presentation structure of the manuscript, the tracked changes marked will be considerable and might be a bit difficult to follow.)

Summary of changes

Section	**Actions taken in the current manuscript**
General	R2: We modified the replication closeness label to "conceptual" for Studies 1 and 2.
Introduction	R1: We elaborated further on the potential psychological processes involved and clarified the definition and issues with the definition of the representativeness heuristic. We added a note explaining the methodology of our literature search. R3: We elaborated further on the issues surrounding the definition and theories of the representativeness heuristic.
Methods	R1: We clarified the decision to not conduct exclusions and addressed the difficulties with sensitivity analyses for Studies 1 and 2. We removed the mention of MTurk and CloudResearch in the survey. R2: We modified the sensitivity analyses and modified the target alpha to .005 to reflect the concerns of multiple comparisons. R3: We clarified Prolific participants' characteristics. We modified the measures of statistical knowledge for Study 7. We clarified the deviations from the target article regarding Studies 1 and 2.

Section	**Actions taken in the current manuscript**
Results	R1: We added the simulated results of the Bayesian analyses.
Reporting	R1: We clarified the wording of the replication classification in the Abstract. R2: We reorganized the entire manuscript to report in order of studies. We clarified the wordings regarding definitions, experimental groups, page numbers, etc. R3: We clarified the wording regarding experimental design.
Supplementary materials	R1: Images that require permissions to be produced were removed. JAMOVI files for Bayesian analyzes were uploaded. R2: We removed unnecessary text in the Qualtrics. We implemented the option of using groundhog package in the code. R3: We modified the measurements for statistical knowledge for Study 7.

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. Rima-Maria Rahal

> **I have now received three reviews of your submission on a replication project addressing Kahneman and Tvsersky (1973). In line with my own reading of your manuscript, the reviewers highlight important strengths of your outlined approach, but also note some areas for further improvement. In line with these suggestions, I would like to invite you to revise the manuscript.**
>
> **Most salient are the need to clarify questions regarding the 7-in-1 approach of conducting multiple replication attempts in one study, regarding the sampling plan, as well as regarding the nature of the replication and evaluations of replication success. These issues fall within the normal scope of a Stage 1 evaluation and can be addressed in a careful and comprehensive round of revisions.**

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

Reply to Reviewer #1: Dr./Prof. Regis Kakinohana

> **Thank you for the opportunity to read this manuscript. I hope my comments are useful for authors to improve their work.**
>
> **The authors aimed to replicate and extend the studies of Kahneman and Tversky (1973), which demonstrate the representativeness heuristic. The role of heuristics in judgment and decision-making studies and the discussion about the importance of replications make this research very interesting. However, there are some points of the work that could be improved.**

Thank you for the positive opening note and the constructive feedback.

> **1) The first observation is in the Abstract. An important initial information for the reader is the replication classification. So instead of simply " we replicated Studies 1 to 7" it could be " we close replicated Studies 1 to 7".**

Response: We agree that the replication classification is important initial information. We received suggestions about the replication classification from other reviewers, and we decided to classify our replication as a close replication for Studies 3 to 7 and a conceptual replication for Studies 1 and 2.

Action: In the Abstract, we changed the phrasing to:

> "We conducted a conceptual replication of Studies 1 and 2 and a close replication of Studies 3 to 7 from Kahneman and Tversky (1973)".

Table 5 (previously 11) "Classification of the replication, based on LeBel et al. (2018)" was adjusted to reflect that Studies 1 and 2 were reclassified as a conceptual replication.

> **2) There is an important conceptual issue to be addressed right at the beginning of the introduction. The representativeness heuristic is different from the availability heuristic, but the text can confuse the reader. I strongly suggest that the authors evaluate an improvement in the description of the representativeness heuristic A so as not to mix it with the availability heuristic.**

Response: We understand the concern, and other reviewers have also raised suggestions of improving on the introductory definition of the representativeness heuristic.

Action: The opening two sentences of the introduction have been changed to:

> "Kahneman and Tversky (1973) introduced and reviewed the "representativeness heuristic" as a mental shortcut in which people tend to make predictions, evaluations, or classifications more based on representativeness - the degree of resemblance of essential features of the target (e.g., fit with stereotypes, belonging to categories) - than based on objective evidence and statistical information. This is related to yet different from the "availability heuristic", a mental shortcut in which people tend to make predictions, evaluations, or classifications more based on the ease-of-recall of related examples than based on objective evidence (Tversky & Kahneman, 1973). While both heuristics may be helpful in some circumstances, they may result in systematic biases with real-life implications."

> **3) The literature review focused on the methodological aspects and did not address the psychological processes involved. Unless I am mistaken, there is no mention, for example, of the dual model of cognitive processes. This also makes it difficult to understand the extent of the replications. For example, why do the authors expect statistical knowledge to be associated with the persistence of the representativeness heuristic? I suggest that the authors include more information about the theoretical bases that support their hypotheses. There is much literature on the psychological processes involved in cognitive bias. As a suggestion, I mention a few:**
>
> - **Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8(3), 223–241. https://doi.org/10.1177/1745691612460685**
> - **Kahneman, D. (2011). Thinking fast and slow. New York: Farrar, Straus, and Giroux.**
> - **Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. Thinking & Reasoning, 24(4),**

423–444. https://doi.org/10.1080/13546783.2018.1459314

- **Toplak, M. E., & Flora, D. B. (2021). Resistance to cognitive biases: longitudinal trajectories and associations with cognitive abilities and academic achievement across development. Journal of Behavioral Decision Making, 34(3), 344–358. https://doi.org/10.1002/bdm.2214**

Some studies have also explored the relationship between cognitive biases and statistical or numerical knowledge. As a suggestion, I mention a few:

- **Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: numeracy and metacognition. Judgment and Decision Making, 9(1), 15–34.**
- **Kakinohana, R. K., & Pilati, R. (2023) Differences in decisions affected by cognitive biases: Examining human values, need for cognition, and numeracy. Psicologia: Reflexão e Crítica, 36(1), 26. https://doi.org/10.1186/s41155-02300265-z**
- **Reyna, V. F., & Brainerd, C. J. (2023). Numeracy, gist, literal thinking and the value of nothing in decision making. Nature Reviews Psychology, 2(7), 421–439. https://doi.org/10.1038/s44159-023-00188-7**
- **Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. Thinking & Reasoning, 27(1), 38–68. https://doi.org/10.1080/13546783.2019.1708793**

Action: Indeed, our replication is focused on the methodology of repeating the same studies to assess how well we are able to replicate the results reported in the target article. The theoretical discussions and the empirical details are provided in the target article, with a prolific line of literature that followed which discussed those theories in great detail. The hypotheses we outlined are of course not our hypotheses, we either included the target's hypotheses or tried to deduce the hypotheses from their descriptions, methods, and/or findings.

We aimed to stay focused on the replication, and we feel that discussing the dual model of cognitive processes would distract from that. We appreciate the citations, yet these are very well-known and too broad for our fairly simple and straightforward replication task of specific classic demonstration of a single heuristic (e.g., Kahneman, 2011, currently has 48900+ citations and covers many other decision-making aspects).

