Reply to decision letter reviews: #176

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

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

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

A track-changes manuscript is provided with the file:
PCIRR-RNR-Fischhoff et al 1978-replication & extension-main manuscript-track-changes.docx

Summary of changes

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

Section	Actions taken in the current manuscript
General	Ed: We attempted to improve details on study rationale, clarified methodological details, sampling, design and analytic decisions, and elaborated on deviations from the replication study. R1: We made numerous amendments to clarify and elaborate the reasons for the replication and the deviations from the original. R2: We amended to clarify statistical analyses and deviations from the original. R3: We made amendments to clarify statistical analyses, deviations from the original design, and limitations to the study. R4: We made amendments to discuss various limitations, clarify changes to the original, clarify instructions and debrief in Qualtrics.

Section	Actions taken in the current manuscript
Introduction	R1: We shortened the Choice of study section and reorganized and amended the Overview of the replication section to clarify the justifications for the replication. R2: We amended the Overview of the replication section to explain in more detail the improvements in our statistical analysis R3: We amended the Overview of the replication section to clarify our expectations regarding our main hypothesis and the reasoning behind them. R4: We expanded and clarified our reasoning for the extension in the Overview of the replication and extensions, Extensions section.
Methods	R1: We added detail on why we expect to find support for the main hypothesis, clarified the Power and sensitivity and Participants sections, and elaborated on deviations from original and evaluation criteria for replication findings. R2: We elaborated on the revised list of 18 items and updated the sensitivity analysis to account for a reduction in sample size due to exclusions. R3: We amended and clarified the Power and sensitivity analysis section to update for two-tailed tests and added more detail to the Measurements and data analysis section. R4: We provided more clarity on the deletion of original items and clarified Qualtrics instructions based on review suggestions.
Results	R2: We amended the independent samples t-test reporting to remove reference to Bayes analysis. R3: We updated the independent samples t-test reporting to remove reference to Student t-tests.
Discussion	We added several paragraphs to address limitations of the current project raised by reviewers, including regarding comparison across studies, ordering effects and confounds, Covid-19 generalizability, and individual numeracy.
Supplementary materials	R2/R3/R4: We added a table to explain the rationale for deletion of each of the 16 items deleted from the original study. R4: We reorganized the table presenting the items in the survey to make comparison of the list across studies easier.

Note. Ed = Editor, R1/R2/R3/R4 = Reviewer 1/2/3/4

Response to Editor: Prof. Chris Chambers

> **I have now received detailed and constructive evaluations from four reviewers. As you will see, the reviews are broadly enthusiastic about the submission and are rich in suggestions for optimising both the study design and quality of reporting in the Stage 1 manuscript. Among the wider headline issues, the reviews prompt for greater consideration of study rationale and background literature, clarification (and addition) of a range of vital methodological details, justification of sampling, design and analytic decisions, and justification of procedural deviations from the replication study. From an editorial perspective, all of the issues raised seem addressable, therefore I am pleased to invite you to address the comments in a comprehensive revision and response.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit. We revised the manuscript and addressed the feedback. Below we answer each of the reviewers' feedback point by point.

Response to Reviewer #1: Prof. Richard Brown

> **I enjoyed reading this detailed and well-written document. This registered report provides a good account of detailing a replication attempt. I believe that Fischoff et al. (1978) is a suitable study to replicate, and the addition of COVID-related items provides an interesting modern context. I have identified several areas which I feel require extensive revision to clarify to the reader 1) why a replication is needed at this time, 2) how will this replication provide new relevance in addition to that conducted in 2016, and 3) why there are considerable deviations from the original study protocols.**

Thank you for the positive opening note and the detailed and constructive comments. We sincerely appreciate the effort and time spent in your review.

> **You only use 14/30 of the original study items. These deviations from the original, and subsequent replication, protocols make me think that it is important to highlight to the reader that this should be considered a partial, or perhaps conceptual replication of the original study.**

Thank you for the suggestion. Yes, we categorized this study as falling between a close and far (i.e., conceptual) replication according to the criteria by LeBel et al., (2018). We added the following sentence to the final paragraph of the "Background" section of the to make this more clear.

> "Due to the number of deviations from Fischhoff et al. (1978) we categorized this study as falling between a close and far (i.e., conceptual) replication according to the criteria set forth in LeBel et al., (2018)."

Choice of study for replication: Fischhoff et al. (1978) section.

This section is too long and does not seem appropriate for scholarly publication. It gives the sense that you have chosen to replicate this study mainly because it is popular. For example, discussing google scholar citations, and contacting the authors, may be suitable for a graduate thesis, but not a published article. To justify the need for this replication study, greater attention should be given to highlighting why replicating historical findings is important (expand on the final sentence of this section), and why this specific replication is relevant now. The question of timing is particularly important given that you highlight the recent replication in 2016. Is an increased sample your main contribution? Are there COVID-related contextual factors that you think are important to capture with respect to the original findings?

Thank you for encouraging us to improve our introduction to the paper. The structure of this section is largely consistent with many of our previous publications (e.g., Adelina and Feldman, 2021; Efendić et al., 2021) and our other in-process PCIRRs that received an IPA (e.g., Li & Feldman, 2022; Zhu & Feldman, 2022), in terms of reasons to replicate.

To address your suggestions, we shortened this section by moving paragraphs relating to specific improvements to methodology and the statistical analyses to the following section. In addition, we prioritized the importance of replicating historical findings and made the details of our contributions more clear in this section, pointing out specifically the increased power of the study, improved methodology and statistical approach, and extension to the COVID-19 pandemic.

References:

Adelina, N., & Feldman, G. (2021). Are Past and Future Selves Perceived Differently from Present Self? Replication and Extension of Pronin and Ross (2006) Temporal Differences in Trait Self-Ascription. *International Review of Social Psychology*, 34(1): 29, 1–16. http://doi.org/10.5334/irsp.571

Efendić, E., Chandrashekar, S. P., Lee, C. S., Yeung, L. Y., Kim, M. J., Lee, C. Y., & Feldman, G. (2021). Risky Therefore Not Beneficial: Replication and Extension of Finucane et al.'s (2000) Affect Heuristic Experiment. *Social Psychological and Personality Science*. https://doi.org/10.1177/19485506211056761

Li, M. & Feldman, G. (2022) Revisiting mental accounting classic paradigms: Replication of the experiments reviewed in Thaler (1999). Received Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/4ps8m/ [IPA]

Zhu, M. & Feldman. G. (2022). Revisiting the links between numeracy and decision making: Replication of Peters et al. (2006) with an extension examining confidence. Received

> Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/8z6ga/ [IPA]
>
> **Methods section**
> **You should justify why you expect to find support the original negative association between perceived benefit and perceived risk ratings, given that Fox-Glassman and Weber (2016) failed to find support. Why do you think the former and not the latter will be repeated?**

We believe that we will find the negative correlation because it has been found elsewhere (Alhakami and Slovic, 1994; Finucane et al., 2000; McDaniels et al., 1997; Slovic et al., 1987), including by our own team in a recent replication finding near identical effects two decades later in two large samples (Efendić et al., 2021).

To try and better clarify this, we added further details to justify why we believe we will find support for the negative correlation in the Overview of the replication and extensions section.

> "In our replication we focused primarily on the negative relationship between perceived risks and perceived benefits. This relationship has been demonstrated in numerous studies since Fischhoff et al. (1978) (Alhakami and Slovic, 1994; Finucane et al., 2000; McDaniels et al., 1997; Skagerlund et al., 2020; Slovic et al., 1987), most recently in a replication of Finucane et al. (2000) conducted two decades after the original with samples from the US and the UK (Efendić et al., 2021). Accordingly, we expect results to show support for the negative correlation between perceived risk and perceived benefit. Our main test for this hypothesis is by examining participant-level risk-benefit associations in an extension, explained in detail in section "Joint risks-benefits condition" below. In addition, to make the most of the replicated design we will also be conducting independent samples t-tests examining differences in participants' perceived risk and perceived benefit ratings. In the supplementary materials, we summarized the key findings in Fischhoff et al. (1978) and Fox-Glassman and Weber (2016) in Table 2 and our deviations from the original and Fox-Glassman and Weber (2016) in Table 3."

In addition, one possible reason that Fox-Glassman and Weber (2016) found mixed results is likely due to design and methodology reasons that made it unlikely to find support for the link: underpowered design and analyses conducted on the item-level with very few items. We explained this in detail in the Overview of the replication and extensions section, reproduced below.

> "For the core part of their analyses, Fischhoff et al. (1978) and Fox-Glassman and Weber (2016) then used item-level mean ratings to correlate and regress results across these two conditions. However, due to the small number of items used in both studies, the ability to detect significance in the relationship between ratings differences on an item-level would require an extremely large and somewhat unlikely effect given common correlations in social psychology. We believe this may explain the mixed results present in both studies. To be able to address the research question we would

> require either many more items, or an analysis on a participant rather than an item level. To improve the study, we modified the analysis of the two conditions to instead perform the only participant level analysis suitable for this design: an independent samples t-tests comparing the participant-level ratings for each item. We believe this provides more accurate and reliable results with respect to the differences between the perceived risk and perceived benefit ratings."

Finally, we further added the following to the Extension 2 sub-section of the Extensions section to explain that our test for negative correlation will be on the participant level to address this limitation:

> "The third condition (Task 1c explained in detail below) will ask participants to rate both perceived risk and perceived benefit, thereby allowing for testing of correlation between perceived risk and perceived benefit ratings at the participant level as opposed to the item-level. This is an improvement to the design of the original study as it will provide the test needed to address the core hypothesis underlying the original study: the relationship between perceived risks and benefits. We expect this condition to show a negative correlation between perceived risk and perceived benefit consistent with numerous studies since Fischhoff et al. (1978) (Alhakami and Slovic, 1994; Efendić et al., 2021; Finucane et al., 2000; McDaniels et al., 1997; Skagerlund et al., 2020; Slovic et al., 1987)."

