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!

-Katherine Fox-Glassman

1A. The scientific validity of the research question(s). This research question is scientifically justifiable, and based in theory and prior research.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable. 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!)

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

- I like the addition of the new COVID-related measures!

- Completely reasonable assumption that the original effect size is not suitable for calculating a power analysis.

- 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.)

- The measures for limiting the sample to US adults look likely to work fairly well. (Certainly no worse than our 2016 study, and likely better.)

- 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.

- 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!).

- 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.

- 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!)

- Evaluation criteria for replication findings also make sense.

- 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.)

- 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.

- 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?

- 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?

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.

- The clarity and detail here is great—clearly a lot of thought has gone into this plan. With the exception of my hesitance about the t-tests, I think the plans for attempting to compare the current and prior work are as good as they could reasonably be. (That final qualification is due to: expectations of changes over time; inescapable differences in the populations; seemingly minor changes in the study protocol inadvertently having larger effects.)

- 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!)

- 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.)

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.