We previously briefly summarized some of the theoretical discussions and the heated debate on the topic. We added additional paragraphs on the theoretical basis for the relationship between statistical knowledge and cognitive biases under “Study 7: Correlational analysis of regressive intuitions and statistical knowledge” subsection under the “Extensions” section.

We also added a planned discussion for Stage 2 to call for conducting a systematic review of the literature:

> [Planned for Stage 2: Following Dr./Prof. Regis Kakinohana comment regarding the need for mapping of the literature and replications, we will discuss the need for a follow-up systematic review of this prolific literature.]

4) **On page 10, it is written: “Much of the literature has focused on the paradigm in Study 3, yet - as far as we know - with no comprehensive replication of all seven studies described in Kahneman and Tversky (1973) together looking at the effects systematically”. Instead of just “as far as we know”, a more structured search in research databases, such as APA PsycNet, Web of Science and PubMed, would give more robustness to this statement.**

Response: We understand the need for a more systematic search, yet that is not within the scope of a replication project. As is common in a review for an introduction of a replication manuscript, we conducted a non-systematic search of the literature with Google Scholar using the “search within citing articles” function looking for replications or the different scenarios (with keyword patterns such as replicat*, and keywords such as “review”, “freshmen”, “personality sketch”, “lawyers”, “engineers”, “adjectives”, “evaluation”, “translation”, “consistency”, and “aptitude test”). We added that as a footnote. The statement of “as far as we know”, was aimed to be humble and address the possibility that there might be other investigations out there that a straightforward literature search would not find.

As noted in the previous paragraph, we added a planned discussion in Stage 2 regarding the need for a systematic review of the literature.

5) **On Method, the authors aimed for a larger total sample of 890 participants due to possible exclusions of 10% based on their previous experience with the target sample. However, on page 49, on Outliers and exclusions, there is no mention of any exclusion criteria. I suggest the authors describe the exclusion criteria they expect, based on their previous experience.**

Response: Thank you for catching that. Indeed, we do not plan to conduct any exclusions.

Action: We removed the references to exclusions in our power analysis and throughout. The planned sample size has been updated to 800.

> **6)The authors had difficulties with the power analyses and report sensitivity analysis on page 24. However, the authors did not report the expected effect size of all the studies (e.g., studies 1 and 2). Even if they could not calculate the effect sizes of the original studies, the sensitivity analysis of the replications is important information.**

Response: We agree that sensitivity analyses are important even if original effect sizes were not reported/ able to be calculated. However, the correlations for Studies 1 and 2 were on an item-level, which sensitivity analyses would not be helpful with given that the number of items are fixed. We tried our best to include sensitivity analyses for other studies where possible.

> **7) Also on page 24, the text mentions that the participant will be recruited on Prolific. However, unless I'm mistaken, the survey available on OSF also mentions Mturk and CloudResearch. Either the OSF material or the text should be adjusted.**

Response: Thank you. Participant recruitment will only be done on Prolific.

Action: We removed the mention of Mturk and CloudResearch in our survey.

> **8) The authors indicate that in case of not find support for the hypothesis for any of the studies, they would run a complementary Bayesian analysis. However, I did not find more details about this Bayesian analysis in the Method. The simulated results also did not present these analyses.**

Thank you for the feedback. This was a bit difficult to catch, but the ggstatsplot figures included the results for the Bayesian on the bottom right. We should make things more explicit and for all analyses.

We also added explicit placeholders for the Bayesian analysis for:

a) Study 1 and 2: “To quantify support for the null hypothesis that self-perceived accuracy is not correlated with the degree of conformity to base rates, we conducted a Bayesian analysis and found that the support for the null hypothesis was 1.259 (BF_{01}) times stronger than the alternative hypothesis.”
b) Study 3: “To quantify support for the null effect that the condition (high or low engineer) did not affect participants’ judgements of probability, a Bayesian analysis was conducted for the 2 (high versus low) x 6 (descriptions) two-way between ANOVA using JAMOVI. It was found that the support for the null hypothesis was 9.09 times larger than the

alternative hypothesis that there exists differences in judged probabilities between the high and low engineer conditions."

c) Study 4: "To quantify support for the null hypothesis of equal variances, we converted the *F* value from the Levene's tests of homogeneity into a Bayes factor (BF_{01}). At the adjectives level, the observed data favored the null hypothesis compared to the alternative hypothesis by a factor of 7.13 for the comparison of variances between the prediction and evaluation conditions. At the reports level, the observed data favored the null hypothesis compared to the alternative hypothesis by a factor of 11.17."
[Extension]: "To quantify support for the null hypothesis that confidence was not affected by whether an evaluation or a prediction was being made, we ran a Bayesian analysis for the 2 (adjective/report) x 2 (evaluation/prediction) two-way between ANOVA. It was found that the support for the null hypothesis was 14.7 times stronger than the alternative hypothesis."

d) Study 5: "To quantify support for the null effect that within participants means do not differ between the academic achievement and mental concentration condition, we conducted a Bayesian independent samples t-test. We found that the support for the null hypothesis was 9.74 times stronger than that for the alternative hypothesis (Cauchy prior = .707)."
"To quantify support for the null effect that within participants SD do not differ between the academic achievement and mental concentration condition, we conducted a Bayesian independent samples t-test. We found that the support for the null hypothesis was 12.91 times stronger than that for the alternative hypothesis (Cauchy prior = .707)."
"To quantify support for the null effect that within participants correlations do not differ between the academic achievement and mental concentration condition, we conducted a Bayesian independent samples t-test. We found that the support for the null hypothesis was 13.03 times stronger than that for the alternative hypothesis (Cauchy prior = .707)."

e) Study 7: To quantify support for the null hypothesis that statistical knowledge is not related to degree of regression, we conducted a Bayesian analysis and found that support for the null hypothesis (BF_{01}) was 0.138 times stronger when all data was included, and the factor was 0.138 times in favor of the null when overly regressive estimates were excluded (Cauchy prior = .707).

We added a note in the general "Data analysis strategy" section that: "Bayesian analyses for Study 3, 4 (extension), and 5 were performed with the "jsq" module on JAMOVI.", and that file is now included in our planned data analysis folder.

9) Unless I'm mistaken, some images require permission to be reproduced.

Thank you. All screenshots taken directly from the original Kahneman and Tversky (1973) article have been removed from the supplementary materials under the “Analysis of the target article” section. For the tables, we copied the numbers from the target article into our own tables.

This research is very interesting. I hope my comments are useful for authors to improve their work.

Thank you for the encouragement and positive constructive feedback.

Reply to Reviewer #2: Dr./Prof. Naseem Dillman-Hasso

> **I commend the authors for taking on this large registered report replication attempt. It is very obvious that much time and consideration has been put into this project, and while PCI reviews do not evaluate the importance of submissions, my personal opinion is that this is an important replication to undertake.**

Thank you for the positive opening note and the constructive feedback.