References:

Alhakami, A. S., & Slovic, P. (1994). A Psychological Study of the Inverse Relationship Between Perceived Risk and Perceived Benefit. *Risk Analysis, 14(6), 1085–1096.* doi:10.1111/j.1539-6924.1994.tb00080.x

Efendić, E., Chandrashekar, S. P., Lee, C. S., Yeung, L. Y., Kim, M. J., Lee, C. Y., & Feldman, G. (2021). Risky Therefore Not Beneficial: Replication and Extension of Finucane et al.'s (2000) Affect Heuristic Experiment. Social Psychological and Personality Science. https://doi.org/10.1177/19485506211056761

Finucane, M.L., Alhakami, A.S., Slovic, P., & Johnson, S.M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13,* 1-17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

McDaniels, T. L., Axelrod, L. J., Cavanagh, N. S., & Slovic, P. (1997). Perception of ecological risk to water environments. *Risk Analysis, 17*(3), 341–352. https://doi.org/10.1111/j.1539-6924.1997.tb00872.x

Skagerlund, K., Forsblad, M., Slovic, P., & Västfjäll, D. (2020). *The Affect Heuristic and Risk Perception – Stability Across Elicitation Methods and Individual Cognitive Abilities. Frontiers in Psychology, 11.* doi:10.3389/fpsyg.2020.00970

Slovic, P. (1987). Perception of risk. *Science*, 236(4799), 280–285. https://doi.org/10.1126/science.3563507

Power and sensitivity analyses section.
This section is confusing. "We aimed for a sample of 1000 participants, 333/4 in each condition, which a sensitivity analysis indicated would allow the detection of independent samples t-test of the two conditions" This first suggests 3 conditions (1000/3) then states 2 conditions. You later state "Participants are randomly assigned to either Task 1a, Task 1b or Task 1c" suggesting 3 conditions. Please clarify and word this more clearly. Also, your stage 1 snapshot states you are going for .95 power not .8 – make sure these are consistent.
Given that you have stated that the core of this replication is the relationship between perceived risk and perceived benefit, this should serve as the basis for your main power analysis. Highlight this. Also, why is this based on correlations of .14? This is not a fatal point, just specifying this value needs explaining.

Thank you for the opportunity to elaborate and clarify.

Our study has three conditions and we are aiming for .8 power. The .14 (now amended to .15) represents the minimum effect we can detect based on the 1000 size sample, which should be sufficient to detect a relationship if it in fact exists.

We amended the "Power and sensitivity analysis" section to make this point more clear as well as address a number of other reviewer comments:

> "We aimed for a sample of 1000 participants, to be evenly split among our three conditions (Task 1a, Task 1b and Task 1, as explained below), resulting in 333/4 in each condition. As explained more fully below, a data analysis strategy contemplates conducting independent samples t-tests on the results from Task 1a and Task 1b. A sensitivity analysis indicated that a sample size of 333 participants in each of these two conditions would allow the detection of independent samples t-test with an effect of d = 0.32 (given 333 in each condition for two condition comparisons, power = 80%, alpha = 0.1%, two-tailed), traditionally considered a medium effect. Separately, for Task 1c, we will be conducting a correlations analysis. A sensitivity analysis indicated that would allow us to detect correlations of r = 0.15 (given single condition of 333, power = 80%, alpha = 5%, two-tailed), traditionally considered a small to medium effect."

> **Participants section.**
> **Instead of listing CloudResearch options and Qualtrics functions (which will be a little value to the reader), it is important that you explain the criteria for participation of your target population. Can anyone participate? Will it be a representative sample? Are there any relevant restrictions on participation? Will there be any criteria for quotas/exclusion of participants?**

We were aiming for full transparency to allow others familiar with the platform to understand the data collection context. Participation in the survey subject to our standard requirements using the CloudResearch/MTurk platforms for approval rating, number of completed tasks, and location.

We added the following to the second paragraph of the Participants section (previously in the supplementary).

> "We will define the HIT for participants that (1) have a HIT Approval Rate between 95% and 100%; (2) have between 5,000 and 1,000,000 tasks, and (3) are located in the United States."

We will also be updating the "Additional information about the study" that includes all information about the data collection process, following the data collection.

With respect to whether the sample will be representative of the broader United States population, this was not something we were aiming for nor would we be able to provide a prediction, and we will not take any steps to ensure that. However, Fox-Glassman and Weber (2016) also used Amazon Mechanical Turk and recruited a sample that was argued to be generally representative. MTurk platform is one of the most widely used platforms for behavioral research in the last decade, and CloudResearch helps address and overcome many of the challenges in conducting research on MTurk.

> **With respect to who your sample will be, and the extent to which they are representative of the broader population, you should refer back to the sampling of Fischoff. You have highlighted that the original study was too small in sample size. Additionally, it was a sample of a specific group of people affiliated with a political group known for enacting societal change. In my opinion, this makes it highly likely that their opinions may not be representative. This may explain why the original study sample broadly reported that 'serious action' was needed to mitigate the level of most risks, whereas this was not replicated in the 2016 study. Again, further explanation is needed to fully explain the sizeable jump in sample (beyond what may normally be expected in replication attempts).**

Thank you, this is a good point. We appreciate the suggestion.

We added the following to the end of the Participants section to point out this additional benefit of using MTurk:

> "As noted in Fox-Glassman and Weber (2016), Fischhoff et al. (1978) did not provide a breakdown of its sample population's demographics. However, it did note that the participants were all members of the Oregon League of Women Voters, which was described in Fischhoff et al. (1978) as "a generally liberal, environmentally minded group". Accordingly, in addition to allowing an increased sample size, the use of Amazon Mechanical Turk offers the potential for a more diverse sample population than the original study."

> **In terms of remuneration, this amount appears to be well below half of what might be considered a living wage hourly rate in the US. Are there any ethical concerns around valuing participants time?**

We applied the US federal minimum wage. We followed the typical approach used for fair wage, also implemented by other platforms that enforce a similar minimum (e.g., Prolific). In addition, at the end of all our studies, we include a question asking about pay fairness, keep track of responses, and take these into account in our pretest when setting the wage and paying bonuses when needed (described in the methods section).

If needed, we are happy to revisit this given clear editorial guidelines.

> **I got the sense that the dramatic reduction in study duration was due to the fact you had already chosen the online surveying platform. Therefore the choice of method (MTurk) seems to supersede the replication methodology (justification for dramatically reducing study duration from 90-120 minutes to 10-20 minutes). I understand the practicalities of this in terms of funding etc, so I'm not suggesting you change this. Perhaps try to word it more that there is a trade-off between needing to dramatically increase the sample to identify whether there is really an effect going on, but that the cost of this is that you need to alter the protocol to reduce time.**

Thank you for this suggestion.

We should first note that while this seems especially urgent for online samples, we consider this change relevant and needed for all samples and participants. A two-hours study would be a burden to anyone, and is likely to greatly affect responses due to fatigue and loss of concentration, especially given how repetitive this survey is.

A second point to note is that the changes in length were mostly by focusing our investigation on the items that we felt were most relevant, and having participants rate 2 out of 9 characteristics in that section of the survey. We aimed to focus our examination on the core risk-benefits link. We therefore consider both these changes as improving on the original's design and reducing cognitive burden on participants, thus making it far more likely to reduce noise and find support for the core effect, with little to no trade-off.

We updated the "Choice of study for replication" and "Overview of the replication and extensions" sections aiming to explain more clearly the changes we made to the methodology and the reasons for them.

We also made many other changes to the section per your suggestions on your second comment, so have not reproduced the changes here.

> **There are also changes to the original protocol such as the original had judgements on a 10 point scale, yours in 1000. Is there a justification for this?**

Yes, we appreciate the suggestion to improve on explaining these points.

The scale in Fischhoff et al. (2016) required a minimum of 10, but participants were free to select their numerical ratings within any range without any limitations other than instructions to make the ratings relative to and consistent with each other.

We believe our design improves on the original for a number of reasons, including (i) allowing zero (i.e., no benefit or no risk) to be a response option, (ii) standardizing the range of possible responses, and (iii) reducing cognitive burden through the use of a sliding scale.

We detailed each of these deviations and the reasons for them in Table 3 of the supplementary materials and added the following reference in the "Overview of the replication and extension" section:

> "We provided a full list of deviations and explanations for the deviations between the original study and Fox-Glassman and Weber (2016) and the present replication in Table 3 in the supplementary materials."

> **Extensions section.**
> **Of the 4 listed items relating to the pandemic, I struggle to see the direct relevance of your inclusion of 'Biological weapons' and would suggest more obvious activities relevant to mitigating the risks of the pandemic like mask wearing.**

Thank you for the feedback to rethink this. We appreciate this suggestion.

We agree that "Biological weapons" is not the best fit with the rest of the pandemic items and therefore replaced it throughout the study with "Experimentation with biological viruses". Separately, we excluded more obvious activities such as mask wearing due to the original study's narrow definition of risk, which was limited to "any risk of dying as a consequence" of the relevant item. While we expanded the definition to "any risk of dying or increased likelihood of dying", we determined to not expand it further to a more general definition of risk and do not think it is broad enough to capture more general items.

> **In the analysis section in the original study, they highlight their rationale for using geometric not arithmetic means. Is there a justification for your deviation from the original analysis?**

In the original study, which allowed participants to create their own ranges, the use of geometric means was to exclude extreme values from the data analysis. We accomplish this by limiting the scale responses to 1000. In addition, using arithmetic means allows us to use zero as a possible response for participants, which we believe is an improvement as participants may believe that some items have no perceived benefit or perceived risk. We detailed each of these deviations and the reasons for them in Table 3 of the supplementary materials and added the following reference in the "Overview of the replication and extension" section:

> "We provided a full list of deviations and explanations for the deviations between the original study and Fox-Glassman and Weber (2016) and the present replication in Table 3 in the supplementary materials."

> **Evaluation criteria for replication findings section.**
> **This section needs elaboration. Which specific statistical tests, with which specific variables, which directions of effect etc? These need to be stated**

> **much more clearly so there can be no doubt as to what will and will not constitute a successful replication.**

Given the number of deviations from the original our comparison for purposes of evaluating the replication will be limited. We will not be able to compare effect sizes and will instead indicate whether we found a signal in support of the hypothesized effects. For Task 1a/1b, this will mean t-tests comparing mean individual ratings for each item individually and the 14 replication items together. We expect these t-tests to show that the two groups do indeed rate perceived risk and perceived benefit differently. For Task 1c, we will be able to conduct correlation and linear regression analyses, however, the design of the task is fundamentally different from the original study and will not be directly comparable. That said, we do expect the analysis to reveal a negative correlation between perceived risk and perceived benefit.

We amended the section to read:

> "We aimed to compare this study with the original findings in the target article. Given the number of deviations from the original we would not be able to compare effect sizes and will instead indicate whether we found a signal in support of the hypothesized effects and whether it was in the same direction as in the original study, instead of comparing effect sizes. In particular, we will conduct independent samples t-tests for Tasks 1a/1b in order to determine whether participants rate perceived risks differently than perceived benefits. We expect these t-tests to show that the two groups do indeed rate perceived risk and perceived benefit differently. For Task 1c, we will conduct correlation and linear regression analyses, however, the design of the task is fundamentally different from the original study and will not be directly comparable. However, we do expect the analysis to reveal a negative correlation between perceived risk and perceived benefit, consistent with other studies since Fischhoff et al. (1978) (Alhakami and Slovic, 1994; Efendić et al., 2021; Finucane et al., 2000; McDaniels et al., 1997; Slovic et al., 1987)."

