Reply to decision letter reviews: #165

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

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

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

A track-changes manuscript is provided with the file:
"PCIRR Peters et al 2006 replication & extension-main manuscript-track-changes.docx"

Summary of changes

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

Section	Actions taken in the current manuscript
General	We added more details and amended the statistical methods to improve clarity and transparency.
Introduction	R1: We revised Table 4-7 (study design) and made some changes to improve clarity. R1 & R2: We provided justifications for why the continuous approach is better than dichotomization in new part "Extension: Numeracy as a continuous measure".

Section	Actions taken in the current manuscript
Methods	R1: We updated the deviations section. R1/Original authors: We adjusted the task order and placed the numeracy scales after the completion of four tasks.. R1: We added an extra check prompt to the questionnaire and will exclude participants who self-report looking for answers of numeracy scale. R1: We updated our reporting and statistical analyses for a median split on the measure of four studies.
Results	R1: We simplified our initial regressions analyses and changed those to reporting correlations, and supplemented those with Spearman rho for non-parametric. We will reported assumptions checks for ANOVA and t-tests post data collection. R1 & R2: We created a new variable of randomized numeracy scores of both numeracy scales. Therefore, we replicated the original data analysis. In addition, we replaced the linear regression in Study 1, 2 and 4 with correlation.
Discussion	R1: We will add a section on limitations of methodology. R2: We will add a section on potential weakness and improvements of individual differences between measures. R1 & R2: The discussion is only to be completed in Stage 2 following data collection
Supplementary materials	R1: We added details of Qualtrics contents and settings.

Note. R1/R2 = Reviewer 1/2

Response to Editor: Prof. Chris Chambers

> **Two expert reviewers have now evaluated the Stage 1 submission. The reviews are generally encouraging and the initial submission is already within striking range of meeting the Stage 1 criteria. The reviewers do, however, highlight a considerable range of areas that would benefit from refinement, from the clarification of methodological details, to strengthening the justification (and providing expanded discussion of limitations) concerning specific design decisions, to confirming the reliability of key measures, and improving the clarity of presentation. On the basis of these reviews, I am happy to invite a comprehensive revision that addresses all points.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit. We revised the manuscript according to the feedback, and we believe the revised version is much stronger as a result.

Response to Reviewer #1: Prof. Elena Rusconi

> **This is part of a larger project assessing replicability of popular findings in decision making research and involving a formative component. I applaud the scientific effort and its great formative value. I agree with the authors on the relevance of a replication of Peters et al.'s (2006) study and I am listing a number of comments/suggestions for improvement (a few minor, a few more substantial) below.**

Thank you very much for the encouraging and positive opening note and the time invested and detailed feedback in helping us improve.

> **SNAPSHOT-Hypotheses**
>
> **Some of the original hypotheses have been reframed (e.g. numeracy as a continuous variable) but for a replication it may be better to use first the original formulation and list the re-framed hypothesis as an extension.**

We understand this concern, and we thought about this long and hard, and decided we would much rather keep the continuous hypotheses given that they are more accurate of the study design (rather than the original's analysis of that design). Given that this was not an experimental design in either the target article or the planned replication, framing the hypothesis based on the chosen dichotomizing analysis of the original may lead to misinterpretations, because it may read to audiences as if the original manipulated numeracy.

To address this suggestion, we added the dichotomizing framing below each hypothesis, with a note indicating our preference for the continuous framing.

For instance, above "Higher numeracy is associated with weaker positive-negative framing effect" in Table 2, we added that the less numerate participants show a stronger framing effect than the highly numerate participants.

SNAPSHOT-Methods (differences):

The authors are planning to recruit participants online via Amazon Mechanical Turk. This aspect is problematic as it cannot be checked if participants use online shortcuts/a calculator to answer numerical questions.

Yes, good point, this is a limitation of the study that we tried to address. We and many other labs ran several such replications that were tested online for general knowledge or calculations that can be easily Googled or computed, as for example with anchoring, and these were often concluded as successful replications.

We employed several best practices for such adjustments to online samples:

1. We added a pledge block at the beginning of the questionnaire to ask participants not to look for answers.
2. We added timers to questions to allow for exploratory reaction time analyses.