> **At it's current stage, I do not believe that this project is ready for data collection. I have a number of concerns, suggestions, and comments that I believe will greatly improve the end product. Even though this is my first time reviewing a registered report (although I have been involved in and am leading one right now), it is my understanding that the review process is meant to be constructive and collaborative. If it seems that any of my comments below are blunt, please do not take them as indicative of anything other than my quick jotting down of ideas.**
>
> **While my point-by-point comments can be found below, I wish to draw out a few themes and mention some concerns I have.**
>
> **.1. First, I believe that the manuscript can be organized in a much clearer way. I found that there was a great deal of repeated information due to the way that sections were outlined, and a lot of jumping back and forth was required by the reader. I left some suggestions below, but largely, I would try to work on cutting a great deal of text and consolidating repeat information. One major step would be going through all of the components of any one study in order, as opposed to components (i.e. Study 1 manipulations, measures, deviations, Study 2 manipulations, measures, deviations, etc.; as opposed to Manipulations study 1, study 2, study 3, etc., Measures study 1, study 2, study 3, etc.). While I did not note all occurrences of this, I do believe restructuring the entire manuscript would help with readability and reduce total length.**

We appreciate the feedback. We realized that our initial structure was a bit difficult for readers to navigate. We rearranged the manuscript as suggested, so that each study is presented in order of all its components.

> **.2. Second, I think that there should be more consideration put towards sample size, sensitivity analyses, and power analyses. There was a recent article in PSPR around considerations of power (Giner-Sorolla et al., 2024). Pulling from that article, and related research, I urge the authors to consider what power and sensitivity mean. Power is the probability of detecting an effect if there is an effect there, and is effect specific rather than study specific. The same holds through with sensitivity analyses (which is truly just a different mathematical representation of the same equation): sensitivity is related to an effect not a study. Multiple comparisons and analyses all have their own sensitivity analyses (or power levels). I would encourage a consideration of how multiple comparisons and analyses for any given study may affect the reliability of the sensitivity analyses reported, and what can be done about them.**

Response: We have devoted much consideration towards sample size with an analysis of what we could from that target article and details of sensitivity analyses given a large target sample of 800, well powered to detect far smaller effects that are typical of decision making and this literature.

It is unclear from your comment what component you thought was missing, yet we understand that you had concerns regarding multiple analyses and multiple comparisons, given the multiple studies and comparison groups that we included in our analyses.

Action: Given that we have 7 studies, with varying designs, we thought that in the spirit of the Bonferroni correction, we can generally divide the target alpha by 10 and instead set our target alpha for all analyses to .005.

We therefore conducted an additional set of sensitivity analyses using alpha = .005 for all the effects we could analyze, and the differences in the detectable effect size is fairly minor, and we concluded that we would still be well-powered. We added the following in the "Power and sensitivity analyses" section:

> We conducted a sensitivity analysis using Gpower (Faul et al., 2007), which indicated that a sample of 800 with a target alpha of .005, we could detect f = 0.16 for the 2 by 6 ANOVA in Study 3, f = 0.16 for the 2 by 2 ANOVA in Study 4, d = 0.39 for the independent samples t-test in Study 5, d = 0.16 for the paired samples t-test in Study 6, and r = 0.16 for the bivariate correlation in Study 7. These are commonly considered small to moderate effects in social psychology (Jané et al., 2024).

For comparison, this was the previous version's section:

> A sensitivity analysis using Gpower (Faul et al., 2007) indicated that a sample of 800 would allow the detection of *f* = 0.13 for the 2 (between) by 6 (between) ANOVA in Study 3, *d* = 0.31 for the independent samples t-test in Study 5, *d* = 0.13 for the paired samples t-test in Study 6, and *r* = 0.13 for the bivariate correlation in Study 7 (all 95% power, alpha = 5%, two-tail).

We updated the "sensitivity analyses" section in the supplementary with the screenshots from G*Power.

We also updated the alpha we previous set for order effects .005 to .001:

> 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.

We also added the following reminder to all Result sections:

> [Reminder for Stage 2: Alpha is set to .005]

> **.3. Third, I do not see this as a close replication, but rather a conceptual one. I discuss this more below but if the authors wish to argue that this is a close replication, I think there are some changes that have to be made. One of those is my fourth major concern.**

Response: We use the most common criteria for replication classifications we are aware of by LeBel et al. (2018) which helps align people's views to an objective standard. The objective standard can be debated by the community, yet for this replication it would be best to focus on its implementation.

As we described in the manuscript, we have conducted several of these types of replications before grouping many studies of a target review/multi-study paper, some with PCIRR which received endorsement for both Stage 1 and Stage 2, all of them concluded as mostly successful, and they have been classified as a close replication, not as conceptual replications (e.g., Au & Feldman, 2020; Hong & Feldman, 2024; Li & Feldman, 2024; Mayiwar et al., 2024).

That said, we do recognize that in some of the studies in this replication project, we made larger modifications and so this warranted rethinking our criteria for Studies 1 and 2.

Action: We changed to the following in the abstract:

> "we conducted a conceptual replication of Studies 1 and 2 and a close replication of Studies 3 to 7 from Kahneman and Tversky (1973)".

Table 5 (previously Table 11) “Classification of the replication, based on LeBel et al. (2018)” was adjusted to reflect that Studies 1 and 2 were separately classified as conceptual replications.

> **.4. Fourth, I do not love the design of this replication for two reasons. I am strongly opposed to all 6 (or seven, depending on how you want to count it) studies run by each participant. Much of the cognitive biases and heuristics literature does show how knowledge and practice reduce the effects that biases have on individuals. I would venture a guess that having participants run through a number of prediction tests may influence their answers on later questions, and even if randomizing order and looking at order effects is done, I think that this reduces the ability to detect an effect if an effect is there.**
>
> **Additionally, the population in question (Prolific participants) is not a naïve group. I just downloaded some of my own data from Prolific. This study included U.S. residents who were over 18 and proficient in English and were not involved in our pilot test. Out of 1226 participants, the mean number of prolific studies completed was 2194 (median 1719, SD = 1914). I think there is something to be said that this population has had more experience and exposure to heuristics and biases, and may not respond the same way. I would argue that a participant on prolific who has completed 300 studies is not representative of the population, and over 75% of participants in my sample have. While there is nothing that can be done about the non-naïvate of prolific participants, I would suggest that participants do not run through multiple studies (or only run through a subset of studies). If there is flexibility with the budget, I could even envision some participants running through only one study, some running through 2-4, and some running through all if there are interesting questions there. But I think that having participants running through all 6 studies is a bad idea.**

We understand and have acknowledged this concern, and have tried to pre-empt and address it in our initial submission.

First, there is no need to venture any guesses. As we explained in our plan, this is not the first time we are doing this kind of project. We mentioned four such projects that are very similar to the one we are doing here: Au and Feldman (2020); Hong and Feldman (2024); Li and Feldman (2024), and Mayiwar et al. (2024). Some of those are not only a similar setup, but are of a similar phenomenon and also by the Kahneman and Tversky duo. For example, Mayiwar et al. (2024) is a replication and extensions of the problems reviewed in Kahneman and Tversky (1972). Hong and Feldman (2024) is also a PCIRR Stage 1 endorsed Registered Report with mostly successful replications of problems reviewed in Tversky and Kahneman (1971). Li and Feldman (2022) is another PCIRR Stage 1 endorsed Registered Report with mostly successful replications of over 20 problems reviewed in Thaler (1999). All of these were conducted on online labor market (MTurk/Prolific/Connect), all of these with highly experienced participants, all of them combining many problems into a single setup, all of them replicating most of the problems (or finding future directions with reasons for why they did not replicate), and none of those with clear order effects. Furthermore, 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, with 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 PCIRR, and there are many more.

