Reply to PCIRR 2nd revise and resubmit decision letter: Tversky and Kahneman (1971) replication and extensions

We would like to thank the editor for the 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 comments are in bold with our reply underneath in normal script.

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

A track-changes manuscript is provided with the file: "PCIRR-RNR2-Tversky-Kahneman-1971-replication-extension-mainmanuscript-track-changes.docx" (https://osf.io/b4sju)

Reply to Editor: Dr./Prof. Moin Syed

> **Thank you for submitting your revised Stage 1 manuscript, "Revisiting the "Belief in the law of small numbers": Conceptual replication and extensions Registered Report of problems reviewed in Tversky and Kahneman (1971)," to PCI RR.**
>
> **As indicated in the previous decision letter, I reviewed the revised manuscript myself rather than returning it to the reviewers for further comment. I appreciate the careful and thorough approach you took to the revision, and believe that the current version is much improved and will be ready for a recommendation after some minor revisions.**
>
> **Your study materials are such that they could be tinkered with endlessly to attempt to improve them. What you have crafted, following the helpful reviewer comments, is a solid set of questions that is certain to be informative and to stimulate further work. Thus, I am happy with the question wording and have only suggestions that are mostly cosmetic and aimed at clarifications.**

Thank you for the positive constructive feedback, and the invitation to revise and resubmit.

> **.1. I understand your concern about the challenges of having many robustness analyses. However, in this case, it is not "many" but only one (per analysis), namely the removal of outliers. Moreover, I question the use of the analysis only if you fail to find support for your hypothesis, as this ignores the possibility that you found support because of the included outliers.**
>
> **In terms of integrating findings, if you find support in one analysis and lack of support in another, that provides some indication that the finding is "fragile" in some way, either because it is a weak signal or because it is highly dependent on specific conditions. This is a more informative result than a single test without a robustness check. Unless you can offer a strong and convincing argument to the contrary, I suggest you follow the procedure I describe here.**

In our previous round Dr./Prof. Romain Espinosa suggested winsorizing instead of exclusions, which we adopted. In that discussion we referred to the challenge of trying to understand what

outliers represent, whether they represent 1) thoughtful reasonable responses, or 2) inattentiveness, lacking comprehension, a deliberate biasing agenda, or disregard for the survey.

We noted our choice for winsorizing was because we felt that participants in this target sample are attentive and therefore we regard outliers to be thoughtful and reasonable responses. Given that we severely restrict the range of answers (for example, likelihood, the most common answer type in our survey, can only be 0-100), outliers are less of a risk of giving unreasonable answers (for example, in an open ended question asking how much one would pay for a coffee, someone would indicate a million), but rather that they are likely to be in the range of answers that make sense. We therefore first treat any answers as if they represent real thoughtful answers, and therefore consider statistically manipulating these answers to either exclude them or standardize them represents a shift from what took place in the sample. We therefore need to consider that when we choose to winsorize, we are changing the meaning of what some participants indicated, for the mere fact that they gave different answers from what others have answered.

In addition, adding more information with different analyses and transformations may unnecessarily confuse readers, and make it more difficult to conclude what the findings are in cases where the different analyses show different conclusions. This is why we aimed to try and assume the data is what it is, and only proceed to conduct additional analyses if we fail to find support for the hypotheses, with needed compensation for alpha.

If we find support for the hypotheses with outliers included, then we think that it is reasonable given the assumption that those represent real data points. If the effect does not hold when we remove those outliers or winsorize them, one might similarly argue that we were artificially forcing the sample to be something it is not and failing to capture an effect that would normally be found in the general population that includes those members.

Though the number of robustness checks is a concern, the bigger concern is that each of those "increase[s] complexity and decrease[s] the interpretability of findings when robustness check findings do not align" (from our previous reply). The project is already quite complex, with many problems, dependent variables, and tests.

