Reply to PCIRR decision letter #410: Tversky and Kahneman (1971) conceptual replication and extensions

[Important note: We are very grateful for the immensely constructive and positive feedback from the reviewers, which has helped us catch and address oversights. Yet, that also meant our revision round took longer than we expected.

We therefore feel it necessary to note that this submission is part of a MSc thesis project with the thesis submission date currently set to the end of June. We therefore note that we will have to proceed to pre-registration and data collection by June 21st the latest, regardless of whether we receive an in-principle acceptance from PCIRR or not, in order to ensure timely thesis submission. We do hope to be able to proceed to data collection with the community's endorsement, however we want to align expectations in case that is not possible.

In case we do not receive the community's in-principle acceptance in time and we proceed to pre-register and collect the data, then based on our previous discussion and correspondence with recommender Chris Chambers, this would mean an adjustment of the PCIRR control level towards "RRs involving existing data" from Level 6 to a lower level (per "Guide for authors" on the PCIRR website).]

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 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/YcuXJLAvOjDt

A track-changes manuscript is provided with the file: "PCIRR-RNR-Tversky-Kahneman-1971-replication-extension-mainmanuscript-trackchanges.docx" (https://osf.io/xvu8t)

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: Clarified the reason for exclusion of Q7 from the original paper and the reason for classifying the current study as a conceptual replication. R1: Revised languages on some questions to make them easier to understand, also updated the exclusion criteria of participants. R2: Added the original stimuli to Table 3 according to reviewer's recommendations. Also updated our adjusted stimuli for easier understanding for laypersons and to avoid confusion.
Introduction	Ed: added the Nosek and Errington (2020)'s perspective of conceptual replication on top of the original LeBel et al.(2018)'s view. Also explained why the theme of Q7 was repeating that of Q5 and Q6. R1: Rephrasing certain sentences to make them more generalizable to the population.
Methods	R1: revised the question scenarios and prompts to make them more comprehensible for layperson, also added the type of responses for each question in Table 3. Updated the exclusion criteria of participants to reduce misunderstanding. R2: major and minor changes to Table 1 and Table 3, which include rewriting the hypotheses to minor updates to the questions regarding language use.
Discussion	Many added points raised by reviewers added as planned discussion after data collection.

Note. Ed = Editor, R1/R2 = Reviewer 1/2

[We note that we are not familiar with the titles and ranks of the reviewers, and looking for that information proves tricky. To try and err on the side of caution, we refer to all reviewers with the rank Dr./Prof. . We apologize for any possible misalignments and are happy to amend that in future correspondence.]

Reply to Editor: Dr./Prof. Moin Syed

> **The reviewers and I were all in agreement that you are pursuing an important project, but that the Stage 1 manuscript would benefit from some revisions. Accordingly, I am asking that you revise and resubmit your Stage 1 proposal for further evaluation. I do not expect to return the revised version back to the reviewers, but will act on the manuscript myself.**

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

> **The reviewers provided thoughtful, detailed comments with align with my own read of the proposal, so I urge you to pay close attention to them as you prepare your revision. A few points that require special attention:**
>
> **1. The reviewers and I all had questions about your treatment of outliers, exclusions, and multiple hypothesis testing. Rather than having results-dependent approaches to these issues, you should treat the decisions as constituting a set of a priori robustness analyses. See reviewer comments for specific issues. I also question your decision to only include data from participants who completed the entire study, as there is an extensive literature on missing data highlighting how listwise deletion can often result in the largest bias.**

We received some great constructive feedback, for which we are very grateful, and we did our best to answer your and the reviewers' comments in detail below.

Per the question regarding inclusion of data of participants who have not completed the study, there are many reasons why not to include those who dropped out. For one, their compensation relies on a completion code that is provided at the end of the survey, therefore those who do not complete the study have likely not been compensated for it, or have not gone through the procedures of funneling, demographics, and debriefing. We provide the option of dropping out and providing a dropout code to be compensated in the consent form, yet from our experience this has rarely been used. In addition, our experience with this target sample is that those who tend to drop out do so mainly because they have been distracted while taking the task or that the task has not met their expectations in some way, and both tend to have severe implications for these participants' data quality. This is not a simple case of missing values because of random inattentiveness or participants electing to skip some items (sidenote: we force responses on all items), this is about participants dropping out completely midway and returning the task to the platform indicating that they do not wish to be part of it. The list goes on and on. In short, our experience with many studies conducted with this target sample and this platform - some with

IPA or Stage 2 endorsement from PCIRR - is that to ensure high data quality it is best we only include those who completed the study.

The issue we see and have experienced in other peer review of replications regarding adding many robustness checks in NHST is that they increase complexity and decrease the interpretability of findings when robustness check findings do not align. Though we would hope that as a community we would have strict and clear practices as to how to handle such cases, we have come to learn the hard way that this is not the case, and that often parties tend to interpret the very same conflicting findings in very different ways. As mentioned by the reviewers below, these also tend to increase the perception of capitalizing on chance. We aimed to keep things as simple and straightforward as possible, with the strategy of clearly defining everything before data collection. Our understanding is that result dependent analyses are valid, if specified in advance before data collection, such as in the case of normality testing or interaction probing. That strategy has served us well in the other PCIRR endorsed replications.

We are open to the possibility that we are wrong, and would gladly revise this further if given clear editorial directive and guidelines for how you would like this implemented.

> **2. The Introduction section would benefit from some additional text about how you are conceptualizing the replication in relation to the target. That is, you frame the study as a conceptual replication but do not provide many details about why it constitutes such and the implications of the deviations. I know that the direct/conceptual distinction is widely used and accepted, but I tend to favor the Nosek & Errington (2020) perspective that shifts attention from the procedure to the claim, and thus does away with the distinction.**
>
> **After reading over the Tversky and Kahneman paper, they are quite loose with their claims, sometimes constraining them to psychologists/researchers whereas other times the claims see to be applied to all people. If you take the former claim, then yours is a test of generalizability, whereas if you take the latter claims yours is a test of replication. I don't raise this issue to force you to think about it as I do, but to highlight how and why it would be beneficial to clarify the nature of the replication.**

Thank you for this comment and suggestion. First we will clarify what we did and why, and then address the request, opting for mentioning both paradigms, the one we used, and the one you suggested. We also see the value of discussing this point in the introduction rather than in the methods section, and have moved the table to the introduction under a dedicated subsection of "Replication closeness evaluation".

We tried to be very careful with how we conceptualized this replication, and provided much details regarding the classification using the most commonly used replication criteria by LeBel et al. (2018) in Table 5 (now Table 1, given the move to the introduction). Using their paradigm, which we included in the supplementary for reference, there are fairly simple rules of thumb that help clarify how to classify a replication. We included these details and classification in the methods section, because much of the LeBel et al. paradigm is related to methods related issues.

We are familiar with the arguments made by Nosek and Errington (2020) and appreciate the view that we may benefit from shifting from evaluations of procedure to evaluations of the claim. We are also familiar and appreciate your advocating for this approach in Metascience 2023 conference (https://osf.io/6xsmc/), and for the most part this approach makes sense and can help address some of the issues we have faced in the past in our replications.