All this evidence so far can inform us of what we can expect here. There is no reason to assume, guess, or suspect, when we have so much evidence to the contrary. More importantly, all of our investigations have shown that this design is a strength, not a weakness, because rather than assuming there are order effects or that participants learn, we can test it. If there are order effects, then we would want to know, and then we can retest controlling for them, as we have outlined we would do. If there is no indication for order effects, that is also informative because we then know that we need not be so concerned about those.

Finally, we do not think it needed given all the evidence we have accumulated in our team about this design, but there is also broader evidence not from our team that we far over-estimate participants' ability to guess, learn, or improve, even in within-repeating designs. Our participants are professional survey takers who take their work seriously and do this as a living, and most are focused on the one task of completing a high-quality submission in time. Guessing what we intended, what condition they are in, what is the link between each scenario, inferring what is the right answer, or learning anything from that is a very difficult if not impossible task (you can try it out with your colleagues or students, it is an informative fun class/seminar exercise). Sample citations:

- Lambdin, C., & Shaffer, V. A. (2009). Are within-subjects designs transparent?. *Judgment and Decision Making*, 4(7), 554-566.
- Aczel, B., Szollosi, A., & Bago, B. (2018). The effect of transparency on framing effects in within‐subject designs. *Journal of Behavioral Decision Making*, 31(1), 25-39.

> **Thank you for the opportunity to review this project. I am looking forward to seeing the next version, and to seeing this project through. If there is anything I can clarify, let me know. Additionally, please understand that everything in this review is my opinion, so take things with a grain of salt.**

Thank you for the detailed and thoughtful feedback.

> **.5. Minor suggestions/comments:**
>
> **· Page 7: First sentence feels like it's trying to define both representativeness heuristic (hereafter RH) and give results from 1973 study all at once. I might suggest splitting it into two sentences, one that gives a clear one-sentence definition of RH followed by the 1973 seven studies example.**

Action: The opening sentences have been changed to:

> Kahneman and Tversky (1973) introduced and reviewed the "representativeness heuristic" as a mental shortcut in which people tend to make predictions, evaluations, or classifications more based on representativeness - the degree of resemblance of essential features of the target (e.g., fit with stereotypes, belonging to categories) - than based on objective evidence and statistical information. This is related to yet different from the "availability heuristic", a mental shortcut in which people tend to make predictions, evaluations, or classifications more based on the ease-of-recall of related examples than based on objective evidence (Tversky & Kahneman, 1973). While both heuristics may be helpful in some circumstances, they may result in systematic biases with real-life implications.
>
> In their review, Kahneman and Tversky (1973) presented the results of seven studies which showed that when making predictions, people shift evaluations towards their predictions of representativeness based on available information of the assessed target. This suggests that people seem to disregard prior probabilities according to Bayes' theorem and the accuracy or relevance of the evidence. Kahneman and Tversky (1973) concluded that the representativeness heuristic affects both categorical and numerical predictions, is influenced by the consistency of input variables for prediction, and is difficult to avoid despite having relevant statistical knowledge or knowing about the effect. Their seminal work has an immense impact on psychology, economics, policy, and beyond, and is considered the foundation of and related to the Nobel Memorial Prize in Economic Sciences recognition Kahneman received for his work in 2002.

The sentence regarding the availability heuristic was added due to a concern raised by another reviewer that the old description has the potential to be confused with the availability heuristic for readers.

- **Page 8: groups in S1 are unclear, I might name them (as K&H did) to clearly indicate that the key results were between-subjects correlations between similarity and likelihood (and likelihood vs base rates). Could be read as within subjects currently.**

Action: Thank you. The wording was changed to:

> "Participants in the similarity group ranked the same nine fields in terms of how similar Tom W. is to a typical graduate student in that field, and those in the likelihood group ranked the nine fields in terms of the likelihood that Tom W. is now a graduate student in each of the fields."

- **End of page 8/beginning of page 9: page number for direct quote would be helpful**

Page number was added and the quote was changed to:

> "'high intelligence, although lacking in true creativity. … a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place.' (p. 238)".

- **Page 9: As opposed to wording "representativeness heuristic predicts…", give the actual findings, and say "…, [in]consistent with what the representativeness heuristic would predict."**

We agree, thank you. The wording was changed to:

> "This seems consistent with the idea of the representativeness heuristic."

- **Page 11: Sentence starting with "At the time of writing…" is a run on, I would split it up for readability**

Changed to the following:

> At the time of writing (January, 2024), there were 9803 Google Scholar citations of the article and a few important follow-up theoretical and empirical articles. For example, Gigerenzer et al. (1988) engaged in a heated debate regarding the results of Kahneman and Tversky (1973) about the very nature and interpretation of heuristics and biases, and Koehler (1996) commented on the theoretical and practical problems of the effect size and applicability of the representativeness heuristic.

- **Table 2 Exp. 6: Indicate from where N was inferred (presumably df of t-test + 1?)**

We added this to the table note:

> ** inferred from df of t-test +1; not reported in target article

- **Page 21: I would choose another phrase than “Bayesian line” to describe S3. There is no one Bayesian line, maybe pull from original article wording which states “The curved line displays the correct relation according to Bayes’ rule.”**

Thank you. Description of Study 3 in the “Overview of replication and extension” subsection, Table 1, and in the results section were changed to:

> curved line of the correct relation according to Bayes’ rule

- **Page 21: for S4, should read “either adjectives or reports”**

We changed to:

> In Study 4, participants were given either adjectives or reports (between-subjects)

- **Page 21: S5 is unclear if it is between or within subjects. I would edit to read: “representing academic achievement, mental concentration, or sense of humor”**

We changed to:

> In Study 5, participants were given input scores supposedly either representing academic achievement, mental concentration, or sense of humor of ten students (between-subjects).

- **Table 3: Geographic origin of KH 1973 participants were recruited via student paper at University of Oregon unless stated otherwise (see footnote on page 238 of original study**

Response: Although we were aware of this when considering the geographic origin of the participants, we thought that this information could only indicate that participants were highly likely to be US American. However, we agree that it would be helpful to include this information as well.

Action: The following footnote was added under Table 4 (previously 3):

> “Original authors reported that participants were recruited by means of a student paper from the University of Oregon (see footnote 3 in p. 238), but did not further specify sample characteristics.”

- **I would remove all text mentioning “HIT” in the Qualtrics survey, as HIT is specific to MTurk and not Prolific.**

- **Related to the point above, the Qualtrics mentions “MTurk/Prolific/Cloudresearch” at the end. Are data being collected from multiple sources? If not, I would remove any references to other platforms.**

Thank you. We removed all text mentioning “HIT” or MTurk related terms.