We appreciate the feedback encouraging us to try and do better here, so to further address this concern, we took the following actions:

1. We added a question at the end asking participants "Our research depends on you using your intuitions to answer our questions, so it is very important for us to know: Did you look up any questions? Did you use any aid to answer these questions? You will be paid regardless, and there is no penalty, but for the sake of our research, we need to know "I DID NOT use any aids in answering this survey" "I DID used external aids to answer this survey".)
2. We added this as an exclusion criteria.

In addition, we will discuss this as a limitation of our study in our discussion following data collection. We added a placeholder in the revised manuscript's discussion.

SNAPSHOT-Conclusions

It is mentioned that d will be transformed to r, in order to compare with the original effect size. However, this applies to the extension (the replication results may be directly comparable, and should drive conclusions on the replication outcome).

We will conduct two statistical analyses. The one using Cohen's d was used in original article, which we could compare to the original results directly. The other using r will be used in extension analyses, which we use an approximation conversion to allow some comparison with the original effect. We report both pre and post conversion effect sizes.

STUDY DESIGN TABLE

From this table it becomes clear that the original analyses will be performed too. The replication outcome should be decided on the basis of those analyses.

Does the power analysis take into account the extensions or will these be regarded (and discussed) as a secondary/exploratory component of the study?

We conducted power analysis and sensitivity analyses based on the target article's reported effects, though we expect effects to be stronger with our extensions using continuous variables. We then multiplied that by 2.5, and added extra for possible exclusions. This should be much more than needed to address any additional analyses.

To our sensitivity analyses we also added a reference to correlations: Our sensitivity analysis indicated that a sample of 850 would allow the detection of $f = 0.12$ (one covariate, groups = 2, df = 1, 95% power, alpha = 5%, one-tail), an effect much weaker than any of the effects reported in the original, and the detection of $r = 0.12$ in our continuous measures extension, an effect considered weak in social psychology (Lovakov & Agadullina, 2021).

Reference:

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51(3), 485-504.

Relationship between numeracy as a continuous variable and decision making outcome/confidence: investigated via linear regression but the relationship might not be linear. Clarify that confidence judgments refer to the decision making tasks. Conclusions about the continuous variable analysis outcome should not overrun the conclusions about the replication Peters et al.'s study.

Thank you, good point.

We added the report of simple assumptions checks for correlation to both the code and the results section. We also mentioned that we will add assumption checks for ANOVA and t-tests in both manuscript and supplementary.

We clarified that the focus of the replication assessment would first be on the original analyses conducted in the target article, with a note on the additional analyses.

We also worked to simplify our analysis, and moved from regressions to correlations, supplementing the Pearson r with Spearman rho as a nonparametric measure to capture monotonic relationships.

Main Text

Introduction

The section on "affect and numeracy" could be better tuned on the manipulation of Peters et al.'s experiment 4, where the focus is on the disadvantage of being highly numerate under certain conditions.

In a replication we aim to revisit all studies. We were not sure regarding the logic behind emphasizing one study over the other. The suggestion of which of the studies to focus on felt a bit subjective and unclear to us, and we were aiming to cover all studies.

Section "choice of study for replication". Overall statements should be backed up more robustly.

We were not sure what "more robustly" indicates here. Some guidance would have been appreciated. We tried to elaborate a bit further. We could add more given more specific editorial guidelines.

> **Low power: please clarify what the target effect of interest could have been and why the specific number of participants was insufficient for each study – here or later, under each specific study sections (please note that some additional or supplementary analyses reported in the original paper may be exploratory/of secondary importance and it should be taken into account when evaluating power in the original study).**

It is very difficult to assess what the phenomenon's target effect size is, so any kind of analysis here would feel forced and subjective. We toned down the references to “low-power” and “underpowered” and instead mentioned the exact samples used and the need for larger samples.

> **Methods: Peters et al. stated that dichotomization was introduced to address data skewedness in Study 1 and maintained it for the other experiments as well; had the sample been larger for Study 2 and Study 3, it looks like the authors would have considered other splits (such as a split based on quartiles). Please clarify if/why this procedure is not suitable and/or ideal. It would be useful to mention here whether you will calibrate your replication so that it will achieve sufficient power to test for a relation between a continuous numeracy variable and decision making biases, and that you will also replicate the original analysis for a direct comparison with Peters et al.’s study (i.e. with numeracy dichotomized).**

We added a new section in Introduction, “Extension: Numeracy as a continuous measure” to clarify why such a procedure is not ideal. We tried to argue the disadvantages of dichotomization (e.g. power loss) and advantages of treating numeracy as a continuous variable (e.g. lower required sample size to achieve the same level of power).