We also amended the “Comparing this to original findings” section to read:

> “Since the simulated dataset generated random noise, the comparison between this study and findings in the original is irrelevant, and will only be completed after data collection. We will aim to compare the results of the replication to the original findings to the extent possible based on the criteria by LeBel et al. (2019) (see supplementary materials for more details). However, given the number of deviations from the original we will not be able to compare effect sizes and will instead indicate whether we found a signal in support of the hypothesized effects and whether it was in the same direction as in the original study, instead of comparing effect sizes. In particular, we will conduct independent samples t-tests for Tasks 1a/1b in order to determine whether participants rate perceived risks differently than perceived benefits. For Task 1c, we will conduct correlation and linear regression analyses, however, the design of the task is fundamentally different from the original study and will not be directly comparable.”

> **Introduction section.**
> **When introducing the work of Fischoff at the beginning, it might be useful to also briefly discuss the work of Chauncey Starr to highlight where this study sits in the context of the historical literature and the difference between ‘revealed preferences’ and ‘expressed preferences’.**

Thank you for the suggestion. We considered including a more detailed historical background of the study as it provides helpful context for understanding the use of a psychometric design. In the end we decided to leave it out in order to focus on the perceived risk/benefit relationship.

> **You state the “cconduct an independent replication of the negative correlation” - the goal should be to replicate the study protocols, not specifically to reproduce a result.**

Thank you for pointing this out. We changed the references to replication of the negative correlation to match your suggestion.

> **Sometimes you refer to the relationship between perceived risk and perceived benefit, sometimes just to the risk/benefit relationship. It is important to be consistent and highlight to the reader that it is the perceived, not objective, relationship that is being studied.**

Thank you for pointing this out. We adjusted the manuscript to make all references to perceived risk and perceived benefit consistent.

Pre-registration and open-science section.
You don't need a section for this, similar to my comment about the choice of study replication section, this reads more like a student thesis than publishable article, referring the reader to the osf link is sufficient.

Thank you for the suggestion. This seems like a matter of personal taste, so we hope that you will understand us deciding to keep this, as this is standard in all of our replication publications. For examples, please see Adelina and Feldman (2021), Korbmacher et al. (2022), and the recent project receiving IPA from PCIRR - Li and Feldman (2022) and Zhu and Feldman (2022).

If needed, we will gladly amend this given clear editorial guidelines,

References:

Adelina, N., & Feldman, G. (2021). Are Past and Future Selves Perceived Differently from Present Self? Replication and Extension of Pronin and Ross (2006) Temporal Differences in Trait Self-Ascription. *International Review of Social Psychology*, 34(1): 29, 1–16. http://doi.org/10.5334/irsp.571

Korbmacher, M., Kwan, C., & Feldman, G. (2022). Both better and worse than others depending on difficulty: Replication and extensions of Kruger's (1999) above and below average effects. *Judgment and Decision Making*. Retrieved from: https://osf.io/7yfkc/

Li, M. & Feldman, G. (2022) Revisiting mental accounting classic paradigms: Replication of the experiments reviewed in Thaler (1999). Received Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/4ps8m/ [IPA]

Zhu, M. & Feldman. G. (2022). Revisiting the links between numeracy and decision making: Replication of Peters et al. (2006) with an extension examining confidence. Received Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/8z6ga/ [IPA]

Additional suggestions
You may wish to consider the influence of individual level numeracy. Recent research has highlighted that people can struggle to numerically express their beliefs and perceptions of the level of risk they experience. This is particularly true when trying to use large numbers (for instance providing a score out of 1000). For example, Raude et al. (2021) recently reported that the magnitude of the primary bias in risk (overestimating uncommon risks and underestimating common risks) varies as a function of the respondents' individual level of numeracy. Given this recent finding that a famous effect within the risk perception literature is heavily influenced by numeracy, you may consider adding a short measure of

numeracy to consider whether the Fischoff effect is also partially driven by this factor.

Thank you, we appreciate the suggestion and gave this much consideration when revisiting the study design. The original study's instructions contained a somewhat complicated description regarding how to calculate the relative numerical risk and benefit scores for the items. We believed this to be somewhat challenging for MTurk workers and revised the instructions to be much more straightforward. While we agree that a numeracy measure may be helpful, we are hesitant to add an additional element to an already lengthy and complex study.

That said, we added a Future directions subsection pointing this out and suggesting it for future research:

> "Recent research has indicated that people have difficulty with numerical expression of their own risk judgments. In particular, Raude et al. (2021) found that individual numeracy plays an important role in the magnitude by which people overestimate the perceived riskiness of certain common illnesses. In the current study, we did not measure participants' individual numeracy and as a result, we are unable to report if the effects we observed are influenced by numeracy. Future research adopting a similar methodology may consider an individual numeracy measure to test whether, as has been shown elsewhere, numeracy affects the perceived risk and perceived benefit relationship."

Additionally, having looked at your Qualtrics survey, you may wish to consider randomising the order of the 18 items to avoid any order effects. These are common to occur in questions with large numbers of sliding scales.

Thank you for encouraging us to better clarify our decision on this point.

Given the study design of participants using the same 18 items across multiple tasks involving multiple scales, our primary concern was reducing cognitive burden on participants. We tried different designs and keeping item order seems to be the clearest and easiest.

Accordingly, we chose to group similar items together rather than randomize them for each task. We updated the Methods section with the following to explain this point:

> "Fischhoff et al. (1978) did not specify the order in which the 30 items were presented to participants. In order to control for the potential impact of ordering effects, Fox-Glassman and Weber (2016) randomized the order of presentation. In the current study, we grouped items together based on similarity and presented them uniformly across all three tasks of the study. For instance, we grouped together "nuclear power" and "electric power", "motor vehicles" and "general aviation", and "contraceptives", "prescription antibiotics", "surgery", and "X-rays". While this may create the potential

> for the impact of ordering effects, we believe this is an improvement in the study design as it should significantly reduce cognitive burden when participants are moving from Task 1 through Task 3 to deal with the same 18 items across different scales."

> **You may also look to highlight, either in the existing introduction or in the later discussion, some of the developments in understanding risk perceptions since 1978 that may influence our interpretation of the results. For example, my colleagues and I continue to explore the environmental and informational cues of risk perceptions, which can often differ depending on the demographic characteristics of the sample. Given your extension of the replication to COVID behaviours, see our recent work on risk perceptions during the pandemic and assessment of protective behaviours https://link.springer.com/article/10.1007/s10389-021-01543-9 & https://www.tandfonline.com/eprint/BSGYMZUI79CSNGD9VGRH/full?target=10.1080/13669877.2021.1908403.**

Thank you for sharing this work with us - it is indeed relevant. We added the paragraph below in the Extensions section and will also come back to it once we have our data and results for drafting the discussion section.

> "Indeed, the relationship between COVID-19 risk perception has been associated with adherence to pandemic prevention measures (Brown and Pepper, 2021) and further insight may be instructive."

Response to Reviewer #2: Prof. Toby Wise

> **This study intends to replicate the finding demonstrated in Fischhoff et al. (1978) that perceived risk and perceived benefit are inversely correlated. This is a good candidate for replication, as it is clearly a highly cited study, but as yet there have been no well-powered and/or pre-registered attempts to replicate its findings. Additionally, the COVID-19 pandemic has illustrated the real-world relevance of such results making it timely.**

Thank you for the positive opening note and the detailed and constructive comments.

> **The study design appears largely well-thought through and clearly described, although I would like to see some more details on how the methods differ from the original study. There are also some potential statistical issues that could be addressed.**

Thank you for the suggestion. In our initial submission we already provided details in the supplementary materials with a table summarizing these differences. To make that easier to find in the main manuscript, in the "Overview of the replication and extensions" section, we added a clear references to the table:

> "We provided a full list of deviations and explanations for the deviations between the original study and Fox-Glassman and Weber (2016) and the present replication in Table 3 in the supplementary materials."

> **The planned sample size doesn't appear to account for potential exclusions due to poor data quality, will it still be well-powered even if e.g., 10% of subjects need to be excluded?**

Based on the current expected sample size of 1000 participants, we expect 333 participants per condition. A 10% reduction in the number of participants would result in ~33 fewer participants per condition, meaning ~300 per participant. This would still capture an effect of .34 for independent samples t-test and .16 for correlation, which we believe would be sufficient for the stated purposes in the study.

We updated the "Power and sensitivity analyses" section with the following:

> "Following data collection, we will provide an updated sensitivity analysis for any reduction due to exclusions. A 10% reduction in the number of participants would result in ~33 fewer participants per condition, meaning in ~300 per condition. This would still capture an effect of .34 for independent samples t-test and .16 for correlation, which we believe would be sufficient for the stated purposes in the study."

> **Subjects will be recruited using MTurk – there are a couple of papers showing that Prolific provides better data quality (e.g.,**

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3765448) and it may be worth considering the choice of platform to maximise data quality

Thank you for the chance to elaborate.

If you read closely the article that you cited (the final version is on: https://link.springer.com/article/10.3758/s13428-021-01694-3) you will notice that the concern is for running MTurk participants without CloudResearch. When collection data on MTurk using CloudResearch, that paper summarizes equally high quality data collection (in their "discussion" subsection):

> "Between Prolific and CR, there seemed to be an advantage for CR participants in relation to passing ACQs, but an advantage to Prolific participants in all the questions that tested honesty. In other words, while CR participants were more attentive than Prolific participants, they also took more opportunities to act dishonestly. However, despite these differences being statistically significant, both platforms showed high rates of attention (especially compared to MTurk), and the difference in cheating was also not very large. In addition, the platforms showed comparable results on the other aspects of comprehension and reliability, and had very similar overall data quality scores. Thus, it can be concluded that when the data quality pre-screening filters are on, data quality from both Prolific and CR is similarly high."

As we explained in our manuscript we not only used the advanced filtering in CloudResearch, but went above and beyond that in applying all our experience and gained expertise in ensuring high quality data.

We receive this comment quite often from reviewers. We are in the process of writing a manuscript aimed to address this specific issue and help others use the platform and achieve high-quality data collections. In that manuscript, we cited and referred to many of our other completed replication projects using this very approach. We will try and summarize our experience in short below.

We completed over 80 replications of classic findings in judgment and decision making using MTurk online samples (see https://mgto.org/pre-registered-replications/), and our experience has been that these samples are very reliable, at least for replications in judgment and decision making.

There is much that we can share on that, but briefly:

1. Our successful replication rate is currently at 68% (+12% mixed/inconclusive), higher than most other replication rates in other domains. Even in the ones that are

mixed/inconclusive or seemed to have failed we identified reasons that are not related to the samples.