In addition, please consider that there could be countless ways to run robustness checks, which leaves every researcher with the question: which robustness checks? how many robustness checks? what do each of those represent? how to interpret those against each other?. For outlier handling alone there are many possible methods for deciding who to exclude and why, with much flexibility for researchers, and no real consensus on which of the methods is right in what context (Leys et al., 2019). "Outliers detection" from Easystats R packages specifies 15 different methods for outlier detection. Recent Twitter discussions indicate the recommendations for (statistician Isabella R. Ghement): "1. Carefully check outcome values which seem "extreme" but keep them in the analysis if they are valid (e.g., not a mistake or aberration); 2. Perform the

analysis using ALL observations;". Given enough flexibility one can find or fail to find anything. Similar recent threads show the complexity of outlier detection and exclusions (e.g., Nick Holmes: "don't just take my word for it, here's a true god of reaction times, Jeff Miller, who said, 30 years ago, that outlier removal is: "very dangerous""). What fragility or failing robustness checks might mean when we windorize/exclude outliers based on some outlier criteria is uncertain, if we are not sure whether that is needed and what we are doing when. This is why we consider it much better to ensure high quality responding and range restrictions, and focus on the much clearer, inflexible, and conservative full sample.

We ask for your understanding and support of this strategy.

However, we understand the concern and therefore added the following to the "Data analysis strategy":

> **Robustness checks: Outlier handling**
>
> Following feedback from peer review, we will conduct an initial outliers analysis on Problems Q1, Q2 and Q3 (replication DV1) to examine the differences regarding the conclusions from analyzing the full data and from analyzing the data with winsorizing. In case the conclusions for the full sample are in support of the findings and the winsorized sample is not in support of the findings, we will conduct exploratory analyses to examine possible explanations for how the participants whose responses were winsorized differ from the larger sample (attentiveness, demographics, etc.), and will conduct robustness checks with full versus winsorized comparisons for all effects reported in the manuscript. Regardless, our conclusions regarding whether we found support for the effect will depend on the findings using the full sample, yet we will add a section in the general discussion to discuss limitations and implications. Our reasoning for this choice is because we restricted the range of response for all items which makes all responses reasonable, and our experience with the target sample is that it is of high quality responding and attentiveness, and therefore outliers are likely to represent real and thoughtful responses.

In the results section:

> Robustness checks
>
> [See data analysis strategy "Robustness checks", to be updated in Stage 2.]

To the General Discussion:

> [Planned discussion regarding robustness checks, implications, and limitations.]

.2. Please elaborate on the "law of large numbers" in the Introduction (p. 13). It is introduced in passing at that point, and mentioned again later in the Introduction section, but the clearest statement about its importance for the study comes in the Method section (p. 36) where the potential competing hypothesis is discussed. Given that this appears to be a major motivation for the extension, a clearer and more consolidated treatment in the Introduction is needed.

We first would like to note that the law of large numbers is not needed in order to motivate this replication or our extensions. We felt it necessary to conduct a stronger causal test to go beyond Tversky and Kahneman's single case example demonstrations with occasional descriptive statistics to directly manipulate sample size in order to test their assumptions. Given that their hypothesis is a null hypothesis, which is tricky with Null Hypothesis Significance Testing, we reframed that to an alternative hypothesis, and mentioned that there has been some debate which of these hypotheses hold and under what circumstances. It was mentioned in passing, because it is not the core of the manuscript, and the empirical evidence in support of either hypothesis is surprisingly very limited.

We however agree that a discussion of this debate is relevant and would be helpful after observing the results, especially for our extension, and therefore added the following to the discussion section:

> [Planned discussion: Implications for the results for the debate regarding "belief in the law of small numbers" versus the "empirical law of large numbers" (Sedlmeier & Gigerenzer, 1997), especially in the context of the causal test in our extension]

In page 13 in the introduction it is mentioned briefly, yet the longer elaboration on this point was added in the section about our extension.

We therefore first added a brief link in the earlier parts of the introduction (page 13) to the further discussion later in the introduction (addition marked in bold):

> This holds especially relevant given controversies in the literature in a now classic debate with a different group of scholars maintaining that people actually hold a "belief in the empirical law of large numbers" (Sedlmeier & Gigerenzer, 1997; <u>see discussion below in our extension section</u>), arguing for the need for reframing these effects, identifying possible moderating factors such as study design (single/between versus within) and presentation format.

And, we adjusted our discussion of the extension introduction to the following:

The alternative hypothesis is that people do take sample size into account. One theoretical account that does not assume a null-hypothesis is the belief in the "law of large numbers" which predicts that the larger the sample, the more likely people are to perceive it as representative of the population (Sedlmeier & Gigerenzer, 1997). Previous research demonstrated that people can intuitively infer that the larger the sample size the more likely it is to resemble the characteristics of the population, with youngsters as early as the age of 11 showing indications of having this intuition (Piaget & Inhelder, 1951/1975) and initial similar demonstrations in adults (Peterson & Beach, 1967).