- **I would encourage consideration of any exclusion criteria (i.e. US residents 18+, proficient in English, etc.) and to include those in text, including the attention checks**

We appreciate the detailed considerations of exclusion criteria. All Prolific participants are 18+ and we are filtering for current residents born and raised in the US, so all participants should be proficient in English. We added the following clarification to our method section:

> We targeted US Americans using Prolific’s filters. We restricted the location to the US using “standard sample”, we set it to “Nationality: United States”, “Country of birth: United States”, “Minimum Approval Rate: 90, Maximum Approval Rate: 100”, “Minimum Submissions: 50, Maximum Submissions: 100000”.

- **Page 49: How was the prior of 0.707 generated?**

Response: This is the default Cauchy prior. The Cauchy prior is weakly informative and its distribution is heavy-tailed, which reflects not having strong assumptions of the effect in question whilst allowing for the possibility of capturing small and large effect sizes. Subjective priors can skew the results, and so as replicators we would rather assume a passive stance that does not assume an effect or lack of to any extent.

Major suggestions:

.6. · Intro: I think that all of the information is there, but I don't love how it's organized. While reading through the first time, I found myself jumping back and forth to remind myself of what was said previously. I would suggest a reorganization along the following lines: Start by overviewing the findings from all studies in KH 1973. Something similar to table 1, but outline them all at the start. Then, move into different replication attempts (primarily around S1 and S3) and inconsistencies there and the theories for why replications are less/more successful (salience of randomness, etc.). Move into importance of replicating this specific paper, and then include the overview of replication/extension. End with a section outlining deviations/extensions in replications.

We moved the summary of the seven studies in the target article earlier and included it as part of the theoretical introduction on the representativeness heuristic in the "Representativeness heuristic" subsection in the introduction. Following this section is a brief discussion about replication attempts, then the importance of the replication, and finally the extensions, deviations, and an overview of the replication.

.7. · Page 10: unclear if actually testing reproducibility, given that reproducibility is the reliability of a prior finding using the same data and same analysis (Nosek et al., 2022)

Reproducibility has many sides. You are referring to outcome reproducibility, but reproducibility also has process reproducibility that applies to the reproduction of procedure, materials, and stimuli. From Nosek et al. (2022) that you cited:

> A process reproducibility failure occurs when the original analysis cannot be repeated because of the unavailability of data, code, information needed to recreate the code, or necessary software or tools. An outcome reproducibility failure occurs when the reanalysis obtains a different result than the one reported originally. This can occur because of an error in either the original or the reproduction study.

Conducting replications tests reproducibility of the target article, and in this case which had very brief reporting and many missing details, allows boosting the reproducibility of the process, materials, stimuli, analysis, data, and code, from very poor to hopefully very high.

.8.· Why was feedback accuracy not manipulated in S1/2 as it was in KH 1973? I understand the importance of including self-perceived accuracy, but I think there's something to be said about being told that you either were or were not accurate, and how that might influence usage of representativeness on likelihood. While I get that the original study found null effects of this manipulation, I feel that it is not true to the replication to remove it due to the deception being found "unnecessary and unconvincing." I would suggest keeping it in to remain truer to the original study (which would also allow for assurance that there is not an unequal distribution in perceived accuracy). I might have perceived accuracy first, followed by (deceptive) feedback.

Response: We understand, yet we please ask to proceed with this modification and without this manipulation.

When we do replications we need to evaluate the costs and benefits and make decisions. We did not feel that this type of manipulation was core to the idea of the article, we did not anticipate it having any impact on the results given the target's null findings, and we generally thought it was unconvincing in its implementation and would much prefer not to deceive our participants unnecessarily.

We do agree that a change like this should be better reflected in the our evaluation of the degree of deviation from the target's article, we have decided to change the replication evaluation label of Studies 1 and 2 from "direct-close" to "conceptual-far"

Action: Throughout we changed the classification of Studies 1 and 2 to conceptual.

.9. · Similar to the introduction, I would consider a restructuring of the methods section to go study by study as opposed to a section for manipulations, one for measures, and one for extensions. The cognitive load of switching back and forth between studies might be lessened if it's organized by study, with sub headers for manipulations/measures. This would also cut down significantly on the text.

Thank you, we understand this concern, appreciate the feedback, and made the requested changes.

> **.10. · Page 43: I am unconvinced by the determination of successful/mixed/failed replication. I suggest a consideration about if there are certain studies that hold more weight or less weight, and what that might mean for representativeness heuristic. I would argue that this determination can be made almost only post-hoc, due to the vast differences in potential outcomes (7 studies, multiple analyses, etc.). Perhaps a better metric could be whether a specific study (or analysis) replicated, and whether the evidence in aggregate seems to indicate that RH as a whole replicates. I don't have an answer on how to generate metrics for the second option (whether RH as a whole replicates) but the cutoffs given are unconvincing to me. What would happen if every study gave non-significant findings in the same direction as hypothesized? Someone might argue that power was too low but on aggregate the effect exists… Or what would happen if there are effects of the order or something of that sort?**

The main point of Registered Reports is to try and minimize bias and define things in advance so that we can track what was predicted and confirmed and what was decided post-hoc or analyzed as exploration.

Weights are especially subjective, and this is not a debate that we would like to engage in as replicators. Please consider what would happen if in the Open Science Collaboration (2015) or Many Labs, they would try and agree which of the studies hold what merit or how representative each is of the social psychology literature. This is not a debate that would be helpful or can be easily resolved. Therefore, what the OSC, Many Labs, and we try to do is to provide a replicability overview of a subset of selected studies. Here, the overall replicability rate for all studies we aimed to replicate may say something about the studies that this review chose to cover. In OSC (2015) and Many Labs 2 etc., they provided aggregates of very different studies, and we believe that many found the summary of a percent of successful replications meaningful. They could have just listed the 100 studies and then discussed which of those is important and what replicated or did not, yet the way most refer to those articles and make sense of what was done there is by an overall assessments of an aggregate of different studies in something that ties them together (domain, journal, etc.). What we tried to do here is to satisfy that need.

In OSC (2015) and Many Labs, they shared all their materials and data, and then other meta-scientists revisited their data with new analyses. We aim to do the same. All are welcome to take our findings and analyze this further in whatever way they would see fit, and employ their own value criteria. Yet, for our replication, we feel that this goes beyond the scope of our investigation.

As for the cutoffs, we are well-powered to detect weak effects and employ the best evaluations we know of for replications to assess replicability (LeBel et al., 2019). The importance of a Registered Report is that we - editor, reviewers, and authors, have agreed on an unbiased evaluation criteria before hand, and that based on that criteria they proceeded to data collection and evaluated the findings based on that.

> **.11. · Page 43: I would not argue that this is a close replication by LeBel et al.'s criteria. The population is different, the context is different, the setting is different, the procedure is different, and there are differences in operationalization and stimuli. To me, this is a conceptual replication. There is nothing wrong with that, but I would represent it as it is. LeBel et al. state that a close replication is when the IV or DV stimuli are different, but a conceptual (far) replication is when the IV or DV operationalization or population is different.**

Our previous similar replications mentioned above were classified as close replications with similar changes, and for a good reason, these changes are either not classified as critical to the distinction between close and far, are addressed to be as close as possible in our replication, or seem to be of no theoretical or practical importance. One might summarize it this way - for most of these studies this is as close as we can ever get with a replication, given that any replication is in a different time, a different context, using a different sample, and with some methodological adjustments, especially so with a replication of findings over 50 years old with little to no details about most of what they did.