2. When conducting 8 replications in two different online samples, Americans on MTurk and British on Prolific, we found the results highly consistent across the two samples.
    1. Published examples from our team with MTurk versus Prolific:
        i. Efendić, E., Chandrashekar, S., Cheong, S., Yeung, L., Kim, M., Lee, C., & Feldman, G. (2022). Risky therefore not beneficial: Replication and extension of Finucane et al. (2000)'s Affect Heuristic experiment. Social Psychological and Personality Science.
        DOI: 10.1177/19485506211056761
        [Article] [Preprint] [OSF]
        ii. Brick, C., Fillon, A., Yeung, S., Wang, M., Lyu, H., Ho, J., Wong, S. & Feldman, G. (2021). Self-interest is overestimated: Two successful pre-registered replications of Miller and Ratner (1998). Collabra: Psychology, 7(1), 23443.
        DOI: 10.1525/collabra.23443.
        [Article] [Preprint] [OSF]
        iii. Imada, H., Chan, W., Ng, Y., Man, L., Wong, M., Cheng, B., & Feldman, G. (2022). Rewarding more is better for soliciting help, yet more so for cash than for goods: Revisiting and reframing the Tale of Two Markets with replications and extensions of Heyman and Ariely (2004). Collabra: Psychology, 8 (1): 32572.
        [Article] [Preprint] [OSF]
    2. See summary tweet:
    https://twitter.com/giladfeldman/status/1215175786543534090?s=20
3. In a number of replications, when we conducted replications on both students samples and online on Mturk, we found the findings consistent across the two samples. Published examples from our team:
    1. Chandrashekar, S. P., Yeung, S., Yau, K., Cheung, C., Agarwal, T. K., Wong, C., Pillai, T., Thirlwell, T. N., Leung, W., Li, Y., Tse, C., Cheng, B., Chan, H., & Feldman, G. (2021). Agency and self-other asymmetries in perceived bias and shortcomings: Replications of the Bias Blind Spot and extensions linking to free will beliefs. Judgment and Decision Making, 16(6), 1392-1413.
    [Article] [Preprint] [OSF] [Open access]
    2. Chen, J., Hui, L.S., Yu, T., Feldman, G., Zeng, S., Ching, T., Ng, C., Wu, K., Yuen, C., Lau, T., Cheng, B., & Ng, K. (2021). Foregone opportunities and choosing not to act: Replications of Inaction Inertia effect. Social Psychological and Personality Science, 12(3) 333-345.
    DOI: 10.1177/1948550619900570
    [Article] [Preprint] [OSF]

4. When we ran the exact same replications on Mturk in two time periods, with a time gap of several months to two years, ensuring different participants from the same online platform, we found highly consistent results. Published examples from our team:

    1. Fillon, A., Kutscher, L., & Feldman, G. (2021). Impact of past behavior normality on regret: Meta-analysis of exceptionality effect. Cognition and Emotion, 35(1), 129-149.
    DOI: 10.1080/02699931.2020.1816910
    [Article] [Preprint] [OSF]
    2. Xiao, Q., Lam, C., Piara, R., & Feldman, G. (2021). Revisiting status quo bias: Replication of Samuelson and Zeckhauser (1988). Meta Psychology, 5.
    DOI: 10.15626/MP.2020.2470
    [Article] [Preprint] [OSF]

    It's not entirely clear how the items used differ from those in the original study – it seems as though there are fewer items (14 + 4 rather than the 30 used in the original paper) and the activities/technologies asked about have also changed. It would be good to clarify this a little.

Great point. Thank you for the opportunity to clarify this point.

As background, we shortened the list of items from 30 to 14 primarily to reduce the overall burden/duration of the study. The subset of items was selected based on various criteria that we now explain in the added paragraph and table below.

> "The list of 18 items was based directly on Fischhoff et al. (1978) but reduced to 14 items to reduce overall study duration and cognitive burden. The subset of items was selected based on various criteria including, relevance to current society, relevance to a broader population, duplicativeness, and clarity. For instance, we found that items related to transportation were overrepresented and have deleted "bicycles", "commercial (private) aviation", "motorcycles", and "railroads", while retaining "general aviation" and "motor vehicles". Similarly, a number of items were relevant only to a smaller or limited population due to geographical requirements or other reasons. For instance, we deleted "high-school and college football", "hunting", "mountain climbing", "power mowers", "skiing", and "swimming". In the supplementary materials, we provided the full list of items used in Fischhoff et al. (1978) in Table 4 and the list of deleted items and rationale for deletion in Table 5. In addition to the 14 items taken from the original study, we added four Covid-19-related items to the list as an extension: COVID-19 vaccines, experimentation with biological viruses, lockdowns to address the COVID-19 pandemic, and social distancing to address the COVID-19 pandemic. We provided the full list of items used in the current study in Table 5, and summarized the deviations from Fischhoff et al. (1978) and Fox-Glassman and Weber (2016) in Table 3 in the supplementary materials."

We also added the following table to the supplementary materials stating our rationale for each deleted item.

Table 5

Rationale for deletion of items from Fischhoff et al. (1978)

Deleted Item	Rationale for Deletion
Bicycles	Duplicative of transportation items (motor vehicles and general aviation)
Commercial (private) aviation	Duplicative of general aviation
Fire fighting	Less relevant in the context of a risk/benefit analysis due to the essential nature of the service to society
Food coloring	Duplicative of food preservatives; no longer as relevant given relative prevalence and knowledge of food coloring
High-school and college football	Not relevant to broad segment of population
Home appliances	Less relevant in the context of the risk of dying
Hunting	Not relevant to broad segment of population
Large construction	Less relevant in the context of the risk of dying; Not relevant to broad segment of population
Motorcycles	Duplicative of transportation items (motor vehicles and general aviation)
Mountain climbing	Not relevant to broad segment of population
Police work	Less relevant in the context of a risk/benefit analysis due to the essential nature of the service to society
Power mowers	Not relevant to broad segment of population
Railroads	Duplicative of transportation items (motor vehicles, general aviation)
Skiing	Not relevant to broad segment of population
Spray cans	Previously relevant due to the link between spray cans and ozone depletion, which is no longer relevant
Swimming	Not relevant to broad segment of population

One concern with the number of items is potential lack of power, given that the planned analysis for Tasks 1a/1b involves averaging within item and then performing linear regression (assuming I've understood this correctly) – ultimately, it doesn't matter how many subjects there are, if there are very few items and they aren't strongly correlated this will be underpowered. With 14 items, a correlation of r=.68 could be detected with

> **95% power, but this is probably a lot higher than would be expected. It might be worth adding in more items and reducing the number of subjects.**

Yes, this is one of our main concerns about the original study and the motivation for our different statistical analysis and the addition of a third condition (Task 1c). As you have pointed out, an item-level analysis will likely not have sufficient power to detect a correlation that one would expect in the social sciences. We believe this likely contributed to the mixed-results in both studies. We therefore added a condition that asks participants to rate both risks and benefits, which will allow us to run linear regression on the participant level with sufficient power to test the perceived risk/benefit relationship. For the original design, we changed the original analyses using that design to t-test analyses on the participant level for Tasks 1a/1b, which should give us some indication on whether there are any differences between participant's perceived risk/benefit judgments.

We amended the "Overview of the replication and extensions" section of the manuscript with the following:

> "To improve the study, we modified the analysis of the two conditions to instead perform the only participant level analysis suitable for this design: an independent samples t-tests comparing the participant-level ratings for each item. We believe this provides more accurate and reliable results with respect to the differences between the perceived risk and perceived benefit ratings. In addition, we added a third condition, detailed below as Extension 2, displaying both risks and benefits to participants, which will allow us to run linear regression on the participant level with sufficient power to test the perceived risk/benefit relationship."

> **The items related to COVID are described as exploratory but quite a lot of detail is provided regarding the planned analysis – it would probably be best to either remove this detail (to be filled in once data is collected, avoiding the appearance that it was a planned analysis), or explicitly treat this as pre-planned**

The main difference between exploratory and confirmatory as we understand it has to do with clear hypotheses and predictions. Even for exploratory analyses, it may be valuable to lay out the general strategy in analyzing the direction. In this specific case, we included details about the analysis in this section because in effect the analysis is identical to that of the other main replication items as they will all be part of the same questionnaire.

Though the analysis is essentially the same, we are treating the results as exploratory - that is, whether participants will view the perceived risk and perceived benefit of the new Covid items the same as the other items. We see the inclusion of the analysis plan for the added Covid items to be of very high potential benefit in coordinating us and the reviewers/reader/future-selves and with very little risk, as long as we all align expectations as to what these analyses are meant for.

> **More detail on how outliers will be detected might be useful – e.g., are there statistical tests that will be used to determine whether a data point is an outlier?**

Thank you for the opportunity to clarify this.

After consideration and following on feedback received in other PCIRR submissions handled by the same editor as this manuscript , we decided to remove the outlier strategy in favor of not excluding potential important data. We will not be making any corrections to raw data, and we will be reporting results for both pre and post exclusions, with a comparison in the supplementary. We amended the Outliers and exclusions section of the main manuscript and the Handling outliers: Strategy in the supplementary materials with the following:.

> "The current replication will focus on analyzing and reporting the results of the full sample size and will not attempt to identify outliers. We will not be making any corrections to raw data, and we will be reporting results for both pre and post exclusions, with a comparison in the supplementary. Our generalized exclusion criteria are detailed in the "Exclusion criteria" subsection of supplementary materials."

> **It looks as though Bayes factors will be used to determine support for null hypotheses based on the figures, but this doesn't seem to be described clearly in the methods.**

Thank you for pointing this out.

We will not be using Bayes analyses, that information was provided in the plots as part of the default ggstatsplot package details. We updated our plots in the Results section to remove reference to Bayes factors.

> **It might be worth running the original analyses (e.g., those using geometric means), even if they are flawed, just to enable more direct comparison between studies.**

Thank you for the suggestion.

Given the number and degree of deviations between this study and the prior two, comparisons no longer seem relevant. When comparisons are made, they will be primarily limited to qualitative statements.

> **It's not entirely clear to me how the regression for Task 1c is going to be conducted. The best approach might be to use a multi-level model, allowing**

> **for random slopes across subjects, as this would use all the data available without needing to average anything.**

Thank you for the opportunity to clarify this point. We will be conducting simple correlation/linear regression using the participant-level results for each item and for all items together, without averaging anything. We updated the Perceived risk and perceived benefit: Extension 2 (Task 1c) sub-section of the Results section with the following:

> "We will then test for support for the negative correlation between risk and benefit by conducting correlation and linear regression on the participant-level perceived risk and perceived benefit ratings for each item."