That said, this suggestion raises a host of issues and debates and an almost philosophical debate which we would rather not go into, and which deserves a separate in-depth discussion. It also requires clear falsifiable theories and clear direct links to testable predictions, which is not the case we are facing here and with many of our replications of the JDM classics which were mostly empirical demonstrations of a specific phenomenon, only later followed with phenomenon-based theories interpreting and making sense of a body of evidence. We have been tracking replications for long, and have yet to see a well-executed use of this paradigm in social psychology and/or JDM, that has succeeded in addressing the challenges the paradigm aims to overcome. If you feel this is needed, then we would very much appreciate citations of replications that have successfully implemented this paradigm in our research domains.

We thought that the point you made about calibrated claims important, and in our revision we tried to make our claims more specific and better calibrated, as for example, we modified Table 2 to change the “Generalized hypothesis” column to specifically mention laypersons. We also added exploratory questions aimed to tap statistics knowledge in our population, so that we can make better calibrated claims in our Stage 2 discussion.

Lastly, we wish to note that despite the understandable criticism we found the Lebel et al. (2018) replication classification paradigm to be an extremely helpful tool in communicating with reviewers and broad audiences. We have previously successfully employed the LeBel et al. criteria in many of our team’s replications and it served as a helpful language to communicate with others about replications. For the most part the classification has helped clarify core aspects of the replications, note deviations from the target, and align expectations regarding its interpretation.

The new section in the introduction now reads:

> **Replication closeness evaluation**
>
> We provided details on the classification of the replications using the criteria by LeBel et al., (2018) criteria in Table 1 (see section "replication closeness evaluation" in the supplementary materials about the classification). We were strongly grounded in the target article's claims and empirical demonstration, yet made major changes to allow for a test of its generalizability to laypersons. We summarized the changes we made in Table 5, and decided to classify this as a "far" conceptual replication based on the criteria by LeBel et al. (2018) given our many adjustments to the stimuli from expert language to targeting laypersons', and the shift in target population.
>
> Another system of categorizing replications was by Nosek and Errington (2020), which puts more emphasis on the match between claims of the replication and the target rather than the match between the replication and the target's procedures and methods. To be classified as a replication, outcomes from the study that are consistent with the prior claim increase the confidence in the claim, and outcomes that are inconsistent with the prior claim decrease the confidence in the claim. The target article by Tversky and Kahneman (1971) was somewhat vague about the claims, as for example they sometimes referred to scholars and used methods that require some background with statistics and methods, and sometimes the claims seem to be broad and to also make an argument about a wider effect that also holds for the general population.
>
> We see value in both approaches to replications. We use the paradigm by LeBel et al. (2018) to document our deviations regarding process and methods, and at the same time are building on the Nosek and Errington (2020) paradigm in our aim of testing the generalizability of the claims made in the target article to a broader population.

> **3. You indicate that Q7 was omitted because it was a repeated theme of Q5 and Q6, but you do not actually include what the question was and in what way it was a repeat. Additionally, based on the argument, one might wonder why both Q5 and Q6 are included—if Q7 is a repeat of both, is Q6 not a repeat of Q5? These questions can be briefly addressed by including the question and explaining how the theme is repeated.**

Good point and we appreciate the feedback. In hindsight, indicating this as a "repeat" was not accurate. Rather, this was a combined issue of statistical jargon and knowhow referring to required "t value" which is difficult to translate for laypersons, and of some overlap in the theme with the extension questions we added to the Q5-Q6 combo.

We therefore added the following to the methods section:

> We note we did not include Q7 as it covered a similar theme to the questions we added to Q5 and Q6 and referred to "value of t" that required statistical knowhow and was difficult to translate to laypersons. Specifically, the target's Q7 was as follows
>
> > "An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, t = 2.46. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data.
> >
> > You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate?"
>
> Problem Q7 therefore aimed to get at one's perception of the statistical threshold for "t" that needs to be met to conclude a failure to replicate. In both Problem Q5 and Problem Q6 we added questions on top of those presented in the target which aim at ideas related to those from Problem Q7, asking about confidence in the findings, required sample size, etc.

Reply to Reviewer #1: Dr./Prof. Romain Espinosa

> **Overall, I think that this is great work and will make a nice Registered Report. The authors spent a lot of time explaining the theory, their objective, and what they plan to do. I have only a few suggestions that I detail below. The more important concerns are: (i) outlier exclusion, and (ii) multiple-hypothesis testing.**

Thank you very much for your time and effort in reviewing our proposal. Your constructive comments have helped considerably improve our submission, and for that we are grateful.

> **My main concern is about multiple-hypothesis testing (MHT). The authors mention it a bit at page 36 when they say that they will rerun the analyses with a stricter alpha if they fail to support the core hypothesis. First, if they reduce alpha, their probability to reject H0 will be even lower, so I don't see how this could change their findings.**

In that specific section, the point was about outliers and exclusions. We were aiming to reduce multiple "dipping" into the dataset using the same alpha threshold, when rerunning the analyses with exclusions. Some claim that excluding outliers and using stricter quality qualifiers in the dataset leads to less noise and therefore - if there indeed is an observable effect- to higher accuracy likelihood of finding that effect after exclusions. However, if we were to simply rerun the analyses without compensating for alpha, our understanding is that this would increase the random chance of detecting a signal for rejecting the null even when there is none. The method of stricter alpha (of .005) for that case has been previously recommended to us in such cases and was therefore applied here. Therefore, it aims to balance between rerunning with exclusions supposedly increasing the value of the data with compensation for the risk of capitalizing on chance.

We explain further on this point and how we addressed it below. We are happy to change and improve on this if given clear editorial guidelines on how to better approach exclusions and outliers.

> **Second, I do not think that reducing alpha arbitrarily is a good way to account for MHT. In general, I think that the authors should present uncorrected and corrected p-values all along the way.**

We agree that presenting corrected p-values alone is an issue. We were planning to report p-values regardless of alpha criterion next to the effect sizes and confidence intervals. In our submission we made no reference to p-value corrections, but rather only to a change in the alpha threshold, yet we would be happy to accommodate reporting both corrected and uncorrected

p-values in our extension ANOVA analyses on the manipulated sample size. We added a column of uncorrected p-values to Table 11.

> **As their Study Design table demonstrates, all hypotheses aim to test the same question, i.e., whether there is a bias. So, all hypotheses are from the same family. I would strongly encourage the authors to present p-values corrected for a Family-wise error rate of 5% (or something in this direction). I suggest using Romano-Wolf correction as the pvalues might correlate (they test the same concept underlying the data). (We discuss it a bit here: https://link.springer.com/epdf/10.1007/s40881-022-00123-1?sharing_token=VtVkhrhefQXtJt6VmuO1Lfe4RwlQNchNByi7wbcMAY4_IUtgns7HNovQ1hN1urYA9vXJe08o7XheK7cOrisvQ0ok5l2bU99xK7tmGkTrWh3lXyv9jnetbS-DK-DXeNTvuRdRmz_PSjG6U4Ohm_TGiMJgi6mxB8Ptdr0CM0yBVcU=)**

Yes, this is a tricky issue with NHST and p-values, and there are diverging views regarding how to handle these. This is also partly why we are reporting Bayes Factors for the problems (which also addresses Reviewer 2's request).

As for the p-values and adjustments to multiple comparisons, rather than making corrections and adjustments to p-value, we opted for reporting the raw test p-values and only adjusting the target alpha we use to classify p-value as a signal.

In setting the alpha, we differentiate between tests on the *replication* dependent variables and tests on the *extension* dependent variables. Given that there are seven problems, and the typical alpha threshold is .05, we set our target alpha at .01, slightly higher than the conservative Bonferroni .05/7 suggestion and a round clear target.