Whenever possible, we adhered to the target as closely as possible (with the exception of not implementing deception). LeBel et al. (2018) were quite flexible in their definitions, as - for example - “Population refers to major population characteristics, such as age and whether the sample is drawn from the community or a special clinical population.” (notes for Figure 1 in the referenced paper). The target article was very vague about their participants, and in the one that mentions University of Oregon we mirror that with US American adults, of a similar population. For the representativeness phenomenon, it would be disappointing if any of these factors were to matter, and from our own experience - we had successful replications of Kahneman and Tversky’s work from the same time of similar phenomena that showed that these factors do not matter, and there is not reason to expect representativeness to be any different.

.12. · I found the discussion of power/sensitivity/sample size to be removed a bit from the analyses: sensitivity analyses do not take into account the multiple effects generated. I would suggest a careful consideration of what effect the sensitivity analysis is generated in relation to, and what to do to adjust for multiple comparisons.

Thank you. We now addressed this with an adjustment of our sensitivity analyses to an alpha of .005. Please see our reply on this issue above.

Stats Comments

.13. · I would highly recommend using the groundhog package (https://groundhogr.com/) to ensure reproducibility of all code. This would allow for version control of packages.

Though we appreciate the idea behind groundhog, it has some drawbacks. It essentially forces your environment to install the versions of the date set for the code you are running, and our experience is that it has not always worked smoothly across changes in R and RStudio versions. We ensure reproducibility by making clear and citing all packages and their versions, and by providing a knitted version of the Rmarkdown that shows the code next to its output.

To address this, we added a commented line that would allow anyone to use the groundhog versions of our packages, should they want to:

```
if (!requireNamespace('groundhog', quietly = TRUE)) {
        install.packages('groundhog')
}
library("groundhog")
# groundhog.library(requiredPackages, "2024-03-14", tolerate.R.version='4.2.1')
```

.14. · I would suggest to not use “99” as a code for missing age if that is a valid age in the dataset. Consider an obviously implausible value such as “999”, or commonly used metrics scuh as “NA” or “.”

Is 99 a reasonable age for Prolific participants? We believe that a missing value for age of “99” is already implausible given the target sample and the tools used. We validate the age response, and so increasing the allowed range beyond the 18-99 might result in some innocent mistakes that we would not know how to interpret (at which value would you start considering the age “implausible”?).

.15. · I would recommend installing tidyverse instead of individually ggplot2, haven, kintr, dplyr, purr, etc.

No need to install things that are unneeded. It also further complicates your other point above about reproducibility. In any case, we figured it would be helpful to list out the individual packages so that those running it would have a better idea of the functionality used.

.16. · If the data is collected via Qualtrics, why is it imported via a .sav file as opposed to an open format such as .csv?

.SAV is a commonly used file format for datasets. It is used by the open-source PSPP (SPSS clone; explained here), by R (using haven), JAMOVI/JASP, and many other stats software/packages. This file format addresses the many issues and weaknesses in .CSV and .XLSX files and allows for better integration of labels and values. Of the many issues in .CSV we will just mention the issues with text fields that include commas, tabs, and line-breaks that may cause the entire load to break, and the different behavior caused by different packages that load .CSV files with such formats. XLSX files do annoying things like convert numbers to dates. .SAV was built for and meant for datasets.

Reply to Reviewer #3: Dr./Prof. Peter Anthony White

> **Introductory. The representativeness heuristic proposed by Kahneman and Tversky has been and continues to be hugely influential. Not only has the 1973 paper inspired much further development of the research topic and been cited many times, it is a standard component of undergraduate modules on thinking and reasoning. It forms part of a general thesis developed by the authors that much thinking and reasoning is guided by heuristics rather than by, for example, strict rules of inference or knowledge of statistics and probabilities. The present authors propose a slightly altered replication of the set of studies reported in the 1973 paper. The registered report manuscript is nothing if not thorough, with detailed specification of the methods and analyses that would be used in the research. The thoughtful preparation for the research is admirable. However I do have some concerns which I will do my best to express.**

Thank you for the positive opening note and the constructive feedback.

> **titu.1. The opening paragraph does get across the prediction that prior probabilities will (often) be disregarded, but it does not state the qualifications to this, nor does it define representativeness. In fact there is no unambiguous definition of the representativeness heuristic and just characterising it as a heuristic draws a veil over the kind of processing that is actually going on when people make judgments of the sort exemplified in the research studies.**
>
> **Kahneman and Tversky (1973) (Hereafter "K&T") say that people "select or order outcomes by the degree to which the outcomes represent the essential features of the evidence" (pp. 237 - 238). That is the closest thing I can find to a definition in the 1973 paper. But the phrase "select or order" is odd, given that the research is concerned with judgments of likelihood. And what are "essential features" of evidence? Why are prior probabilities and other statistical information not part of that? People do judge by prior probabilities when they have no other information, or when the information they have seems not to be informative or relevant. Judging from the studies, it is really a contest between statistical information and individuating personal information and the latter usually wins. So perhaps they should have said, "judgments of likelihood about individuals are determined by relevant individuating information when available, not by prior probabilities". That makes it look less universal and less like a heuristic, more like a statement of people's ignorance about probabilities**

> **and how to use them in judgment. People more expert than I have written more extensively on these issues - Gigerenzer, whom the authors cite, is an example - and I think their work is very relevant to this manuscript and merits closer attention. As the authors are proposing a replication study they don't have to do a thorough critique of the representativeness heuristic and they don't have to agree with my analysis of it, but I do think they should address the problematic issue of what the representativeness heuristic is, how far its use generalises beyond the topics of the studies in the 1973 paper, and whether an unambiguous definition of it can be formulated. That much is important to understanding what is going on in the studies. Readers should be given a clear idea of what is really being tested, if possible. I should think a paragraph or two should suffice.**

Response: We agree that there are many issues regarding the definition and underlying theory for the representativeness heuristic. There exists a great controversy in the field with many scholars critiquing the generalisability and even the existence of a representativeness heuristic. Discussing all that goes far beyond the scope of a replication.

Action: We added a paragraph in the “Replication attempts” section under “Representativeness heuristic” to discuss the literature regarding the definition and theories related to the representativeness heuristic. We also added this to our planned discussion for Stage 2:

> [Planned for Stage 2: Following Dr./Prof. Peter Anthony White comment regarding the lacking definition and clear scope of representativeness, we will discuss the need to discuss and better define representativeness with clear measures and falsifiable criteria.]

> **.2. The remainder of the introduction does a good job of reviewing relevant replications and critiques of the representativeness heuristic research. However my main concern is that I'm not sure what is the point of doing a replication of the studies, given the amount of water that has flowed under the bridge since they were carried out. The research literature has moved on, as the brief summary of relevant subsequent research makes clear, so what could we learn from replication of the original studies that would make a real contribution to the literature? If the proposed studies do indeed replicate the results reported by K&T, that just confirms that they suffer from the problems identified in subsequent research. If the proposed studies don't replicate the original findings, what would that mean? The authors should give some thought to that. In general, they need to make a case that there really is a need for the proposed replication.**

Yes, we understand your view. We are unsure just how much water flowed and under what bridge, the findings in this article are still used as seminal examples for representativeness in books, courses, and talks, and the debate that ensued afterwards is far from resolution, and - if anything - is far from having been concluded as problematic.