To try and resolve the two seemingly opposite findings, Sedlmeier and Gigerenzer (1997) suggested that it might be dependent on the type of question, such as a distinction between frequency distribution ("a distribution of values from one sample", p. 36) and sampling distribution ("distribution of means from independent samples of fixed size, drawn from the same population"). In so doing, it seems that Sedlmeier and Gigerenzer (1997) accepted the premise that the results of the target article hold and generalize, and that their resolution relies on the statistical presentation of the problem. Yet, it is possible that the phenomenon is dependent on the experimental design, and that a direct manipulation of the sample size in these scenarios would allow us a stricter empirical causal test to contrast "belief in the law of small numbers" suggesting null differences against the possibility of "belief in the law of large numbers" suggesting an adaptation to evaluations based on sample size. In the case that people do adjust their evaluations according to sample size, a manipulation would allow us an initial estimation of the direction and the extent to which people make such adaptation. The implementation of that manipulation across the different scenarios would allow us to examine whether such adjustments differ based on the question at hand. Together, these may suggest a more nuanced view of the debate regarding the seemingly contradictory findings.

.3. The subsection of the Introduction labeled "Exploratory directions" includes only that it will be updated in Stage 2. This should be removed completely. Although I am personally not super strict about modifications to the Introduction following IPA, some people are, and it is best to not intentionally introduce such into the process. If you know what the exploratory directions might look like and why they are useful, then say so, otherwise leave it be and introduce them as exploratory when doing the analysis.

Understood. We removed the section.

> **.4. Similarly, the "Exploratory Extensions" subsection of the Method is too vague, with reference to "several dependent variables" being included. Here, given that it is the Method section and you plan to use these variables, you should include additional detail rather than remove them. The "exploratory analyses" section should also give some idea of what you plan to examine.**

Given that these were marked exploratory, and to some extent depend on the results of the replication, we felt it better not to specify too much in advance in the previous submission.

That said, we appreciate the nudge to do better and specify as much of these in advance, and have worked to add needed content and revise our Rmarkdown analyses of the extensions. The new Rmarkdown is much improved and includes the planned extension analyses.

We added the following to the Method section under "Exploratory extensions->Statistics intuitions dependent variables":

> We added several dependent variables to some of the Problems to further explore the scope of the phenomenon and examine possible biases in statistics related lay-intuitions. We summarized the extensions for the statistics intuitions in Table 4.
>
> [extensive Table 4 added here]
>
> **Replication success likelihood**
>
> The target article's Q1 "Replication success" was aimed at assessing the way people estimate the likelihood of a successful replication. We aimed to further explore those estimations in the other problems and therefore added replication likelihood questions to: 1) Q3: "Infants" (DV2); 2) Q5: "Exploratory analyses" (DV1 and DV2).
>
> The target article's Problem Q5 focused on whether the original study should be rerun and with what sample size. We aimed to supplement that with examining the perceived likelihood that a replication would show the same results, and whether it would make any difference whether the replication was conducted by the same experimenters or by an independent lab. Therefore, we will conduct a mixed ANOVA (within:same scholar vs. different scholar; between: sample size manipulation) examining laypersons' evaluations of independent verifications, that results are more likely to be replicated if conducted by the same scientists than by other scholars. This direction is exploratory, and framed as assuming differences, yet we suspect that people might underestimate possible scientist biases and perceive the two as fairly similar (null hypothesis, effect size Cohen's $d < 0.2$).
>
> **Estimations of required sample size**

We aimed to further explore laypersons' intuitive power analyses to assess their estimations regarding required sample size to test empirical questions. We added questions to: 1) Q3: "Infants" (DV3), 2) Q5: "Exploratory analyses" (DV4), 3) Q6: "Failed replication" (DV2).

Likelihood of effects

The original Problem Q4: "Population correlation" focused on people's expectations regarding effects similar to that of the population given an estimated required sample size. We wanted to assess not only the likelihood of finding support for an effect but more specific expectations regarding found effects. We included both specific effects, one the same as the population (0.35), one higher than that of the population (0.40), and one lower than the population (0.30), as well as ranges - equal or higher than population (0.35), equal or lower than population (0.35).

This direction is exploratory, yet we suspect that people might tend to overestimate the likelihood of very unlikely exact point estimates, especially the one closest to the population (0.35) and underestimate the likelihood of the much more likely range estimates, especially that with weaker effects than that estimated in the population (equal or lower than 0.35).