> **One of the major methodological differences between the proposed replication and the original study (online vs in person testing), should be mentioned here and discussed. Another important methodological difference (all participants will take part in all the tasks), should be also mentioned, along with its possible advantages and disadvantages.**

We noted the two methodological differences in the deviations section in the method section.

The extension about study-specific self-efficacy is very interesting. However, the introduction of a summary confidence judgment would certainly not interfere with the replication effort if a between-participants design was adopted and if participants were not informed about having to provide a confidence judgment at the end. Could its introduction affect decision making processes on following numerical tasks in a within-participant design, as the proposed one? (e.g. Boldt, Schiffer, Waszak & Yeung, 2019; Confidence predictions affect performance confidence and neural preparation in perceptual decision making).

Given that numeracy is considered an individual-difference measure, and we are measuring associations between traits and various decision-making tasks, we consider these effects stable beyond minor contextual elements, otherwise it would undermine the entire meaning of this trait-decision associations. We consider all our adjustments minor, and in this specific case, per each of the studies we are adding a question after the question to help address a new research question. In this case, a within-subject unified studies design is a major strength as it would allow us to assess associations between the different tasks, and to examine order effects, to be able to empirically examine whether this is indeed concern by conducting analyses with order as a moderator, and with that gain valuable insights if this is indeed a factor. Should we fail to find support for the original findings, we can then explore to examine possible reasons for that, taking order into account.

Methods

Power analysis: p.21 "using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) for the statistical tests in each of the decision-making risk paradigms separately (i.e. framing effect, frequency-percentage effect, ratio bias and bets effect)."
The four studies are not conducted with separate groups of participants. What is the familywise alpha across the four main tests? Please state whether you have included any corrections to your power calculations.

We are addressing every study independently, and analyze the effects of each study independently, and given that we are replicating a specific analysis in the target article, we are using their criterion, and therefore we did not adjust our alpha or include any corrections our analyses or to the power analyses.

The baseline power analysis were aimed to detect the smallest effect of interest of an association between the trait and the decision-making tasks of any of the studies. Our final sample is several times the required sample size according to baseline power calculations, so we are far better powered than needed addressing any concerns that may arise from additional analyses: our

sample of 850 should be able to detect $r = 0.14$ and $f = 0.15$ (one covariate, groups = 2, df = 1, 95% power, adjusted alpha = 0.6%, one-tail).

> **p.21 state what result (effect within the two-way between-subject ANOVA) was the one requiring a larger sample size.**

The largest sample size was calculated from two-way between-subject ANOVA between numeracy and affect in Study 4. Please see details (i.e., input and output via GPower) under the section, Power analysis of original study effect to assess required sample for replication, in the supplementary (p. 16).

> **p.22 Please list all the recruitment criteria and quality data checks beforehand (at the moment a list is provided but it ends with "etc."). Please provide any threshold you will use to exclude participants after data collection, if based e.g. on Qualtrics diagnostic scores.**

We removed the etc., and we provided all the details we know at the moment (both in the manuscript and the Qualtrics). As indicated in that section, we will be pretesting everything prior to data collection to ensure it all goes well, and in case it does not and adjustments are needed then we will provide all updated information about adjustments after the data collection.

> **Table 3: The medium of Peters et al.'s study is paper and pencil, most likely in person, as the questionnaires were described as being "administered" rather than self-administered (p. 408) (perhaps this could be confirmed by Peters?). The year in which Peters et al.'s participants were tested is 2005 (if not earlier; p. 413), please amend.**

We are unsure what the exact issue is here. We simply followed what was written in the target article - "administered". We do not consider this an important issue, and would gladly remove or amend this given clear editorial guidelines.

We amended the test time to 2005. Thank you.

Methods

Design: Replication and Extension

In addition to the comment under the introduction section: the confidence rating is presented on the same page as the other questions, and all the answers can be changed after seeing/replying to the question about confidence. For a closer replication of Peters et al, the confidence rating could be required on a separate page, only after completion of a scenario/task.

Great point, thank you for bringing this up. Much appreciated!

We added page breaks before the confidence questions. Therefore, for each given study, the extension is separated and in a next page from the replication, and the participants cannot go back and amend the results when seeing the confidence extension. We further used piped text to display to the participants what they indicated in previous questions and then asked them the confidence questions.