The main issue from a replication point of view is that at the moment we do not know whether their findings or the findings of those who raise issues about the findings are replicable. Both sides need to establish replicability. We feel that it would be best to start from as early on as possible and build up to see what replicated across all the seminal findings in this research domain.

Consider, for example, our Chandrashekar et al. (2021) replication (https://doi.org/10.15626/MP.2020.2474) of the adversarial collaboration between the two camps regarding the Linda Problem and James Problem where Mellers helped Kahneman and Hertwig test different versions of the conjunction bias. Which of those camps is right? Their conclusion was that both, or neither, but regardless of what that means consider that we were unable to fully replicate their joint findings for the James Problem. This suggests that even seminal effects might be context sensitive, and that different findings and interpretations could be due to the sensitivity in the replicability of the findings (under-powered samples, noisy measures, specific scenarios that are more sensitive than others).

We cannot simply dismiss replications just because there is already so much that has been done without replications. Instead, we should call for more replications and of everything that has been done in this domain. Without replications, the whole debate is subjective and theoretical, without any solid empirical grounds.

> **.3. The various justifications for the replication given on p. 11 struck me as rather vague. On "the potential for improvements in methodology", that seems to me to be contradicted by the need for a replication study to use the same methods as were used in the original research. If there are going to be "further extensions examining the effect of consistency in numerical predictions", what do the authors hypothesise about consistency and why, and how does that issue fit into the existing literature? Extensions should be theoretically motivated and should test hypotheses. They mention "the absence of direct replications" but, unless replicability is likely to be a serious issue here, does that really matter, given the multiple studies that have added to or critiqued the research literature on representativeness? So I think there needs to be a stronger justification for replication, given the extent to which the field has moved on in the last 50 years.**

In the words of a seminal paper on replicability by Makel et al. (2012; https://doi.org/10.1177/1745691612460688) showing a replication/novel ratio of 1.07% (only 1.07 of every 100 published articles is a replication):

> "As an arbitrary selection, if a publication is cited 100 times, we think it would be strange if no attempt at replication has been conducted and published"

It seems to undermine the scientific validity of our field that we have not systematically revisited most of this seminal Noble-prize associated work to ensure that we understand it, that we can reproduce it, that we can replicate it, and that we have an updated recent estimate of its effects and their consistency. Replicability is always an issue for science, it cannot not be an issue, especially so for such impactful findings. Therefore, we consider the need for replications to become mainstream as a given.

Going beyond that, in a number of cases we indicated that when we started analyzing the original demonstrations we realized many issues. From unclear procedures, through lost stimuli for most scenarios, to methodological issues (e.g., Studies 1 and 2 with a comparison of different groups that represent different conditions without random assignment). The replication aims to revisit those studies and address some of those methodological weaknesses to allow the community to better understand what exactly this target article did and to what extent this replicates.

As for the extensions, the consistency was meant as an exploratory direction, for us or the community, as a follow-up of value. Given that Stage 1 is said to be mostly focused on confirmatory analyses, we did not include that analysis. For an example of such analysis, please see our replication of the nine problems in Kahneman and Tversky (1972), in https://osf.io/28ypu.

.4. On p. 13 I thought the section "Selection of studies..." was unnecessary. The preceding section could just conclude with a sentence saying that all seven studies in the paper would be replicated. Table 1 is entirely adequate as a description of the studies, and the numbering is very useful given that the studies in K&T were not numbered. Table 2 is also a clear and useful summary of the results.

We revised the entire section as you suggested above.

.5. p. 14 The rationale for the extension to study 4. I understand why the authors would want to investigate confidence more explicitly. However, saying "which we theorized would give a more straightforward estimate of participants' confidence" is a bit vague. Confidence judgments have been used in large numbers of studies on various topics, but they are explicit judgments, which are not always trustworthy. They could, for example, be prone to response biases such as self-presentation. I wonder if the authors should check whether explicit confidence ratings are generally regarded as valid. Also, if it turns out that the confidence ratings don't predict the standard deviations in judgments, how would that result be interpreted?

“Trustworthy” to what end? What other measure of someone’s confidence would you suggest?

If we go back to the classics in decision making, Fischhoff, Slovic, and Lichtenstein (1977) were among the first to demonstrate overconfidence. How? By asking participants to self-report their confidence and then measuring those against their accuracy. 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). This is as trustworthy as we can ever hope to get, from a replicability stand-point. 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”.

This is all to say that asking participants to self-report their own confidence is a common method in judgment and decision-making, that is considered reliable and for the effects that we’ve tried - replicable, even 45 years later.

.6. On p. 21, "In Study 4, participants were given either adjectives and reports...". Should the "and" be "or"? If not, there is an "or" missing.

Action: Thank you, it should be an “or”. We adjusted accordingly.

> **.7. In the next paragraph, again "either" is used but no "or" appears later in the sentence, so something is not right with that.**

Action: We assume this is referring to the phrase “representing academic achievement, mental concentration, and sense of humor of ten students”. This has been changed to “representing academic achievement, mental concentration, or sense of humor of ten students (between-subjects)”

> **.8. On p. 31 it is stated that the studies will be run as an online Qualtrics survey. My experience of supervising final year project students over the COVID period, when Qualtrics was a common option, has not impressed me: many participants do not engage with the tasks and the data have been very noisy. I am concerned about this and I think the authors should discuss data quality. For example, what would count as evidence that a participant had not engaged with the task and what rules would there be about excluding participants that don't engage properly with it? What would be evidence of lack of engagement?**

Prolific participants take pride in their work, and in our experience most do careful serious work, perhaps even more so than the participant pool undergraduates who are coerced into participating in order for them to receive course credit to allow them to graduate.

To help communicate a serious survey and the need for attentiveness, we employed many measures to try and ensure attentiveness. One of those is that in the study outline before embarking on the study participants must indicate their consent to several questions, and the choices are randomized as to ensure both consent and attentiveness to options, as you have seen when you tried out our survey. We added the following clarification to the manuscript:

> Three of the four questions also served as attention checks, with the order of the options being rotated (yes, no, not sure).

One of those is the need to confirm being attentive, and copy-paste of a declaration that they understand the need for careful reading and attentiveness. In our experience, this proved to be a very effective attention check which would not require exclusions, given that inattentiveness leads to indicating no consent, and therefore not even starting the study.

Finally, the issue of data quality of those platforms has been discussed extensively, MTurk/Prolific and Qualtrics are some of most widely used tools in psychological science. Yet, we need not rely on others’ data, because in our manuscript we cited many examples of our team’s concluded high-quality replications that were conducted using MTurk and Prolific participants, many of those successful, many of those with comparable (or higher quality) results to those with in-person samples. We have completed over 120 of these replications with overall very positive experience with these platforms.