Response to Reviewer #3: Prof. Katherine Fox-Glassman

> **Hi, authors! Before getting into my detailed responses below, I just want to express how glad I am that you're undertaking this RR! I know all too well how much effort this topic is to work on, so I salute you for it. I have a few quibbles, as you'll see below, but overall I think this is an incredibly well-justified and timely study, with an exceptionally thorough plan. I hadn't heard of registered reports when I worked on my dissertation research that eventually became Elke Weber's and my 2016 paper on this topic, but in retrospect I wish we had taken as systematic and well-documented of an approach to replication as you are doing now!**

Thank you for the positive opening note and the detailed, constructive comments. We are very thankful for your openness throughout the process and for sharing your original materials - they were extremely helpful in preparing this study.

> **The proposed hypotheses seem logical, and address weaknesses in the studies to be replicated (e.g., low power; long task duration). Using the authors' own logic, though, it is very possible that the main hypothesis (risk and benefit being negatively correlated) could be expected *not* to replicate: (a) it's one of the findings that is inconsistent between the two prior studies, and (b) as the current authors (rightly) point out: "Fishchoff et al did not have explicit hypotheses relating to its data and analysis, yet reported many findings."**
> **It might help to discuss early on considerations of what it might mean to replicate (or fail to) results about how people perceive risk in studies conducted many decades apart, and in studies conducted (relatively) shortly before vs. during a global pandemic. Since perceptions of risk are highly relative—judged in comparison to other salient risks at the time of elicitation, therefore meaning that the perceived risk of the same activity is unstable even when measured at the same time if it is measured within different arrays of other activities—then it would be reasonable to expect that (a) the gradual changes in the world (technology, typical activities, media reporting, etc.) over a long period of time or/an (b) the sudden changes due to a stressful and alarming global pandemic could/would influence people's perception of the risks of many everyday technologies and activities. All this means it's very hard to predict whether we should expect certain effects to replicate, even if they did represent true positives in the original study. (I genuinely don't know whether I think the original R/B correlation was real or an artefact... but all this said, I think it's 100% worth running a well-powered, careful replication to see if it exists now.**

> **That result might not say much about the previous study, especially if you don't find that correlation now. But if you did find it, that might be suggestive that Fischhoff et al. were capturing a real effect in 1979, and we (F-G & Weber) just didn't have the power to see it in 2016!)**

Yes, this is certainly something we considered. Based on our review of the two studies, we were also unsure what we would replicate. That said, we expect to find some participant-level support for the negative risk-benefits association because it has been found elsewhere (Alhakami and Slovic, 1994; Efendić et al., 2021; Finucane et al., 2000; McDaniels et al., 1997; Slovic et al., 1987).

We added further details on this point in the "Overview of the replication and extensions" section.

> "In our replication we focused primarily on the negative relationship between perceived risks and perceived benefits. This relationship has been demonstrated in numerous studies since Fischhoff et al. (1978) (Alhakami and Slovic, 1994; Finucane et al., 2000; McDaniels et al., 1997; Skagerlund et al., 2020; Slovic et al., 1987), most recently in a replication of Finucane et al. (2000) conducted two decades after the original with samples from the US and the UK (Efendić et al., 2021). Accordingly, we expect results to show support for the negative correlation between perceived risk and perceived benefit. Our main test for this hypothesis is by examining participant-level risk-benefit associations in an extension, explained in detail in section "Joint risks-benefits condition" below. In addition, to make the most of the replicated design we will also be conducting independent samples t-tests examining differences in participants' perceived risk and perceived benefit ratings. In the supplementary materials, we summarized the key findings in Fischhoff et al. (1978) and Fox-Glassman and Weber (2016) in Table 2 and our deviations from the original and Fox-Glassman and Weber (2016) in Table 3"

In addition, we believe that the reason Fox-Glassman and Weber (2016) found mixed results is likely in part due to running their analyses on the item-level with a low number of items. We explained this in detail in the Overview of the replication and extensions section, reproduced below.

> "For the core part of their analyses, Fischhoff et al. (1978) and Fox-Glassman and Weber (2016) then used item-level mean ratings to correlate and regress results across these two conditions. However, due to the small number of items used in both studies, the ability to detect an effect in the relationship between ratings differences on an item-level required a very large and somewhat unlikely effect given typical correlations in social psychology. This may explain in part the mixed results present in both studies. To be able to address the research question the research design would need to be updated to include many more items, or an analysis on a participant rather than on an item level. First, we modified the analysis of the current study design with the two conditions to instead perform the only participant-level analysis suitable: an independent samples t-tests comparing the participant-level ratings for benefits and

> risks of each item. In addition and more relevant for the testing of the benefits-risks link, we added a third condition, detailed below as Extension 2, displaying both risks and benefits to participants, which allows us to examine the risk-benefit associations with sufficient power."

Finally, we added the paragraph below to the Extension 2 sub-section of the Extensions section to explain that our test for negative correlation will be on the participant level:

> "The third condition (Task 1c explained in detail below) will ask participants to rate both perceived risk and perceived benefit, thereby allowing for testing of correlation between perceived risk and perceived benefit ratings at the participant level as opposed to the item-level. This is an improvement to the design of the original study as it will provide the test needed to address the core hypothesis underlying the original study: the relationship between perceived risks and benefits. We expect this condition to show a negative correlation between perceived risk and perceived benefit consistent with numerous studies since Fischhoff et al. (1978) (Alhakami and Slovic, 1994; Efendić et al., 2021; Finucane et al., 2000; McDaniels et al., 1997; Skagerlund et al., 2020; Slovic et al., 1987).*"*

Regarding the relativity of risk judgments and what it means to compare results across decades - this is a very good point. Given the number and significance of changes between the studies, it is unlikely that a true comparison can be made outside of broad qualitative assessment. Accordingly, our focus is primarily on determining whether or not the negative perceived risk/benefit relationship holds in a well-powered and fine-tuned study.

However, to address this point, we began a draft "Limitations" sub-section of the "Discussion" section by adding the following:

> "We made many changes to the target article's and Fox-Glassman and Weber (2016)'s study design. These departures limited our ability to compare between the current study and those two studies. Our list of items was primarily based on the same items used in Fischhoff et al. (1978) and Fox-Glassman and Weber (2016), yet it is possible that in the intervening years since these studies, people's understanding of these items and their attitudes toward their risks and benefits have changed. Moreover, reporting of risk preferences may be sensitive to context, choice options, and elicitation methods (Frey et al., 2017; Jusev et al., 2020). We therefore advise caution regarding drawing any strong conclusions regarding comparisons of our results and these two studies."

References:

Alhakami, A. S., & Slovic, P. (1994). A Psychological Study of the Inverse Relationship Between Perceived Risk and Perceived Benefit. *Risk Analysis, 14(6), 1085–1096.* doi:10.1111/j.1539-6924.1994.tb00080.x

Efendić, E., Chandrashekar, S. P., Lee, C. S., Yeung, L. Y., Kim, M. J., Lee, C. Y., & Feldman, G. (2021). Risky Therefore Not Beneficial: Replication and Extension of Finucane et al.'s (2000) Affect Heuristic Experiment. Social Psychological and Personality Science. https://doi.org/10.1177/19485506211056761

Finucane, M.L., Alhakami, A.S., Slovic, P., & Johnson, S.M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*, 1-17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science advances*, *3*(10), e1701381. https://doi.org/10.1126/sciadv.1701381

Kusev, P., van Schaik, P., Martin, R., Hall, L., and Johansson, P. (2020). Preference reversals during risk elicitation. *Journal of Experimental Psychology: General* 149, 585–589. http://dx.doi.org/10.1037/xge0000655

Skagerlund, K., Forsblad, M., Slovic, P., & Västfjäll, D. (2020). *The Affect Heuristic and Risk Perception – Stability Across Elicitation Methods and Individual Cognitive Abilities. Frontiers in Psychology, 11.* doi:10.3389/fpsyg.2020.00970

Slovic, P. (1987). Perception of risk. *Science*, 236(4799), 280–285. https://doi.org/10.1126/science.3563507

> **I don't see a mathematical justification to use one-tailed t-tests here. A difference in either direction should be considered surprising/unusual, and as such alpha must be distributed across both tails. (I also don't immediately see a theoretical justification—how would you determine whether you'd expect risks to be higher than benefits for each item, or vice versa? Though really this parenthetical question is moot since even an expectation that A is higher than B would not be sufficient justification for a one-sided t-test as long as it is mathematically possible for A to be lower than B.)**

Thank you for pointing this out. Yes, we agree that these should all be two-tailed tests. We updated the "Power and sensitivity analyses" section to reflect this change. The updated section is provided below in response to your additional question regarding the purpose of the t-test and alpha inflation.

> **It isn't clear in this report (or possibly I haven't gotten to it yet, so if it's explained later and I forget to delete this, please disregard!) how the subset of activities/technologies was chosen out of the 30. But it looks like quite a few of them come from the high-unknown/ low-dread quadrant of the original studies. For the best chance at making comparisons / looking at how item placements have shifted over time, items should be taken from across the dread/unknown factor space, e.g., 3-4 from each of the quadrants, with one nearer the origin and the other two capturing the spread (not necessarily the most extreme, but some of the reach) on unknown and dread, respectively. If you don't have your own preferred list from across the factor space, let me know and I can dig up the subset we used when comparing natural hazards to activities/technologies, were we had to do essentially the same thing.**

Thank you for raising this. A similar comment was made by Prof. Toby Wise, please see our response above.

> **The decision to cut out Task 2 for the R+B participants seems like probably a good idea. In your writeup, it would be worth briefly considering whether you might expect any confounds on Task 3 based on whether Ps have completed Task 2 or not. I'm not sure I necessarily would expect that to be a problem, but it's probably worth looking at the Task 3 results between the R+B group and the other 2 conditions to see if there are any systematic differences (if so, that could be worth some follow-up study with more specific hypotheses grounded in theory!).**

Thank you for the suggestion. This is an exploratory direction we can address once we have data and are drafting our discussion section.

We began addressing this point in the "Limitations" section with the following:

> "In addition, participants in Task 1a (the risk raters) and Task 1b (the benefit raters) completed Tasks 2 and 3 whereas participants in Task 1c (rating both risks and benefits) only completed Task 3 without completing Task 2. It is possible that completing Task 2 somehow affects how participants respond to Task 3, and this can be addressed with an exploratory analysis comparing Task 3 completed by Task 1c participants to Task 3 completed by Tasks 1a and 1b participants. "

> **From this writeup, I'm not entirely sure what the purpose of the "t-tests (participant level)" are telling us for Tasks 1a/1b. The mean risk for "pesticides" is compared to the mean benefit for "pesticides"? What theoretical construct would that difference correspond to? My memory (admittedly hazy, due to time and pandemic) of the relevant literature of Affect Heuristic (etc.) is that the theories assume an inverse correlation between perceptions of risks and benefits, but don't speak to any absolute difference between the two. (And is there even reason to believe that the ways people rate risks and benefits even share a common scale? Do you have any hypotheses for whether average risks should be higher or lower than average benefits?) This added analysis seems to invite alpha inflation (especially the plan to run these t-tests individually on all 18 items —are you correcting for multiple comparisons?), especially with the plan to perform them all as one-sided. My suggestion would be to drop these t-tests; if you're set on including them, then more justification for their purpose is needed, they all need to be two-tailed, and they should be adjusted in some way for the fact that you're doing so many of them on data of a common origin.**

Thank you for the comment and the opportunity to clarify.