For the extensions, given multiple dependent variables in some of the problems, with some problems having 7 dependent variables, we set a stricter alpha target of .001 for all analyses on extensions dependent variables, around the Bonferroni .01/7 suggestion and a round clear number. We are comfortable with the NHST Type 1 and Type 2 error trade-offs for our exploratory extensions.

We updated the “Data analysis strategy” subsection of the Methods section in the main manuscript accordingly:

> **Replication**
>
> Eight of the measures are replications, 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).
>
> The replication tests will be conducted on the control condition. We decided to set the alpha to .01, as .05 is a common threshold used in replications to evaluate a signal in support of the target’s findings, and there are seven problems, with one main replication dependent variable each, therefore slightly higher than the conservative Bonferroni .05/7 suggestion and a round clear alpha target.
>
> **Extensions**
>
> We set our alpha threshold to .001 for all extension analysis.
>
> In some Problems we added several extension dependent variables, up to six additional dependent variables, seven overall. Therefore, .001 meets the strict Bonferroni .01/7 suggestion and a round clear number.
>
> For the extension manipulating the sample size (x versus 10x versus x100), we will conduct a series of one-way ANOVAs with post-hoc contrasts against the control condition for each of the measures. For the ANOVA analyses we will report Holm corrections for multiple analyses and will report both raw and corrected p-values, yet our criteria for signal will use the corrected p-values against the .001 alpha threshold.

> **Outliers: I am not a big fan of excluding the outliers as they might convey some information. Have you thought about winsorizing those data instead? (Note: if you exclude them, do you exclude them for all the data analysis or only on this question?)**

Thank you for this note and the suggestion.

We generally agree about exclusions, which is why in our main analyses we did not plan to do any exclusions, and only planned to run exclusions as supplementary analyses in case we fail to find support for the hypotheses. The way we perceive the debate on whether to exclude or winsorize is that it depends on the perceived issue with outlier values, whether we regard them as representing thoughtful reasonable responses and then winsorizing make more sense, or rather regarding them as representing inattentiveness, lacking comprehension, a deliberate biasing agenda, or disregard for the survey, for which exclusions would make sense.

Given our quality control measures and experience with the target sample, we are less concerned about data quality, and therefore winsorizing makes more sense. We appreciate the suggestion to replace outliers exclusions with windsoring, and have updated our outlier section in the method section from "outliers and exclusions" to "outliers and winsorizing", and updated our R code accordingly using datawizard (easystats) using the zscore method with a threshold of 3 standard deviations from mean. We note, though, that we will only run these analyses if we fail to find support for the target's hypotheses. This also answers the dilemma regarding whether exclusions are per variable or per survey, as windsoring is conducted per variable.

We changed the "outlier and exclusions" subsection in the "method" section to "Outliers and winsorizing" as follows:

> We pre-register that in case we fail to find support for the core hypotheses in our replication of the target article, we will then supplement our analyses with rerunning the analyses with outlier winsorizing. Following peer review and a suggestion from reviewer Prof. Romain Espinosa we will employ winsorizing using the R package datawizard (Patil et al., 2022) using the method zscore with a cutoff of 3 standard deviations plus or minus the mean. In such a case we will report findings of both before and after winsorizing, and document differences in the findings.

Minor comments:

Page 19: it is written "previous research demonstrated that people intuitively know that the larger a sample size is, the more likely to produce a uniform distribution".

—> This sentence is true because you consider here the population to be distributed according to a uniform distribution (in your next sentence). I would maybe suggest rephrasing it in a more general way (e.g., the empirical distribution tends to get closer to the population distribution when the sample size gets larger), no?

Thank you for the comment. Good catch, we agree.

Original:

> Previous research demonstrated that people intuitively know that the larger a sample size is, the more likely to produce a uniform distribution, even young children at the age of 11 can understand it (Piaget & Inhelder, 1951/1975).

Revision:

> 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). Which raises the question - If people have that intuition, then why do people not intuitively understand that a small sample is not representative of the population?

For Q1, I'd like to mention that the authors make two things vary (at the same time I assume?): the sample size for the exploratory study and the sample size for the confirmatory study. I'm fine with that because if the two sample sizes increase, the probability of replication should be larger. But it is not straightforward that both dimensions should jointly vary. (This holds for all other questions that involve two sample sizes.) [This relates to your paragraph on sample size manipulation.]

Yes, this is an important clarification to make. We therefore added the following to the "Extension: Sample size manipulation" subsection of the "method" section:

> We note that in the scenarios we manipulated all the mentioned sample sizes in the question. For example, in Problem Q1, we varied both the sample size of the described original experiment of 20 (to 200 and 2000), and the sample size of the described replication of 10 (to 100, and 1000). An alternative approach could have been to only

> vary the described original experiment or to only vary the described replication. We decided to vary both because we wanted to keep the ratio between the original and the replication constant, to be able to examine the overall use of sample size information.

In addition, we added a planned discussion section citing the above and noting that:

> "Planned discussion: Manipulating original and replication numbers separately. As we explained in the methods: "[...]", and a direction for future research could be to vary both parameters. We will discuss possible insights that can be gained from such an experiment."

> **For Q2: I'm wondering whether increasing the sample sizes is the best option. Why not try to decrease it as well (i.e., going below 50). The main point is that, T&K say in their paper that the correct answer is 101 with a sample size of 100. Well, it is quite close to 100. (In my opinion, the participants' error is very small.) The smaller the sample size, the larger should people change their answers. You might be able to better detect effects going below the original sample size. (Just a suggestion though.)**

We agree that this is an interesting direction, not only for this specific problem, but also for some of the other problems. However, we would like to keep our manipulations consistent across the different presented problems, and maintain our focus on the higher end of the spectrum to examine the implications of substantially increasing the sample size from those detailed in the target article.

To address this, we added a planned discussion of this suggestion as a direction for future research under "Limitations and future directions" in the Discussion section.

> **For Q4: I am wondering whether people understand well what a « likelihood » is. Btw, you are asking here the likelihood of having an association of 0.35 (like a point estimate?). I think that the likelihood to have this precise value is almost zero. In the original question, it is asked whether there is support for an association of 0.35. (So, it is in the confidence interval, which is more likely.) I feel like your extension questions are a bit differently framed from the original question.**

Great comments, thank you.

You raised two insights here. The first is regarding the need to examine a point estimate, and yes, we agree and understand, and one of the exploratory directions we discussed was to examine how people compare a single point estimate, of which the likelihood is extremely small, to a range that includes that point estimate. We do not have a clear hypothesis, yet we suspect that

people may overestimate the likelihood of the point estimate and possibly also underestimate the likelihood of a range that includes that point estimate (in the spirit of the Linda problem conjunction effect).

> **It would be useful to have the type of answer respondents can give for each question. (Numerical input, probabilities, Likert, etc.)**

Thank you, great suggestion. We added a new column in Table 3 (right column) in the main manuscript to indicate the type of response expected.

> **In my view, the replication is between « close to far » and « far ». To be conservative, I'd suggest keeping « far ».**

Thank you. We agree and accept. We adjusted our classification of the replication to "far". Please also see our reply to the editor on this point and adopting a second complementary framework.

