Peer review for PCI RR of "Revisiting and updating the risk-benefits link: Replication of Fischhoff et al. (1978) with extensions examining pandemic related factors", by Jason M. Frank & Gilad Feldman

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).

Scope

The aim of the currently suggested registered report replication is to replicate one of the key findings in Fischhoff et al. (1978). The study has previously been subject to a fairly close replication by Fox-Glassman and Weber (2016), as discussed in the manuscript. The current study collects most of the same variables as in the replication target study. The data collection exceeds what is needed to test for the single stated hypothesis (discussed below). The replication also adds questions about the current COVID-19 pandemic. There are a number of methodological and analytical deviations from the replication target, made for practical reasons as well as to improve the study design.

1A: Research question

Replication of a descriptive study

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.

Stated research question

The stated research question of the current RR is to directly replicate and extend the psychometric risk/benefit assessment paradigm of Fischhoff et al. (1978). The current authors state three research questions, pertaining to (RQ1) the relationship between perceived risk and perceived benefit, (RQ2) the relationship risk and benefit has with the acceptability of the activities, and (RQ3) the relationship between risk characteristics, perceived risk, and perceived benefit?

I think these are valid and relevant research questions for the current project, and it should be possible to contribute to answering them within the project as described. I see no ethical issues for running the project.

COVID-19 research aims

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?

1B: Hypotheses

Missing hypotheses?

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.

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.

Other findings from target article to replicate?

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.

COVID-19 extensions

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.

1C: Methods and analysis plan:

Selection of items

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.

Task 1 response mode

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.

Addition of Task 1c

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)?

Shortened Task 3

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.

Details in Qualtrics survey to consider:

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)?

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?

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?

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?

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

Statistical power:

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.

1D: Closeness of replication and undisclosed flexibility

Closeness of replication

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.

Undisclosed flexibility

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.

1E: Ability to test the stated hypotheses

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.

Minor details in manuscript to consider:

- Page 17: Check for line breaks and bracket parentheses
- Page 21: Missing word inserted: "asks participants to RATE each of the 18 items on both perceived risk and perceived benefit"
- 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?
- 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?
- Table 4 of supplementary materials – it would be good to retain same numbering for each named activity in both lists for easier comparison