The reasoning behind the t-tests is due primarily to the between subjects design of the original study. We understand that this was done to reduce the overall duration of the survey, yet the direct outcome is that we cannot perform participant-level analyses across the two groups to examine the core hypothesis. The original's approach to this issue was to conduct an item-level regression analysis. However, the number of items in the study are too few to be able to reliably detect a relationship in both the target and especially in our study. As a result, the only remaining participants-level statistical analysis to look at a between subjects design such as this is by comparing the two groups using an independent samples t-test.

As you point out, this will really only tell us if participants are rating risks differently than they are rating benefits. Unfortunately, this is all that can be done given the design. This is the motivation for adding the new Task 1c - a within subjects design that will enable us to perform an individual-level regression/correlation analysis.

With respect to the alpha inflation and using two-tailed tests, we updated the manuscript to (i) set our sensitivity analysis alpha level to .001 (which is approximately 0.05 divided by 18) and (ii) changed all t-tests to two-tailed tests. Our amended paragraph in the Methods section is:

> "We aimed for a sample of 1000 participants, to be evenly split among our three conditions (Task 1a, Task 1b and Task 1, as explained below), resulting in 333/4 in each condition. As explained more fully below, a data analysis strategy contemplates conducting independent samples t-tests on the results from Task 1a and Task 1b. A sensitivity analysis indicated that a sample size of 333 participants in each of these two conditions would allow the detection of independent samples t-test with an effect of d = 0.32 (given 333 in each condition for two condition comparisons, power = 80%, alpha = 0.1%, two-tailed), traditionally considered a medium effect. Separately, for Task 1c, we will be conducting correlation analysis. A sensitivity analysis indicated that would allow us to detect correlations of r = 0.15 (given single condition of 333, power = 80%, alpha = 5%, two-tailed), traditionally considered a small to medium effect. Following data collection, we will provide an updated sensitivity analysis for any reduction due to exclusions. A 10% reduction in the number of participants would result in ~33 fewer participants per condition, meaning ~300 per condition. This would still capture an effect of .34 for independent samples t-test and .16 for correlation, which we believe would be sufficient for the stated purposes in the study. Our planned sample is several times larger than both Fischhoff et al. (1978) and Fox-Glassman and Weber (2016) that had 75-6 participants."

> **Outliers and exclusions plans seem reasonable, and are quite detailed. Might be worth specifying whether respondents whose data are >3 SD from the mean on one variable will be removed entirely from analysis, or whether that value alone will be dropped. (Sorry though if this is specified somewhere and I missed it!)**

We appreciate the feedback encouraging us to reflect more on this. The other reviewers had similar other questions about the suggested outlier analysis.

After consideration, we decided to remove the outlier strategy in favor of not excluding potential important data. We will not be making any corrections to raw data, and we will be reporting results for both pre and post exclusions, with a comparison in the supplementary. Please see our detailed reply on this point above.

> **Why are both Student's *t* and Welch's *t* planned? Either you have theoretical reason (or past experience) to expect equal variances in the population, and then should run that analysis and benefit from the higher power, or you don't have reason to assume equal variances and so should run the analysis that way with slightly less power but more confidence that your assumptions aren't undermined. In this case, Welch's *t* is almost certainly the appropriate test. (Though per my objection above to the t-tests being performed at all, maybe this point is moot.**

Thank you for catching this. We are planning to use only Welch's *t* in our analysis.

The manuscript has been updated to reflect this by removing references to Student's t-test in Table 6 and in Table 8.

> **I'm having a hard time predicting what effect (if any) it might have to only ask Ps about 2 of the 9 characteristics of risk. This could be worth setting some expectations out for before running the study.**

Yes, this does limit our analyses. We are unable to address the full original analyses, yet we can still learn much about the relationship between the characteristics and perceived risk and perceived benefit.

We detailed all these in Table 3 of the "Measures and data analysis strategy" section as well as by adding the following just after the table:

> "We will focus on conducting correlational analyses to examine the relationship with perceived risks, perceived benefits, and risk characteristics, yet given our design we will not conduct analyses among the characteristics as reported by Fischhoff et al. (1978) and Fox-Glassman and Weber (2016)."

> **Is there any concern that grouping activities/technologies based on relative similarity might create artificial clustering of risk or benefit ratings on those similar items?**

In structuring the study, we considered a number of issues relating to ordering effects and clustering for the items. In the end, we determined that the benefit of reducing overall cognitive burden (by grouping items) outweighed the risk of ordering and clustering effects.

We added the following to the Methods section to explain this:

> "Fischhoff et al. (1978) did not specify the order in which the 30 items were presented to participants. In order to control for the potential impact of ordering effects, Fox-Glassman and Weber (2016) randomized the order of presentation. In the current study, we grouped items together based on similarity and presented them uniformly across all

> three tasks of the study. For instance, we grouped together “nuclear power” and “electric power”, “motor vehicles” and “general aviation”, and “contraceptives”, “prescription antibiotics”, “surgery”, and “X-rays”. While this may create the potential for the impact of ordering effects, we believe this is an improvement in the study design as it should significantly reduce cognitive burden when participants are moving from Task 1 through Task 3 to deal with the same 18 items across different scales.”
>
> **I’m sure this is in there somewhere but I missed it on first readthrough and now can’t find it: is whether Risks or Benefits are rated first vs. second for the new (extension) group of Ps simply randomized?**

Thank you for pointing this out. The ordering of risk and benefit rating will be randomized across participants to control for order effects. we updated the Methods section under sub-section “Task 1c (Extension 2) - Perceived benefit and risk (within subjects)” with the following:

> “The ordering of risk and benefit ratings will be counterbalanced by randomizing which one is presented first to participants.”
>
> **I do have some personal uncertainty about what it will mean to get data on only 2 risk characteristics from each P. I think that’s a very clever way to reduce the study duration, and given the focus on the R/B correlation I think it makes sense that you’ve cut back on Task 3. But I’m not sure what the data from Task 3 is going to give you... it can’t be compared to the prior studies (as you say in your plan), so what is its purpose? Is it worth considering cutting that part of the study entirely? (I say that with reluctance, since that’s the part of the study that is of most theoretical interest to me, personally—I’d be so curious to see an updated risk factor space with COVID included!)**

We should be able to understand the relationship between perceived risk, perceived benefit, and the risk characteristics, and examining these associations seems valuable. The factor structure of the risk characteristics is secondary to that.

In addition, while not a stated aim of the study, one significant benefit of continuing to collect the data for itis to validate the study design for future use. All of our materials will be available for future studies that may be able to test these research questions in a more robust manner.

> **The justification for using arithmetic mean is that your procedure for accounting for outliers makes it unnecessary to use geometric mean. That seems reasonable on its own, but are there any concerns that using the different type of mean could cause difficulty in comparing the planned vs. prior studies? (Maybe the answer is no, since we'd already expect so many differences for other reasons? But could be worth considering.)**

Yes, we will not likely be able to make comparisons between studies a focus of this study. Given the number and significance of the deviations, comparisons, when we make them, will be limited to qualitative statements.

> **1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).**
>
> **Honestly, I'm not sure it's possible to consider enough controls for this kind of question, for some of the reasons discussed above—in short, there are so many possible things that could have changed over 4 decades that it's unclear what either a successful *or* an unsuccessful replication would mean. But in spite of that, I feel strongly that this study is worth carrying out: either way, a better-powered study is called for here, and on its own the question of how people think about the risks and benefits of COVID is a worthy one.**
>
> **One big question to consider before running this study is what you think it would mean if the R/ B correlation was (wasn't) significant without the COVID-related items, but wasn't (was) with them included. I don't necessarily expect that to be the case, but it seems within the realm of possibility that people think differently about COVID-related risks than they do about other risks in their environment (either because COVID is novel, or because we have all learned a lot about it very fast and maybe not very accurately, or because we're working mostly on descriptive information rather than experienced probabilities, or for another reason). It seems like you anticipate something like this too, since you plan to do the t-tests separately for the COVID- related items. So some a priori expectations could be helpful to lay out at this stage.**

We certainly share your interest and curiosity about the potential differences between the way people perceive Covid risks/benefits and other more routine items. As you indicate, there are many factors that might influence a different treatment in people's minds about Covid risk/benefit, including its novelty, the nature of our information gathering about it, its urgency,

fatigue around pandemic and its impact, etc. Given the variegated impact of all these factors, it is quite difficult to set clear expectations for this, hence our plan to leave this as exploratory.

However, we added the section below in the Extensions section to provide more color around the purpose of the COVID extension.

> "The aim of this extension was to gain insights as to people's evaluations and judgments concerning the benefits and risks of various pandemic responses and policies. In particular, we will explore whether participants view the relationship between the perceived risks and perceived benefits of these items differently than other non-pandemic related activities and technologies, and the differences between perceived risks and benefits. If participants do view the relationship differently, this may provide useful insights as to how to structure pandemic related public communications around the pandemic, especially regarding activities and technologies designed to mitigate the pandemic or protect the public. Indeed, the relationship between COVID-19 risk perception has been associated with adherence to pandemic prevention measures (Brown and Pepper, 2021) and further insight may be instructive. Measurements and data analysis concerning the additional extension items will be consistent with the main analysis in the study."

In addition, we elaborated further in the Limitations section to address the potential lack of generalizability of any COVID-related findings:

> "Our study was conducted in May of 2022, while the COVID-19 pandemic was still very much ongoing. During this time, lockdowns, quarantines, mask-wearing policies, not to mention the medical impact of the pandemic, were impacting the general public. Accordingly, the results of this study pertaining to the COVID-19 pandemic may be constrained in terms of its generalizability."

Response to Reviewer #4: Prof. Bjørn Sætrevik

> **Thank you for the opportunity to function as a Stage 1 peer-reviewer for this project. I'm enthusiastic about the registered report format but have so far only participated in them in a limited capacity.**
> **I have been following Feldman's project of preregistered and registered replications for a while, and I'm very impressed with the scope and productivity of the project. I realize that the timeline may be tight, and I hope my involvement and comments do not delay the project too much. This was also a great opportunity for me to look more closely into the classic paper of Fischhoff, Slovic, Lichstenstein, Read & Combs (1978).**

Thank you for the positive opening note and the detailed and constructive comments.