> **Exclusions: page 36, you did not mention whether you'll include/exclude people who failed the attention checks.**

We realized we might have not been clear enough about our attention checks. We also realized that it might be valuable to give reviewers access to the preview of the Qualtrics so that they can see and experience the survey (if they are unable to import it to their Qualtrics account), and we therefore added the following to the "Design and procedure" subsection of the method section:

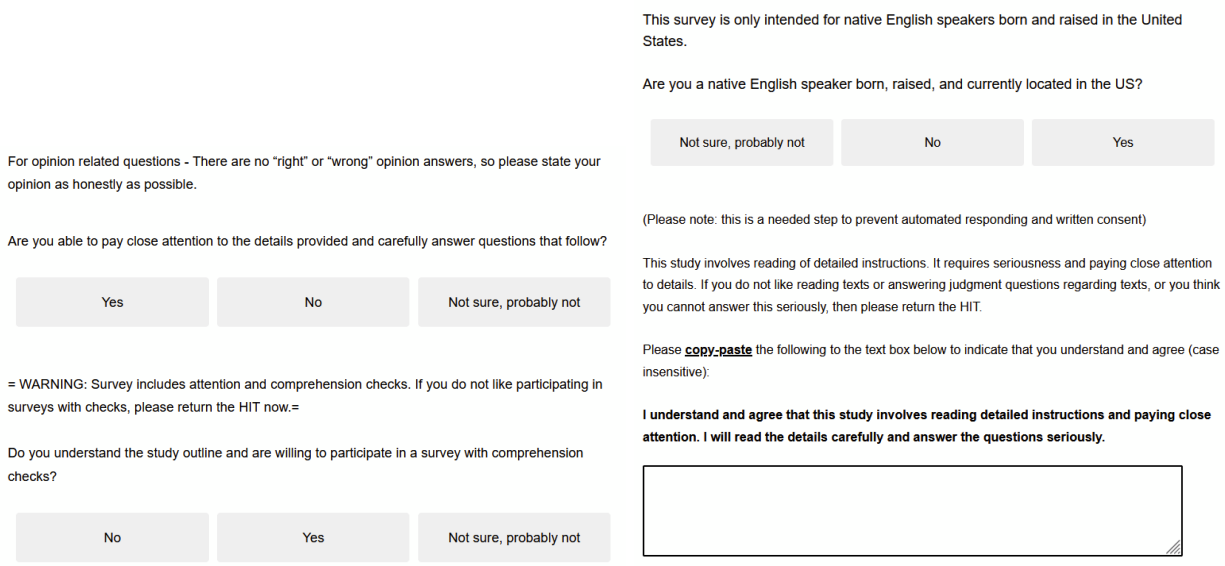
> [For review: The Qualtrics survey .QSF file and an exported DOCX file are provided on the OSF folder. A preview link of the Qualtrics survey is provided on:
>
> https://hku.au1.qualtrics.com/jfe/preview/previewId/b0b9be99-3191-4ae4-9c7d-41212b7f2f92/SV_bDWVv5m9EXpqgxo?Q_CHL=preview&Q_SurveyVersionID=current]

This is important for this section, to clarify what we wrote in our first submission:

> Participants indicated their consent, with four questions confirming their eligibility, understanding, and agreement with study terms, which they must answer with a "yes" and required responses in order to proceed to the study. Three of the four questions also served as attention checks, with the options order being rotated (yes, no, not sure).

This means that our attention checks are combined with consent, we rotate the yes/no/not sure options in the 3 consent/qualification questions, and answering consent with a yes to these questions is mandatory for taking part in the study.

Here is an example screenshot:



We also took the opportunity to clarify the section about inclusion.

Original:

> We will only include responses from participants who completed the entire questionnaire.

Revised:

> In our data analysis we will only include responses from participants who completed the entire questionnaire, have passed the consent checks at the beginning of the questionnaire (participants cannot proceed to the survey without correctly answering those), and rated with "seriousness" >=3 (on a scale of 1 to 5) and English understanding >=4 (on a scale of 1 to 7) in the funneling section at the end of the questionnaire.

Therefore, to return to your question, participants cannot begin the questionnaire unless they've answered the consent/qualifications with a yes, which serve as attention checks in choice rotation. And, beyond that, the funneling questions about seriousness and English understanding are both attention checks and match check, and we worked to make our inclusion criteria better.

In our revision, we also made sure to be clearer about this inclusion criteria in the R code as placeholder for the real data:

```
### INCLUSION CRITERIA CHECK

# (only to be used on final data collection)

# Check that participant completed the survey

# dataset <- dataset [which(!is.na(dataset$Satisfaction)),]

# Check seriousness > 3 (scale 1-5)

# dataset <- dataset [which(dataset$Seriousness>3),]

# Check English > 4  (scale 1-7)

# dataset <- dataset [which(dataset$EnglishUnderstanding>4),]
```

Reply to Reviewer #2: Dr./Prof. Kariyushi Rao

1 Summary of the Research Plan

The authors plan to reproduce, replicate and extend seven "empirical demonstrations" of the belief in the law of small numbers (LOSN) presented in Tversky and Kahneman (1971). The authors make several improvements upon the original paper. First, the authors merge all seven demonstrations into one procedure, presented in random order to a large sample of Amazon Mechanical Turk Workers located in the United States. Second, the authors seek to minimize jargon and statistical terms in the experimental stimuli. Third, the authors plan to perform statistical tests of the hypothesized deviations (whereas Tversky and Kahneman only provided descriptive statistics).

For each demonstration, there is a purported "correct answer" (according to Tversky & Kahneman). The authors have done a commendable job clearly articulating what those correct answers might be, given the target article is often unclear and fails to state the "correct answer" directly. The authors will measure the deviation of participants' responses from the correct answer they've inferred from the target article. The results of the present experiment will be compared to the descriptive conclusions presented by the authors of the target article.

[Note: The authors have revised their original snapshot to exclude a scholar sample. I find their justification for this choice perfectly reasonable. In my opinion, the lay sample is more interesting and important than the scholar sample. I agree that the present research plan is sufficiently complex and challenging, and the present results (that exclude the scholar sample) will provide a significant contribution to the literature.]

1.A Is the research question scientifically valid?

Yes. The present proposal meets the PCI standards. The research question is clearly defined. The research question is scientifically justifiable, and defined with sufficient precision as to be answerable through quantitative or qualitative research. The research question make sense in light of the extant theoretical and empirical literature in statistical reasoning, probability updating, and judgment and decision-making. The hypotheses are capable of answering the research question. The research question falls within established ethical norms. The authors have clearly distinguished work that has already been done from work yet to be done.

1.B Are the proposed hypotheses logical and plausible?

Yes. The present proposal meets the PCI standards. A priori hypotheses are coherent and credible. Hypotheses follow directly from the research question. There is a sufficiently strong mapping between the theory, hypotheses, sampling plan, preregistered statistical tests, and possible interpretations given different outcomes. The authors have explained precisely which outcomes will confirm or disconfirm their predictions.

Thank you, we very much appreciate the kind words and support. Your thorough and positive review has greatly helped us improve our manuscript.

1.C Is the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable) sound and feasible?

The present proposal is of sufficient quality to merit IPA from PCI, but some improvements could be made. The study procedures and analyses are valid, for the most part. The authors have performed a statistical power analysis to the best of their ability given the lack of information in the target article, with appropriate (conservative) adjustments. The proposed sample size is sufficient to provide informative results. The authors clearly state their rules for randomization of experimental participants, and for data exclusion.