We also added examples of the analyses from the Rmarkdown in the results section.

.5. I find it confusing to have the scholar versions included in Table 3, given those versions are not part of the study. Table 3 should only include the original and the lay versions, and the current Table 3 with the scholar versions can be added to supplemental if you think people would be interested in seeing them for future work.

We understand. We thought it important for the reader to be able to see the steps in the translation, yet we accept that this may be confusing and/or overwhelming. We therefore made a copy of the Table 3 with the scholars' version in the supplementary, and adjusted Table 3 in the main text as requested, with a note referring to the full table in the supplementary.

6. The role of the Bayes Factors is not clearly specified. Specifically, how will inferences be made if the Bayesian and NHST results diverge? What is the rationale for using a default Cauchy prior? That BFs are reported by default in the figures is not a sufficient reason to include them, rather they should be fully integrated if they are to be useful.

We appreciate the suggestions and the opportunity to improve clarity and add details.

We reframed that section, and adjusted to the following:

> Our main analyses and empirical focus is on using Null Hypothesis Significance Testing, and this is what we use in order to test for detecting a signal (rejecting the null hypothesis). The conclusions of whether a hypothesis was supported or not will rely on NHST. Yet, in cases in which we fail to reject the null, we will complement all our analyses with Bayesian analyses aimed to try and quantify the evidence in support of the null.
>
> Bayesian analyses are tricky specifically because they incorporate a subjective measure of a prior, which is especially challenging given the competing hypotheses, a replication of an old classic that has not been subjected to many replications, very little in the follow-up literature, and no experience for this research question with our target sample. Therefore, any prior is debatable, and so we implemented a prior of 0.707 that is often used as the default in many Bayesian tools and packages (such as JASP, BayesFactor, and ggstatsplot), generally meant to address ambiguous cases like this. Bayes factor will be reported in our figures, using the R package ggstatsplot.

Regarding the default prior, here is from Schmalz et al. (2021):

> “Software programs such as JASP have a default prior. The default prior for the alternative hypothesis of a t test is a Cauchy distribution (see Figure 3), which is centered around zero and has a width parameter of x = .707. [...] The width parameter of the Cauchy distribution (x) corresponds to the bounds of the range of effect sizes which are proposed (by the researcher) to occur with a 50% probability. [...] The default prior, with the width parameter x = .707, therefore, makes the claim: “We are 50% confident that the effect size lies somewhere between d = .707 and d = .707.”

Given that we do not know what effects to expect, we find that to be a reasonable prior.

7. You alternatively refer to the same process as "consent checks" and "attention checks." I would stick to the former, as those are more descriptive of what they actually are.

Thank you. We added a bit more information to clarify why the consent checks also serve as initial attention checks. We only refer to attention checks in the context of explaining the consent checks:

> "Three of the four questions had the options order being rotated (yes, no, not sure), thereby also serving as attention checks by ensuring that participants carefully read the question options. Participants who failed to indicate their consent and answer the rotated consent questions were asked to return the task."

8. p. 41 – It states that "eight of the measures are replications" but I believe this should be seven.

We think we identified the cause for the misunderstanding and have worked to revise accordingly. We summarized the replications and extensions in Table 5, and eight of the rows were marked as "replication", yet they appear in seven problems that we aimed to replicate.

We changed the paragraph to read:

> Eight of the measures in the seven problems are replication dependent variables, taken from the original study with a translation to laypersons aiming to demonstrate the generalizability of the underlying phenomenon in a sample of laypersons. We will therefore compare our findings to that reported in the target's (for those that reported sufficient details).

9. p. 41 – the first sentence after indicating the alpha threshold of .001 is confusing, referring to "up to six additional dependent variables, sever overall." I was not clear what that meant.

We tried to make that clearer in our revision:

> We added extension dependent variables to Problems Q3, Q4, Q5, Q6, and Q7, summarized in Table 4, with up to 7 analyzed variables per each of the problems (when combining the replication with the extension dependent variables). Therefore, .001 meets the strict Bonferroni .01/7 suggestion and a round clear number.

> **Once again, I will review the revised version myself. I will attempt to so immediately upon submission, and assuming that you are attentive to the issues outlined above, I imagine I will be able to recommend an in-principle acceptance at that time.**

We are very grateful for the comprehensive feedback, and would very much appreciate a speedy approval to allow us to proceed to data collection and meet our strict deadlines.