> **Somewhat unusually, the replication target article was not confirmatory research, in that it did not set out to test any specific hypothesis. Instead, the aim appears to have been descriptive, in identifying relationships between aspects of risk evaluation. The paper was nevertheless related to past findings and theory, as it emphasized the comparison to previous findings and approaches, in particular that of Starr (1969). When the current authors use this as a replication target, I will assume that they set out to test whether similar patterns of responses will emerge in a new dataset. In that sense, it is a confirmatory replication of a descriptive target.**

Thank you for the reframing of the paper in this way. It was indeed interesting to think through what the original paper was attempting to do and how to present this particular replication. We generally agree that this is a confirmatory replication of a descriptive target, however, we do have specific expectations around certain parts of the results, most obviously the perceived risk/benefit relationship.

> **An additional aim of the current study appears to be to examine the risk evaluation of activities related to the COVID-19 pandemic. However, this appears to indicate additional research aims beyond what is specifically covered by the three research questions RQ1-3 listed above. I wonder if it would make sense to also state a specific RQ related to the COVID-19 activities, from which to extract more specific hypotheses?From the stage 1 manuscript, it is not entirely clear to me why the current data collection is being performed during a pandemic, with questions relevant to the pandemic. There could be good reasons for this, but I think they should be clearly stated. It could also be discussed what the costs and benefits of doing so could be. Will it affect generalizability? Are the authors hoping for the COVID-19 measures to provide an applied value of the results above the more theoretical contribution that is offered by the other research questions?**

We agree that there are many considerations that should go into the COVID extension and its analysis. Given the scope and scale of this project, our main focus is on the core hypotheses of the target regarding the perceived risk/benefit relationship.

That said, we appreciate the potential limitations of generalizability of any findings from the COVID extension. We added the following to the Limitations section to address this point:

> "Our study was conducted in May of 2022, while the COVID-19 pandemic was still very much ongoing. During this time, lockdowns, quarantines, mask-wearing policies, not to mention the medical impact of the pandemic, were impacting the general public. Accordingly, the results of this study pertaining to the COVID-19 pandemic may be constrained in terms of its generalizability."

In addition, we attempted to elaborate regarding the COVID extension with the following in the Extensions section.

> "The aim of this extension was to gain insights as to people's evaluations and judgments concerning the benefits and risks of various pandemic responses and policies. In particular, we will explore whether participants view the relationship between the perceived risks and perceived benefits of these items differently than other non-pandemic related activities and technologies, and the differences between perceived risks and benefits. If participants do view the relationship differently, this may provide useful insights as to how to structure pandemic related public communications around the pandemic, especially regarding activities and technologies designed to mitigate the pandemic or protect the public. Indeed, the relationship between COVID-19 risk perception has been associated with adherence to pandemic prevention measures (Brown and Pepper, 2021) and further insight may be instructive. Measurements and data analysis concerning the additional extension items will be consistent with the main analysis in the study."

> **The "Hypothesis" section of the snapshot lists three hypotheses (for ranking of risk and benefits for activities, related to RQ1; for the risk/benefit association, related to RQ2; and for identifying the two risk factors, related to RQ3). It also states that additional hypotheses will be introduced for the COVID-19 items. However, the "PCIRR-Study Design Table" only lists a single hypothesis (for RQ1), while the other analyses are listed as "exploratory". Also in the remaining text, there appears to be only the hypothesis for testing the risk/benefit association. Apart from the misalignment between snapshot and manuscript, I think this is unfortunate in itself, since the study will collect sufficient data to test additional hypotheses, and the authors appear to intend to also address these research questions. I think it would be of great value to do so within the framework of registered hypotheses, rather than a purely exploratory approach to the research questions.**

We agree that these are very interesting research questions to pursue, however, given the scope of the project and our intended focus on the perceived risk/benefit relationship, we decided to treat these as exploratory. We updated the manuscript so that inconsistencies around this point are removed.

> **I agree with the current authors that the association between risk and benefit is one of the main findings of the original paper (RQ1), in particular as it is framed in opposition to the previous finding from Starr (1969). However, a perhaps equally important takeaway (in particular in the context of subsequent psychology literature), was the identification of the factors of "dread" (severity) and "technological novelty" as crucial determinants for evaluating and accepting risks (RQ3). The replication target article found these factors to supersede other risk aspects examined in preceding research, and this finding has often been cited in the literature published since then. It could be argued that replicating the RQ3 effect is equally important as replicating the RQ1 effect.**
> **As I understand the stage 1 manuscript, the authors plan to collect and analyze data relevant to RQ3. I think it would be very valuable to have a confirmatory hypothesis to replicate the original finding also for RQ3. I understand that the planned changes to the design will make such a test less powered than the RQ1 test. It may therefore make sense to mark this hypothesis as a secondary aim of the replication. Table 3 of supplementary materials state that there will be limited power to conduct analyses. But if my thinking is correct, shouldn't there be 222 answers for every risk characteristic? Although not overly powerful, I assume that this will be sufficient for some types of analyses. Similarly, it appears that the planned**

> **design will collect data and perform analyses related to RQ2. Again, I think the study will benefit from stating specific hypotheses for this research question and doing this as confirmatory research. As far as I can see, the planned study will have equal power to resolve RQ2 as it will have for RQ1.**
> **That being said, the authors may argue that including RQ2 and RQ3 as confirmatory RQs will detract from the aim of the replication, that the RQ2 and RQ3 finding are not sufficiently established or have previously been sufficiently replicated or that the planned study will have insufficient power to provide clear answers for RQ 2 and RQ3. Such arguments may reasonably be made, but that would raise the question of why RQ2 and RQ3 data are being collected rather than favoring a more efficient design for testing only RQ1.**

We agree that these are very interesting research questions to pursue, however, given the scope and complexity of the project and our intended focus on the perceived risk/benefit relationship, we decided to treat these as exploratory, and therefore will not be elaborating on these in Stage 1. We agree that an analysis might be possible, though much less powered, and therefore of much limited value.

While not a stated aim of the study, one significant benefit of continuing to collect the data for the RQ2 and RQ3 questions is to validate the study design for future use. All of our materials will be available for future studies that may be able to test these research questions in a more robust manner.

> **The Fichhoff et al. (1978) article also reports a number of other findings about the relationships between different aspects in risk evaluation.**
>
> - **The participants expressed that most of the activities should be made safer, a few of them should be made much safer**
> - **Participants that first rate benefits judge risks to be more acceptable than participants who first rate risks**
> - **Perceived benefit has negative relationship to perceived risk, but positive relationship to the level of acceptable risk**
> - **Substantial agreement in ranking of risks and (particularly) in ranking of benefits**
> - **Degree of voluntariness did not mediate the risk/benefit tradeoff (but did so for the tradeoff for "acceptable risk")**
> - **The level of acceptable risk and of perceived/current risk can be predicted with high accuracy from the two risk factors**
> - **Risks are seen as more acceptable after evaluating the benefits**
>
> **To the extent that these findings can also be tested in the current design, I would encourage the authors to state them explicitly in the stage 1 manuscript. I think it**

> **could be valuable for the subsequent stage 2 manuscript to be able to state whether these findings are replicated or not in the new dataset, while referring to a priory expectations.**

We agree that these are all very interesting points from the original study and the 2016 replication. We will certainly be looking at these and many other points once we have data and may report some of them and perhaps include qualitative comparison, for instance on the frequencies of risk acceptability judgments.

However, given the scope and complexity of the project and our intended focus on the perceived risk/benefit relationship, we decided not to treat these as specific research questions but leave those as possible exploratory directions that we may visit at a later stage after data collection.

> **The analyses of the COVID-19 items are clearly marked as being exploratory. I think this is fine if the authors prefer it to be so, but it does seem like a missed opportunity. If the authors expect the overall findings of Fischhoff et al. (1978) to replicate, it would seem reasonable to also expect similar findings for the COVID-items (and it would be interesting if that should fail to emerge). I would therefore encourage the authors to include hypotheses about the generalization of the main findings to their COVID-19 items.**

We certainly share your interest and curiosity about the potential differences between the way people perceive Covid risks/benefits and other more routine items. However, given the scope and complexity of the project and our intended focus on the perceived risk/benefit relationship, we decided to treat it as exploratory.

> **The rationale for the selection of 14 items to replicate was not clear to me. Was the selection made based on something like representing the different risk characteristics evenly? From a quick glance at comparing the selection with the original items, it seems that involuntary risks (those determined by societal decisions on nuclear, electric, weapon regulations, healthcare, transport, food safety) may be overrepresented. Conversely, risks more determined by choice of leisure or vocational activities (firefighting, police work, hunting, football, bicycles, motorcycles, power mowers, skiing, spray cans, swimming) may be underrepresented. I worry that this methodological deviation from the replication target may offer an alternative explanation if the results should deviate from the original. I would recommend justifying the item selection or trying to balance it as well as possible.**

Thank you for raising this. A similar comment was made by Prof. Toby Wise, please see our response above.

> **The replication target study had all the 30 activities printed on cards, and asked participants to first order the cards, and then assign the number 10 to the lowest ordered activity and higher numbers to the others (with no maximum value given). The instructions also tried to explain how the assigned numbers were to be used (i.e., a rating of 12 indicates 20% more risk or benefit than a rating of 10). Participants were encouraged to double-check the relationship between the values they submitted.**
> **I understand the current authors' desire to have a more efficient procedure, reluctance to use 10 as a starting point, and that an efficient way to implement this may be to use a slider going from 0 to 1000. But note that the new procedure skips the step of first ordering the activities, that may have some effects on how they are evaluated. Also, providing a scale may give the participants the idea that the full scale should be used, and the activities should be distributed along the scale. In the original study the emphasis was on evaluating risks and benefits of the activities relative to each other. The changed response mode in the replication may direct the emphasis more towards evaluating "absolute" values of risks and benefits for the activities.**
> **One may argue that the change does not necessarily impact the central research question to be evaluated (i.e., the negative association between risk and benefit). The current authors acknowledge the difference in measures, and claim it to be necessary for faster responses, scalability and reducing cognitive burden. I wonder whether this assumption has been tested through piloting. I would imagine that even with the current response mode, many participants will mentally order and compare between the different activities. The cost in time and cognitive burden may thus be fairly high also in the revised methods (as it has to be done without visual aids). If technically possible, I would recommend trying to implement an ordering stage first, and then a stage of entering numbers to indicate the relative difference between each ranked activity. The instructions could also emphasize the importance that participants compare their responses to risks or benefits, to make sure that they express the intended relative relationships between the activities.**

Thank you for this question and the opportunity to elaborate.