The authors do plan to rely on conventional null hypothesis significance testing. The authors also intend to interpret negative results from their one-sample t-tests and one-way ANOVAs as evidence that an effect is

> **absent. The authors have not proposed Bayesian hypothesis testing or frequentist equivalence testing (inferential methods better capable of drawing conclusions about the implications of negative/null results). However, the proposed statistical methods are standard in the authors' field (psychology), and reviewers for the two "PCI-Interested" journals (JEP:G and JEP:LMC) will sometimes explicitly request traditional methods (e.g. ANOVA) be performed. So, I do not think the present analysis plan should preclude the authors from receiving an IPA from PCI.**
>
> **The authors also plan to implement the paradigm suggested by LeBel and colleagues (2019) to judge the extent to which their experimental results replicate the original results in the target article.**

Thank you for the comments, this is an important point to clarify. I think it has helped us realize that we might have not been clear enough on some of the points raised here.

We indeed focus on NHST, yet we think there was a misunderstanding regarding your interpretation of our intent as “interpret negative results from their one-sample t-tests and one-way ANOVAs as evidence that an effect is absent”. We were not completely sure which of our many planned analyses this was aimed at, but we think it is important to reframe that, using the LeBel criteria and accurate NHST interpretation, that if we observe null effects where we expected a signal then we conclude failure to find support for an effect. This is different from concluding support for the null. In addition, even when according to Tversky and Kahneman we are to expect a null effect (such that belief in the law of small numbers would predict no differences between control, X10, and X100 conditions for most of the problems), we reframed the hypothesis to an expectation to find differences (belief in the law of large numbers, reviewed in the introduction).

In addition, in our submission’s results section we have run and reported Bayesian effects. The ggstatsplot package that we used to produce our figures automatically includes the Bayes analysis at the bottom. We added that to the results but were not explicit enough in how we plan to use those.

To address this point, we added a section to the methods:

> In this extension, we aim to test the competing hypotheses of expected differences between the three conditions. Tversky and Kahneman’s “belief in the law of small numbers” would for the most part predict that people are not sensitive to sample size, and therefore that there should be no detectable differences between the conditions. A competing hypothesis, related to the “belief in the law of large numbers” mentioned in the introduction, is that there will be differences between the three conditions and that

people will update their answers in response to sample size differences. Our main analysis is using Null Hypothesis Significance Testing, which is focused on rejecting the null hypothesis, yet to address the possibility that we may fail to find support for rejecting the null, we complement all our analyses with Bayesian analyses using a prior of BF = 0.707. Bayes factor will be reported in our figures, using the R package ggstatsplot.

Suggestions:

(1) Exclusion Criteria

Participants are incentivized to lie on each of the self-report measures proposed as exclusion criteria, so it is unlikely these criteria will serve their intended purpose. I recommend the authors run a qualification survey in advance of the focal study, and include the following substitutes for their first two self-report measures in the qualification survey.

1. Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale). Ask participants the following two questions (or something similar):
1. What region of the United States do you live in currently? (Drop down list that includes "Prefer not to disclose")
2. What is your favorite thing about the region of the US where you live currently? Please respond with one complete, grammatically correct sentence. (Question should appear on a different page than the above. Question must be open response. Responses should each be read by the same human reviewer. Exclude all participants who do not provide a complete, grammatically correct sentence. Exclude all participants who provide non-sequitur responses, e.g. "I love my television.")
2. Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale).
Restructure this question using Drazen Prelec's Bayesian Truth Serum. Qualification survey responses should be checked by hand by the same person, to ensure that patterns are detected across responses (e.g. groups of participants colluding on the survey who paste copied text from internet sources instead of writing original responses). Qualified participants should be assigned an approval code, and the subsequent focal study should be restricted to those participants that have been assigned the approval code.

We understand the concern, and appreciate the suggestions. We took a different approach and strategy, which has worked well for us with this sample in our many other replications of similar

surveys. We employ the approved participants in CloudResearch, who have passed and regularly pass qualifications surveys by CloudResearch to ensure high quality responding. We have also taken many measures to ensure attentiveness, some of which - like attention checks - that we clarified above in our reply to the other reviewer and expanded on in the revised manuscript.

Generally, we would rather assume best intentions in our participants, and see no reason to doubt or test their honesty. The questions we are asking are not about private matters (which is what tools like “Drazen Prelec's Bayesian Truth Serum” are intended for), and we see no reason or incentive for respondents to hide their answers or to try and deceive us. We would also rather focus on objective clear measures that do not involve a coder, subjective evaluations of complex qualifications like grammar, and on responses that might not be reflective of what the survey is about.

> **I don't see the purpose of requiring participants to confirm they are native American citizens born and raised in the United States. Participants are prone to lie on these types of questions, and this particular question does not necessarily indicate English proficiency or any particular (relevant) level of education or acculturation.**

Thank you, we appreciate the opportunity to clarify this point.

We agree, there is no way for us to confirm anything regarding our participants, including not whether they were born and raised in the US. These questions at the beginning of our survey are provided as an acknowledgment that they read and understood that this is who we are aiming for, in the hope that it will minimize those trying to take the survey regardless. We also explained in the manuscript and clarified above that we use this as attention checks, given that we rotate the Yes/No/Not sure choice order.

> **(2) Compensation**
>
> **The hourly pay target for US-based Amazon Mechanical Turk Workers is too low. MTurk Workers in the US desire, expect, and actively seek out a pay rate equal to the most generous State minimum wage, which is $15.00/hour. The authors' target of $7.25/hour will result in selection issues, as more highly conscientious and experienced Workers are less likely to accept HITs at lower rates. The authors should also be aware that MTurk Workers can manually set hourly targets in their MTurk Dashboard that are perpetually displayed within the MTurk interface, so the $15.00/hour anchor will be salient for them.**

Thank you for the feedback.

We indicated the US$7.25 amount as a placeholder to indicate that we will meet federal regulations for compensation data collection in the US, with a section in the supplementary that is to be updated after data collection with the data.

We do aim to go beyond federal wages and meet and exceed higher local minimum wages in the different states while striking a balance with making the most of the limited budget we have for the project. We also take several measures to ensure satisfaction, examining duration and asking for pay feedback in a pretest of 30, and adjust pay accordingly. We will report all study parameters after data collection.

> **(3) Comments on Table 1**
>
> **Q3: The LOSN hypothesis should be rewritten as, " If a study reports that 0.8*X out of X infants preferred Toy A over Toy B, then people tend to perceive that as representative of the general population and therefore expect that 0.8*X out of X infants in the general population will prefer Toy A to Toy B. Regardless of what X is.**

We modified Q3 in Table 2 (previously Table 1) to the following:

> "If a study reports that 80% of X infants preferred Toy A over Toy B, then people tend to perceive that as representative of the general population and therefore expect that about 80% infants in the general population will prefer Toy A to Toy B. Regardless of what X is."

> **Q4: Given that you are trying to remove jargon and statistical language, shouldn't you avoid using the concept of a power analysis here? Instead, you should present a layman's explanation of what a "critical significance value" is. The power analysis version of the question might have been interesting when you were going to include a scholar sample, but without the scholar sample it doesn't seem as interesting or appropriate.**

Thank you. We added "power analysis" as a term in parentheses, but that is not needed, and so we removed that. We also appreciate the suggestions to improve on explaining what this analysis is.