> **.9. In the participants section the authors mention recruiting people on Prolific. I had to google that to find out what it was and I think the authors should add some information about it for the benefit of readers in the same state of ignorance as me. What demographic information can they provide about samples obtained using Prolific? And in particular for study 7, do they not need students or ex-students who have done modules or courses on statistics? This needs to be sorted out.**

Good feedback, thank you. The use of Prolific is so common in psychology, that we just assumed everyone knows what Prolific is, but you are very right, this does need clarification. We added a citation that was one of the first articles to introduce it: Palan and Schitter (2018), https://doi.org/10.1016/j.jbef.2017.12.004

We also added information about the filters used in Prolific when recruiting:

> Participants were 18 years old and above and were born, raised, and residing in the US. We targeted US Americans using Prolific's filters. We restricted the location to the US using "standard sample", we set it to "Nationality: United States", "Country of birth: United States", "Minimum Approval Rate: 90, Maximum Approval Rate: 100", "Minimum Submissions: 50, Maximum Submissions: 100000".

We will be using the general population for all studies, and will not be aiming at recruiting students. For Study 7, we noted this as a deviation.

> **.10. p. 26: "We ran the seven studies together in a single unified collection". It appears that this is the plan for the real data collection. Even though the authors say they have done this before, I do not think it is a good idea for all participants to take part in all the experiments. First, it is not the way K&T did it, so it compromises the fidelity of the replication. More important, the danger is that participants' knowledge of the experiments will accumulate as they go through, and that could have effects on their responses in the later experiments. They might, for example, be induced to reflect on what they are doing by the repeated presentations of personality information, and might change their thinking about its relevance. It would be much better to have separate samples for each experiment - but then I am uncertain whether there could be a target n of 800 for each one, because that would entail a total sample size of 5600. This needs to be clarified. The issue is discussed on p. 49. My response to that is**

> **that the authors should examine order as a moderator regardless of the results they get; I do not think this analysis should be contingent on the data.**

Please see our detailed reply #3 to reviewer #2.

The issue with running all these analyses regardless of the results is that this involves at least doubling the number of analyses, impacting readability and interpretability, when it is not clear what the added benefits are. If the studies replicate, as they have in our other similar projects, then these analyses are of limited value.

Consider also, that with 7 studies in random order, there are many ways to analyze such a moderator: is it the positioning in the 7? Is it which of the specific studies came before it?. The benefits are when the moderator might be what is causing the effect to go from signal to no-signal. If the effects replicate well, then there is no need for that. If things fail, then this would allow us to check for possible issues.

> **.11. p. 30 The measure of statistical knowledge is rather vague and subjective. Would it be better to ask the Ps what education they have had in statistics - e.g. what modules at university and at what level - and how well they did? I see on p. 34 and p. 37 that they will be asked directly about their knowledge of confidence intervals. How will they answer that question? Will it be a free verbal report? What will the authors do about Ps who report that they don't know what a confidence interval is?**

Referencing one of our previous projects that has received in principle acceptance from PCIRR (Hong & Feldman, 2024, https://osf.io/mns7j/), we replaced our original self-report question with the following:

> 1) Statistics knowledge: “How would you rate your proficiency in the use of statistics?” (0 = Not at all proficient in statistics; 100 = Very proficient in statistics)
>
> 2) Statistics usage: “How often do you use statistics and statistical inferences in your job?” (0 = Not at all; 100 = All the time)
>
> 3) Statistics training: Do you have any training in statistics? (choose the one most fitting)
>
> - No statistics training (0)
> - Highschool level statistics training (1)
> - College level statistics training (2)
> - Professional training in statistics (3)
> - Academic training in statistics (postgraduate and above) (4)

We will report the correlations with the three measures (statistics knowledge, usage, and training), and the degree of regression in the participants' answers for the confidence interval.

> **.12. Table 4 in the manuscript appears to have combined studies 1 and 2 from K&T, but it doesn't resemble the study 2 reported in K&T. In study 2 in K&T "the experimental materials consisted of five thumbnnail personality sketches of ninth-grade boys" (p. 240). Participants were divided into high and low accuracy conditions, on the basis of a statement saying how often people like themselves make correct predictions. Are the authors planning to do that? Possibly more information is needed there.**

Response: Thank you for the feedback. We reported the deviations from the original Study 2 in Table 3 (previously 10) "*Replication and extension adjustments to the target article's methods and design*", yet realized we could have been clear in describing those changes.

Action: In Table 3 (previously 10) "*Replication and extension adjustments to the target article's methods and design*", the description of the target article in the row about Study 2 materials has been changed to "Participants were told that the descriptions shown were thumbnail personality sketches of ninth-grade boys allegedly written by counselors (authors did not clarify whether these were artificially created)." For the "adjustment in current study" column, the sentence "Participants were not told about how the descriptions were obtained." was added for clarification.

We detailed in the original submission that we are not implementing accuracy manipulations. Please also see our replies above #8 to reviewer #2 regarding that choice.

> **.13. On p. 49 the authors state that they set alpha to .005. How did they arrive at this decision? Did they use the Bonferroni correction? I think some sort of rationale should be given because there is a happy medium to be found between the risks of type 1 and type 2 errors and an arbitary choice might not be in the right place for that.**

First, we would like to point out that the alpha threshold is only one of the factors that we use to assess replication. We feel that generally the psychological science literature has shifted somewhat to emphasizing effect sizes and confidence intervals, and our use of LeBel et al. (2019), whenever possible, compares the confidence intervals of the replication effects to the effects reported in or deduced from the target. We also supplement these with Bayesian analyses when NHST fails to detect a signal to reject the null. Here, given that the target article was so brief and lacking in reporting, sometimes with no statistical tests at all, and with no mention of alphas or effect sizes, we both have more flexibility to determine our statistical tests (and the alpha), but are also less able to make meaningful comparisons.

We are aware of the ongoing debate regarding setting alpha (e.g., Benjamin et al., 2018; Lakens et al., 2018), with many of the researchers engaged in those debates not following their own recommendations. We understand why - it is controversial, it is subjective, and it is complex and confusing. In some ways, any alpha is arbitrary, it would seem that Fischer did not originally plan for his side-note on .05 to become the de-facto cutoff point for NHST.

We generally aim to follow the target article's setup on its own terms with the common alpha of .05, yet some reviewers/editors rightly point out that in some of our designs this might lead to biased conclusions. In our replication projects we keep getting conflicting views from editors and reviewers on how they would like us to approach the topic.

What seems to have served us well when such issues are brought up is to apply some kind of a correction to alpha. The simplest idea is one similar to Bonferroni's, of dividing the alpha by an approximation of the number of analyses. Given that we have seven studies in a unified design with some studies with multiple analyses, we thought that an overall division by 10 would be a simple solution, and so aimed for .005.

In addition, and likely what you are referring to, is setting an even stricter alpha for additional unplanned analyses. Therefore, if we are to run additional exploratory order analyses, with many more analyses than planned will set the alpha to even lower at .001.

> **.14. I confess I don't understand why analyses were run with simulated data and I have nothing to say about that section of the report. To the extent that the fabricated data illustrate the kinds of analyses that will be run and the kinds of tables and graphs that will be generated, I think it all looks O.K.**

Thank you, we appreciate the support.