Our original intention was to construct the task as you suggested - have participants drag and drop the items to rate them then follow up with a text field for participants to numerically rate their already ranked items. However, we found this design to be extremely taxing as well as time consuming. In addition, the original instructions, as you have pointed out, provided detailed

calculation instructions. We found these to be overly complicated and assume a high-level of numeracy. We wanted to keep the task simple and doable, especially given our target sample,

We believe that the present study is already much less susceptible to the influence of individual numeracy than the original design. In addition, our design is easy to understand and allows participants to compare and rate items in one setting.

To help ensure that participants are still comparing items, we added instructions asking participants to make the ratings internally consistent.

The original allowed participants to use any range of numbers. While we agree that limiting our scale to 1000, we believe this is a sufficiently large range to allow participants to vary their answers in a way similar to the original, and limit inflation to allow for better accuracy and interpretability.

> **Task 1c where participants will rate both perceived risk and perceived benefit appears to be a useful modification of the design, and segmenting this to its own participant group appears to be a way to control for these effects without deviating from the replication of tasks 1a and 1b. The only disadvantage I can think of is the reduction of statistical power of 1a and 1b. However, it was not clear to be whether the order of the two ratings in task 1c were to be counterbalanced between participants as a control for order-effects (as opposed to e.g., always rating benefits first and risks second)?**

We conducted power analyses in which we detail our aims for detection of effects. We find the planned sample size to be quite sufficient, and so we do not consider this a concern.

The ordering of risk and benefit rating will be randomized across participants to control for order effects. we updated the Methods section under sub-section “Task 1c (Extension 2) - Perceived benefit and risk (within subjects)” with the following:

> “The ordering of risk and benefit ratings will be counterbalanced by randomizing which one is presented first to participants.”

> **Each participant will rate all risk events on two of the nine scales from the original article. This is done in order to give the study a manageable duration. Such an adjustment may be necessary, but it rests on the assumption that answering the two scales when presented on their own is not significantly different from if they were presented amongst the full set of nine scales. This assumption may presents an alternative explanation for diverging results. I would encourage the researchers to consider alternative designs where a subsample answers the full set of the nine scales, in order**

> **to compare the results from those participants to those who answer only two.**
> **An alternative (but weaker) solution could be to ensure that each participant that answers only two scales, will always answer one from each of the two factors ("novelty" and "dread"). The two solutions could also be used in combination.**

Thank you for the helpful suggestions. We considered a number of options around how to treat this part of the survey. While it is a major part of the original study and has given rise to follow-up research, we decided to focus on the risk/benefit relationship. Accordingly, there will be inherent limitations in what we are able to conclude from this portion of the study, which is the main reason we did not outline our expectations for this research question and are treating it as exploratory only.

> **The task 1 instructions use the terms "net" and gross". The replication target article discusses the possibility that these instructions may not have been correctly understood by the participants. I suspect that the terms may be even less familiar for the average modern MTurk worker than it was for the Eugene, Oregon League of Women Voters in the seventies. Could the original meaning of the item be expressed in simpler terms (without deviating too much from the replication target)?**

Thank you for pointing this out. We were generally concerned with the level of numeracy of participants, which motivated us to amend the original instructions to remove anything that might be too mathematical. We agree that the use of gross and net should be amended as well. We updated the Qualtrics instructions in Task 1a to read:

> "Your job is to assess only the benefits on their own, not the benefits which remain after the costs and risks are subtracted out."

and Task 1b to read:

> "Your job is to assess only the risks on their own, not the risks which remain after the benefits are added."

> **After the check of understanding the instructions, some participants may be unsure whether they responded correctly or not. Perhaps you could mention that they will only get feedback on incorrect answers?**

Thank you for this comment. We updated the Qualtrics instructions for Task1 to reflect this.

> **I found it difficult and confusing to rate the category "electric power" when compared to "nuclear power". Electric power (as part of the energy infrastructure) can be powered in a number of ways (coal, solar, hydroelectric, etc.), including nuclear power. The risk and benefits mainly stem from the source of the energy, not from the electric grid itself. Perhaps this confusion is due to the time passed since the original study. I assume that at the time the "electric power" would have been understood as continued use of the current energy sources, while nuclear power was a novel and fairly unused technology. I would recommend considering changing the category "electric power" to something like either "coal-powered energy" or to "electrical appliances". I think the cost of deviating from the replication target is afforded by the reduced confusion for the participant and increased certainty about what participants actually had in mind when answering.**
> **Similarly, I think that the term "motor vehicles" in the original study would have been interpreted as combustion motor vehicles, but in 2022 the same term may be interpreted to include both gas and electric/hybrid vehicles. This may have consequences for how risks are evaluated in terms of emissions (as mentioned in the instructions). Perhaps this activity should be specified as gas powered vehicles, if you would like to compare the responses to the 1978 results?**

Thank you for raising this point. We agree that there may have been changes in the understanding of and attitudes towards the risk and benefit of certain items in the intervening years. Partly, keeping these the same may help us understand if indeed anything has changed. For purposes of revisiting the target's design, we believe it makes most sense to keep these the same. However, we added the following to the "Limitations" sub-section of our "Discussion" section:

> "We made many changes to the target article's and Fox-Glassman and Weber (2016)'s study design. These departures limited our ability to compare between the current study and those two studies. Our list of items was primarily based on the same items used in Fischhoff et al. (1978) and Fox-Glassman and Weber (2016), yet it is possible that in the intervening years since these studies, people's understanding of these items and their attitudes toward their risks and benefits have changed. Moreover, reporting of risk preferences may be sensitive to context, choice options, and elicitation methods (Frey et al., 2017; Jusev et al., 2020). We therefore advise caution regarding drawing any strong conclusions regarding comparisons of our results and these two studies. "

> **The debrief at the end of the survey sounds highly generic, and almost somewhat misleading: "The experiments in which you participated today were designed to examine how personal and environmental factors may affect human cognition and decision making. In psychology, it has been known that information can affect person's behavior to certain extent and that individual differences affect behavior. The purpose of the study was to know how exposure to stimuli and certain individual differences affect decision making and behavior." Is this the intended debrief, or has there been an error in copying from a previous study?**

Thank you, we updated the debrief to be more specific to the current study.

> "The experiments in which you participated today were designed to better understand the ways that people perceive, respond to, and evaluate, the risks and benefits associated with various activities and technologies in their environment. This type of study using numerical responses to a number of related questions is called psychometric analysis. Analysis of the responses to these scales may help us better understand differences in the way people perceive the risks of an activity or technology in relation to its benefits."

> **In the formatting of the "Common vs. dread" scale, one of the letters in the third word is missing the emphasis.**

Thank you for catching this! We updated the Qualtrics accordingly.

> **The sample size is set to be suited for detecting one-tailed effects of d = 0.19. Although not directly comparable, I think it would be good to compare this to the effect size in the replication target, and later studies using similar approaches. Given that the study's result is described as non-intuitive and not robustly demonstrated, it may make more sense to test for two-tailed effects.**

The typical approach for a replication of this type is indeed to compare it to the effect size in the replication target and, as you suggest, later studies that use similar approaches. The difficulty here is that for various reasons, we do not believe that the original's effect size can serve as a basis for comparison. These reasons include: the adjustments made to the original's design, the original's focus on item level analyses, and our data analysis plan to focus on the individual level. We agree that it makes more sense to test for two-tailed effects and have amended the Power and sensitivity analysis section to reflect this change.

> **Overall, I think the planned study is sufficiently close to the replication target, and the deviations are clearly recognized.**
> **Perhaps it would make more sense to report "Classification of the replication" (Table 4) separately for comparing to Fischhoff et al. (1978) and for comparing to Fox-Glassman and Weber (2016). The design appears to be quite similar to the latter, but with more differences to the former. In any case, the classifications of "Different/Same" is confusing – I'm guessing this refers to comparing the current study to either of the two previous studies.**

Thank you for pointing this out.

We updated Table 4 to remove reference to Fox-Glassman and Weber (2016) and added an additional table, Table 5, to the manuscript to address the classification of the replication with respect to Fox-Glassman and Weber (2016).

> **As far as I can see, the study design and analysis plan are sufficiently clear for the central confirmatory hypothesis related to RQ1. But as I have argued above, there are additional research questions that the researchers plan to measure and explore. Here the study design seems quite rigid, but with an exploratory aim, by their nature these analyses have a lot of flexibility.**
> **The authors could consider recoding the "benefit" and "risk" variables, as well as the activity names and the different risk scales in order to allow a masked analysis. However, the benefits of this may be limited as long as they retain only the expectation of a negative association between risk and benefit in RQ1**.

Apologies, we are not sure what is meant by a "masked analysis" and an example or a citation may have been useful here. Yet, you pointed that this does not seem to be crucial, and the project is already very complex. We therefore proceed in keeping our primary focus on the perceived risk/benefit relationship.

The design and analysis plan appears to be sufficient to test the single stated hypothesis. No obvious candidates for positive controls or parallel measurements comes to mind that would not cause significant deviation from the replication effort.
As argued above, I would recommend additional hypotheses and expectations for the results to be added, which would require additional specification in analysis approach.

Thank you again for the interest in the additional research questions. As noted above, given the scope and complexity of the project, we are limiting our focus to the perceived risk/benefit relationship.

Minor details in manuscript to consider:

Thank you for each of these detailed corrections and suggestions. We made the amendments for each as noted below.

- **Page 17: Check for line breaks and bracket parentheses**

We fixed the page breaks and removed hanging brackets.

- **Page 21: Missing word inserted: "asks participants to RATE each of the 18 items on both perceived risk and perceived benefit"**

We added the missing word.

- **Page 22: "In Task 2 participants are instructed to judge how acceptable the risk level of each item currently is." – Perhaps it could be clarified whether this mean the way that we as a society currently relate to the risks of this technology?**

Thank you for the suggestion. For purposes of the replication, we would like to retain the instructions as close to the original to the extent it can be.

- **I'm confused by Table 2 of the supplementary materials (document page 48), stating the findings in the original article(s). The crucial finding to replicate (a negative association between risk and benefit) does not stand out clearly in this presentation. Perhaps it is the non-significant p-values that confounds the message, and a different reporting standard should be used?**

Thank you for pointing this out. Yes, the original study's findings were non-significant, which is understandable given the low power, yet the effect found is rather typical in social psychology (r = -0.2). We feel like it is important to keep all the statistics as is in that table. We reiterate again

and again throughout the manuscript the problematic design and test of this core hypothesis and our main point in doing this project is to offer avenues to address that confusion.

- **Table 4 of supplementary materials – it would be good to retain same numbering for each named activity in both lists for easier comparison.**

We updated the numbering in Table 4 to match the original items and split out the new and/or modified items separately as well as presented the list as it appears in the survey.