The revised version of Problem Q4 is:

> Scientists who study two personality traits (Trait A and Trait B) expect there to be a positive association (relationship) between these two traits in the general population. In other words, scientists expect people with higher ratings on Trait A to also tend to have higher ratings on Trait B. On the possible range of associations from -1 (fully negative

> association) to 0 (no association) and 1 (fully positive association), the expected association in the general population is 0.35.
>
> You have just read a media report about a new study on the association between Trait A and Trait B. In that study, the scientists conducted an analysis to try and determine how many people need to participate in their study in order to convincingly be able to conclude support for a relationship between Trait A and Trait B, and determined that in order to be able to detect an association of 0.35, they would need to run a study with at least 79 participants.
>
> (Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)
>
> **Q5: The generalized hypothesis should be rewritten as, "If a study reports a surprising phenomenon using any sample, then people tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population."**
>
> **Same for the secondary hypotheses - the word "surprising" should be added in before "exploratory" in each case. The surprisingness of the phenomenon is really key here, especially from a Bayesian perspective. If a finding runs contrary to accumulated human knowledge (even lay knowledge), then our willingness to update in the direction of that finding should be smaller than if the finding does not run contrary to accumulated human knowledge.**

Thank you for catching that. We amended Table 2 (previously Table 1) accordingly:

> “If a study reports a surprising phenomenon using any sample, then people tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population.
>
> People tend to underestimate the sample size required to confirm the surprising exploratory finding.”

> **Q6: For similar reasons to the above, the generalized hypothesis should be rewritten as, (1) "People do not differentiate between exploratory studies that produce surprising results and confirmatory studies that seek to replicate those surprising results," and separately, (2) "People ignore sample size. Participants perceive the following to be equally representative: (1) an exploratory study with a sample size of X, and (2) a confirmatory study that seeks to replicate the results of the original exploratory study using a sample size of 0.5*X."**

Thank you. The aim of the generalized hypothesis column is to make it as generalizable as possible, and so we followed your recommendations regarding the first part and amended accordingly in Table 2 (previously Table 1). We made minor adjustments to make the second suggestion more generalizable to fit the title of that column:

> Laypersons do not differentiate between exploratory studies that produce surprising results and confirmatory studies that seek to replicate those surprising results.
>
> Laypersons ignore sample size when comparing exploratory studies and replication studies and consider them to be equally representative of the population regardless of sample size.

> **Q8: The generalized hypothesis should be rewritten as, "People overestimate the likelihood that a confirmatory study seeking to replicate several correlations found in an original exploratory study will produce support for at least 2/3 of those correlations, even if the confirmatory study has 2/5 the sample size of the original study."**
> **And, the LOSN hypothesis should be rewritten as, "If an exploratory study with a sample of X found support for Y correlations, then people overestimate the likelihood that a confirmatory study will replicate at least 2/3*Y correlations from the original study, even if the confirmatory study has a sample size of .4*X, regardless of what X is."**

We appreciate the suggestion, yet the suggestions deviate from the purpose of those columns. Our aim for those columns was to indicate a generalized hypothesis in the second column and to a hypothesis specific about sample size in the third column. Both your suggestions make the predictions too specific. Yet we feel that adding the information about this being a correlational study to be relevant. We therefore amended to the following:

> Generalized: Laypersons overestimate the likelihood of confirming exploratory correlational findings with a replication with similar or smaller sample size.

> LOSN: If an exploratory study with a sample of X found support for Y correlations, then laypersons overestimate the likelihood of finding support for these associations in a replication with a sample of 40% of X. Regardless of what X is.

> **(4) Comments on Table 3**
>
> **In general it seems odd to use the word "experimenter" with a lay population. I suggest either using the generic "scientist" or "you" (as in Q3 of T&K, 1971) or "toy company executives," etc.**

Thank you, we agree that it would help with clarity and consistently if we use one term throughout all problems. Much appreciated.

We updated all references throughout in both the Qualtrics survey and in the revised manuscript. References to experimenter(s), academic scholar(s), and scholar(s) have been changed to scientist(s).

> **Also, the notes on what it means to "find support" seem to commit a common error in the description of null hypothesis testing (that there is less than a 5% chance of obtaining X result if H1 is not true), and the language should be updated to avoid this error.**

We tried to be very careful about the way we framed this. Specifically, we previously wrote:

> What scientists typically mean by "finding support" is that there is a less than 5% chance of the data showing an effect in support of their theory when there is actually no true effect of relationship in the population being studied.

p-values are notoriously tricky to get right, and even more difficult to translate for laypersons, as captured vividly by “Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly” (Cassidy et al., 2019).

You did not explain nor provided references for what you thought was our common error in our framing. From what you did explain, we are unsure why you interpreted what we wrote to be:

> “that there is less than a 5% chance of obtaining X result if H1 is not true”.

This is not what we wrote. Rather, using your framing and terms, what we wrote is better framed as

> “that there is less than 5% chance of the data indicating support for rejecting the null hypothesis if the null hypothesis of no differences/association is true.”

Our definition and reframing was based on the common framing by Lakens (2021):

> The correct definition of a p value is the probability of observing the sample data, or more extreme data, assuming the null hypothesis is true.

Given that you did not provide any explanation, we are open to the possibility of being wrong about this, and asked the Twitter community for help on improving on this definition and translating it to laypersons (https://twitter.com/giladfeldman/status/1654056588737499136).

The most helpful reply was by Lakens (https://twitter.com/lakens/status/1654111997657489412?s=20):

> Assume a researcher compares 2 groups, such as two groups of students, on a measure, such as their exam score. P-values tell you how small the probability is that both groups differ in their exam score, if both groups of students are in reality equally good at the exam. If this probability is small, scientists act as if the difference is not just due to random variation in exam scores. They reject the idea that the groups are identical, and act as if the groups differ in how good they are in the exam.

Given the community replies, we changed our framing to the following:

> What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern, and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.

We are open to revising this further, and please ask that if you see further improvements needed that you explain clearly what the issues are, preferably with citations, and provide us with a clear suggestion as to how to improve on those.

References:

> Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, 2(3), 233-239.
>
> Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on psychological science*, *16*(3), 639-648.

> **Q1: The phrase "a sample of X people" may be misinterpreted as a subset taken from a group of X people. E.g. "There were 100 people, and we took a sample of 10 out of that 100."**
>
> **The "clarification" note also commits an error in its description of null hypothesis testing.**

Thank you, good point about “a sample of”. We adjusted throughout to remove “a sample of” to change from "a sample of X people" to “X people”:

Original:

> You read a media report about an experiment that was run on a sample of [20/200/2000] people, and the report indicates that the result shows support for their theory. The same experimenters ran the same experiment again (a replication) with a new sample of [10/100/1000] people of the same population.

Revision:

> You read a media report about an experiment that was run on [20/200/2000] people, and the report indicates that the result shows support for their theory.
>
> The same scientists ran the same experiment again (a replication) with [10/100/1000] people of the same population.
>
> (Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)

The clarification was adjusted to be the same as we indicated above.

> **If you are really trying to get away from jargon and statistical language, consider changing the scenario and prompt to the following:**
>
> **Scenario:**
> **"You read a news report about [20/200/2000] people who participated in an experiment to test scientists' theory of X. The report indicates that the results of the experiment support the scientists' theory.**
> **(Usually when scientists say they found support for a theory, it means they ran a statistical test that tells them if they ran the same experiment one hundred times, less than 5 of those experiments would produce the same results they got in the original experiment if their theory was false. Basically, they think it would be really hard to get the result they did if the theory wasn't true.)**
> **The report also indicates that same scientists have just run the same experiment again with [10/100/1000] new people from the same population."**
>
> **Prompt:**
> **"How likely is it that the scientists will find support for their theory again in the experiment they just ran with [10/100/1000] new people? (Indicate your response as a percentage out of 100; e.g. 0% means there is absolutely no chance they will find support for their theory, 100% means they will definitely find support for their theory.)**

Thank you for these suggestions. We made minor adjustments to our previous framings based on some of the recommendations. All that you suggested is tricky for many reasons that we would rather not go into, yet we will give an example to explain the complexity.

You suggested the following clarification:

> Usually when scientists say they found support for a theory, it means they ran a statistical test that tells them if they ran the same experiment one hundred times, less than 5 of those experiments would produce the same results they got in the original experiment if their theory was false. Basically, they think it would be really hard to get the result they did if the theory wasn't true.

Based on our limited understanding of p-values, there are many challenges to this framing, from the possible misunderstanding that p-values say anything about the likelihood of replications, that it allows for any strong inferences about theory, etc.

We are very grateful for you giving this a try, we know how hard it is to try and get this right, and we appreciate that very much. We give this as an example of why this is tricky, and why we tried to stay as close as possible to the original framing with minor changes to correct what is wrong, to improve clarity where possible. We would rather not start from scratch and reframe everything.

We do recognize that our study is a first step, and we hope others would be able to build and further improve on our work in future studies.

> **Q2: There appears to be a typo (underlined and bold in red below), and I think there's a big difference between the (original) phrase "known to be 100" and the (new) phrase "reported to be 100."**
> **I suggest the following: "The average IQ of all eighth graders in a particular city is 100. A scientist randomly chose [50/500/5000] eighth graders from that city to test their IQ. The [first eighth grader / average IQ of the first 10/100 eighth graders] tested out of the [50/500/5000] chosen by the scientist [has an IQ of / is] 150.**
>
> ***Important note*: I think the odds that the first child drawn from a sample of 50 having an IQ of 150 are higher than the odds of obtaining an average IQ of 150 from a contiguous sequence of 100 children drawn from a finite sample of 5000 children (I didn't work out the math on this one, so maybe the authors are right to assume that these odds are the same). If the odds are different then I don't think you are actually asking the same question when you increase the number from 1 to 10 to 100.**

Thank you, good catch with the known versus reported. We changed “reported” to “known”.

Excellent catch with 10/100 instead of the 500/5000. We implemented this correctly in the Qualtrics, but the table in the main manuscript was not updated. Much appreciated!

As for comparing the odds of 1 of 50 versus first 10 of 500 versus first 100 of 5000, this is an interesting point we had not considered. When we asked for advice on Twitter/Mastodon we were provided with a calculation that works for increasing 10 and 100 times.

Daniel Lakens (https://mastodon.social/@lakens/109784123276880564) wrote:

> “49*100+150=5050. 5050/50=101”

If it is that simple, as others suggested it may indeed be, then the calculation for the rest is:

> 490*100+10*150=50500. 50500/500=101
>
> 4900*100+100*150=505000. 505000/5000=101

However, we are open to the possibility that we are misunderstanding something here. We posted this on Twitter: https://twitter.com/giladfeldman/status/1659206111113928707?s=20, and the few replies we received did not come up with good options for improving further, especially given the constraint of the baseline control condition replication scenario, and the need to not overwhelm laypersons with statistical jargon.

We added the following to the methods section under the "Exploratory extension independent variable: Sample size manipulation" subsection:

> In addition, manipulating sample size may have additional factors that we had not taken into account, as for example reviewer Dr./Prof. Kariyushi Rao suggested the possibility that in Problem Q2 the likelihood of one person having an IQ of 150 in a sample of 50 is higher than the likelihood of an IQ average of 150 in the first 10 people in a sample of 500. We noted the complexity inherent in the manipulation of Problem Q2 yet to keep things simple, we decided not to further complicate the X10 and X100 conditions.

We would gladly improve on this point if clarified further, and in such a case would appreciate specific, constructive, clear guidelines on how to do so.

> **Q3: I'm not sure what it means for a "ratio of 80% of infants choosing Toy A over Toy B" to "persist." I think what's inferred by the grouping of the three questions is the following: "Suppose that [8/80/800] out of the [10/100/1000] infants in your second study also preferred Toy A over Toy B. If you were going to run one final study to conclude once and for all that 4 out of 5 infants in the world population prefer Toy A over Toy B, what is the minimum number of infants you would need to include in that final study? (Try your best to estimate.)"**

Thank you, that is a very good point. We are grateful for the suggestion

Original:

> Assume that the ratio of 80% of infants choosing Toy A over Toy B persists. Try and estimate - what is the minimum number of infants that would be required in a new study for the scientists to be able to conclude that infants overall have a preference for Toy A over Toy B?

Revised:

> Suppose that [8/80/800] out of the [10/100/1000] infants in the scientists' second study also preferred Toy A over Toy B. If these scientists were to run one final study to conclude once and for all that 4 out of 5 infants in the world population prefer Toy A over

Toy B, what is the minimum number of infants that the scientists would need to include in that final study? (Try your best to estimate.)

Q4: The use of the words "positive association" and "expected association" are foreign to the lay population. Also, the meaning of 0.35 is ambiguous here, because the grade scale is not specified. Americans are used to a 4.0 GPA scale, or a letter-grade scale from A to F. Using a need for achievement scale that ranges from [-1, 1] with a GPA scale that ranges from 0 to 4 (and sometimes up to 5.0) makes the meaning of 0.35 difficult to understand for a lay person. I suggest using the following scenario instead:

Scenario:

"Psychologists who study two personality traits (Trait A and Trait B) expect there to be a positive relationship between these two traits in the general population. In other words, psychologists expect a people with higher ratings on Trait A to also have higher ratings on Trait B. Both traits are rated on a scale from 0 (does not exhibit the trait at all) to 10 (exhibits the highest level of this trait). Specifically, for each 1-point increase on the Trait A scale, psychologists expect people to exhibit a 0.35-point increase on the Trait B scale.

You read a news report about a study on the relationship between Trait A and Trait B. Before running that study, psychologists performed a statistical test using all of the existing evidence about the relationship between Trait A and Trait B. The test is supposed to determine how many people need to participate in the study in order to accurately detect the relationship between Trait A and Trait B. The result of the psychologists' test indicated that they need at least 79 people to participate in their study in order to accurately detect a 0.35- point increase in Trait B for each 1-point increase in Trait A.

(When scientists say "accurately detect" they usually mean that if they ran 100 experiments they would only detect the 0.35-point increase in less than 5 of those experiments if there wasn't really a positive relationship between Trait A and Trait B.) "

Prompt:

"If the psychologists ran their study with 79 people, how likely is it that

> **they will find support for a 0.35-point increase in Trait B for each 1-point increase in Trait A?**
>
> **(Usually when scientists say they found support for a relationship between two things, it means they ran a statistical test that tells them if they ran the same experiment one hundred times, they would find a relationship between Trait A and Trait B in less than 5 of those experiments if there wasn't really a relationship between those two things. Basically, they think it would be really hard to find a relationship there if it didn't really exist.)**
>
> **Indicate your response as a percentage out of 100 (e.g. 0% means there is absolutely no chance they will find support for a 0.35-point increase in Trait B for each 1-point increase in Trait A, 100% means they will definitely find support for a 0.35-point increase in Trait B for each 1-point increase in Trait A)."**
>
> **For all extension questions, I suggest using the verbose "What is the likelihood that in their sample of [79/790/7900] people, there will be a 0.35 -point increase in Trait B for every 1-point increase in Trait A?" instead of the association language you have now.**

Thank you, we appreciate the suggestions, and yet decided to keep our framing. We understand this is tricky with laypersons, but the incorporation of scale range and trying to describe a correlation in terms of an increase in point in Trait A for Trait B is not - in our limited understanding - an accurate description of a correlation, and we worry it might be even more difficult for laypersons to comprehend and understand. In our previous submission we tried to address the issue by describing the range of the correlation:

> On the possible range of associations from -1 (fully negative association) to 0 (no association) and 1 (fully positive association), the expected association in the general population is 0.35.

We did however incorporate some of your suggestions and tried to further improve on Q3, yet with only minor changes and without any radical changes:

> Scientists who study two personality traits (Trait A and Trait B) expect there to be a positive association (relationship) between these two traits in the general population. In other words, scientists expect people with higher ratings on Trait A to also tend to have higher ratings on Trait B. On the possible range of associations from -1 (fully negative

> association) to 0 (no association) and 1 (fully positive association), the expected association in the general population is 0.35.
>
> You have just read a media report about a new study on the association between Trait A and Trait B. In that study, the scientists conducted an analysis to try and determine how many people need to participate in their study in order to convincingly be able to conclude support for a relationship between Trait A and Trait B, and determined that in order to be able to detect an association of 0.35, they would need to run a study with at least 79 participants.
>
> (Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)
>
> **Q5: Stripping out the words in bold here - "<u>you</u> completed a <u>difficult and time-consuming experiment</u>" - really changes the nature of this question. I don't think you're measuring the same thing anymore with the updated version. Also, the updated version still has a lot of unfriendly language for laypeople. Same thing with taking out "you" in the origininal "you were to run the same study again" changes the nature of the question. A person's confidence in their own ability to do the same thing twice is a very different judgment than a person's confidence that some stranger could do something they have no relevant information about twice. The "Reminder" note commits an error in its description of null hypothesis testing.**

That is a good point. Thank you.

We agree this is an issue and something that has not been looked at in the target article, it would have been best to ask this more broadly than ask people about themselves. In our translation to laypersons, there is no way around it, laypersons do not conduct studies, so this must be about someone else.

We added this as a planned point for discussion following Stage 2 in the “Limitations and future directions” subsection:

> [Planned discussion raised by reviewer Dr./Prof. Kariyushi Rao: Limitation in the target study about the questions being about self versus others’, which may incorporate some degree of self-efficacy. In the laypersons version everything had to be translated to others’, and therefore we indirectly addressed this issue.]

> **Q6: Laypeople don't really understand what it means for results to be "in the same direction," so it would be better to say something concrete here instead. The "Reminder" note commits an error in its description of null hypothesis testing.**
>
> **Q8: Laypeople don't understand factors, associations, and correlations. The language should be updated to use words like "relationships" and phrases like "an increase or decrease in X tends to happen whenever there is an increase or decrease in Y." The "Clarification" note commits an error in its description of null hypothesis testing.**

Thank you. We adjusted some of those in the revision.

For Problem Q6:

> Follow-up question: Suppose that the scientist reran the study that they reported, this time with [20/200/2000] new people from the same population, and the results of the new study seemed to generally be consistent with the previous study, yet did not meet the scientific criteria that scientists typically set for them to be able to conclude that they found support for the findings.
>
> (Reminder: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.).

For Problem Q8:

> You read a news story about a study in the field of personality and social psychology that examined the associations (relationships) between 20 different personality traits, and the study was conducted on [100/1000/10000] people.
>
> In examining the 20 traits, there are 190 possible associations (association of trait 1 to trait 2,3,4…20, trait 2 to traits 3,4,5…20… up to associations between trait 19 and trait 20).
>
> Of the 190 possible associations, the study found support for 27 associations, and with slightly stronger associations for 9 of those 27 associations.
>
> On the possible range of associations from -1 (fully negative association) to 0 (no association) and 1 (fully positive association), the average of the 27 supported

associations was .31. Overall, the scientist thought that the pattern of relationships seemed reasonable and consistent with their theory.

(Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)

(5) If I understand correctly, the Conditions are defined by the magnitude of X. It seems that presenting participants with seven questions all having the same magnitude of X reinforces the validity of X as an appropriate magnitude (especially when you consistently refer to "experimenters" and "researchers" in the question prompts). It was not clear to me why all three versions of each question are not randomized across participants instead of having three different sets of participants each focus on a specific magnitude of X. E.g. It should be possible for a given participant to see one question with magnitude X, another with magnitude 10X, and another with magnitude 100X.

Interesting point. Both options (randomizing once vs. randomizing for each question) have merits and downsides, and we decided on the one that we thought was simplest. Given how different the problems were and that the X10 represent very different numbers, we do not see any reason for concern. We see some benefit to keeping the control condition to be about the replication, to fully mirror the target's statistical properties (even if translated to laypersons'). This design might also allow for easier complementary exploratory analyses comparing not only conditions neutral, to X10, to X100, but also comparing pattern of results across the different problems.

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

Yes. The present proposal meets the PCI standards. The Stage 1 protocol contains sufficient detail to enable replication by an expert in the field and ensures protection against research bias, undisclosed procedural, or analytic flexibility. The protocol specifies sufficiently precise links between the research question, hypotheses, sampling plans, analysis plans, and contingent interpretations given different outcomes.

> **1.E Have the authors 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?.**
>
> **Yes. The present proposal meets the PCI standards. The authors include a statistical sampling plan that is sufficient in terms of statistical power and/or evidential strength. The authors have minimised discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features. The authors describe attention checks. However, manipulation checks are not described.**

Thank you for your thorough review. We appreciate the time and effort you put into evaluating whether the proposal meets the PCI standards. Your constructive detailed feedback and suggestions were invaluable in strengthening the study.

> **Suggestions:**
> **(1) Direct Replication**
>
> **It would be informative to include the original forms of each stimulus, with *no adjustment* from the target article, in the procedure. I do not think every participant needs to see both versions of each stimulus.**
>
> **I can think of two ways to incorporate the original stimuli in your procedure that won't add a great deal of time to the procedure for each participant: (1) participants could be exposed to one duplicate (e.g. a given participant responds to both your version of Q1 and Tversky & Kahneman's version of Q1, and to your versions of Q2-Q8; another participant responds to your version of Q2 and to Tversky & Kahneman's version of Q2, and to your versions of Q1 and Q3-Q8);**
> **(2) for each participant, one of the seven stimuli could be the Tversky & Kahneman version, and the other six could be your version.**

We appreciate the suggestion. This is something that we have been debating for long, and at the end decided to go with only the option of full translation of the stimuli for laypersons. The laypersons version undertaking is challenging enough (as you commented above), there are many other complexities with the scholars version, and mixing the two can confuse and frustrate participants further and add factors that are difficult to anticipate and address.

However, to be clearer about the adjustments we made, we see the value of moving the original version to Table 3 so that readers can compare the original to the adjusted and then to our

translation to laypersons', and researchers can build on our research to conduct future studies with more complex methods like the one you suggested.

We added the following to the planned discussion section:

> [Planned discussion raised by reviewer Dr./Prof. Kariyushi Rao: Can run additional studies comparing the use of the original scenario, to our scholars adjusted scenario, to our laypersons scenario, for both scholars and laypersons, to see how the framing affects interpretation.]