Tables 4-7 (study design): should be checked and possibly reorganised to improve clarity and consistency (e.g. you could use the words "scenario" and "conditions" that are used in the main text, to achieve more clarity and distinguish between variable names and variable levels). E.g. for Study 1: IV2: Framing scenario or Frame (as in Peters et al.); IV2 – condition: Positive; IV2 – condition: Negative; also Numeracy is indicated as a between subjects IV like the Framing scenario though it is not manipulated but measured and no level can be indicated. Perhaps add a header with: manipulated variables for the first column and measures collected for the second. After reading the introduction, it is unclear whether the numeracy measure indicated in these Tables is the same as the original measure or a novel one – or both. For Study 2, IV2 should probably be indicated as Risk scenario (or Format, as in Peters et al.) rather than as "frequency-percentage effect" (condition 1: Frequency, condition 2: Percentage). Study 3 does not seem to have a manipulated IV. In Study 4 the manipulated IV could be Bet scenario (or Bet type) rather than "bet effect".

Terrific suggestions, thank you!

We made many revisions to Tables 4 to 7 (study design) and made some changes to improve the clarity. We added more information about the manipulations, we cleared text from the description of the dependent variables, and reformatted. We added information about numeracy to refer to both between-subject and continuous, and to both the original scale and the extension. "Bet effect" was changed to "Bet type" in Study 4. "Frequency-percentage" was changed to

"Frequency-percentage description (risk format)" in Study 3. Study 1's manipulation was changes to "Positive-negative framing".

Procedures

The numeracy was presented after the decision tasks in Peters et al. (2006) but this aspect is not preserved in the current replication, why are the authors planning to present the numeracy scales first?

Thank you for raising this. Much appreciated. This was an oversight regarding the task order. We also received similar comments from Peters after sending our preprint (i.e., put the numeracy task at the end). We adjusted the task orders in Qualtrics and amended the manuscript indicating that participants will finish both numeracy measures after completing tasks of four scenarios.

Instructions ask for participants to answer via their gut feelings/intuitions (this is partly functional to prevent them from checking their answers online). But was this prime present in the original instructions? Would it not be more ecological to let low- and high- numerate individuals choose their preferred approach?

Thank you, we removed this instruction.

Manipulations

Table 8 should be updated with the deviations from protocol outlined above.

We updated Table 8 including all the deviations written above.

Page 32: please provide reasons why you think that dichotomization is the main weakness (over and above sample numerosity?

Yes, we added a new part "Extension: Numeracy as a continuous measure".

Please specify whether linear regressions will be tested with both numeracy scale scores. And what is plan B, if the assumptions for linear regressions are not met by your data?

We simplified our analysis and replaced the regressions with correlations for both numeracy scales. We also supplemented the analyses with reporting Spearman correlations for more robust reporting.

Results

The dichotomization was originally performed on the basis of data distribution (median split). Why do the authors plan to apply Peters et al's

thresholds rather than a median split (as in Peters et al.'s study) to their own distribution of scores? The latter might improve their chance of having even groups to compare – important, given the use of parametric stats. It could well be that their median split will overlap with that of Peters et al.'s but it could also be that their sample is more (or less) numerate on the whole (information about demographics in Peters et al. is scarce). Related to this point, a weakness in Peters et al.'s paper is the (likely) uneven distribution of experimental conditions between numeracy groups (the division in groups was performed after assigning conditions to participants) – the numbers in each group are not reported in the original paper.

Thank you for raising this. Much appreciated. You are very right.

We realized this was an oversight in our description of Peters et al.'s reporting regarding their dichotomization. Peters et al. indeed conducted a median split in Studies 1 through 4, and after recognizing their median, performed the 2-8 versus 9-11 split.

Rather than relying on their results' 2-8/9-11 split, we perform a median split on the measure of four studies. We amended our reporting in the revision.

Clarify what numeracy scale you are using for regressions.

We are now reporting correlations for both numeracy scales.

"The results of the Rasch-based numeracy scale generated by simulated data were hard to analyze. 958 out of 1000 participants achieved zero marks and the rest of them all achieved one mark only."

Re: Rasch-based numeracy scale, could a list of 1000 random numbers ranging between min and max score be generated in Excel instead?

Yes, thank you. This is an excellent suggestion, and we appreciate the nudge to do better here.

To address this, we created a new variable where we randomized (using R) a numeracy score from 0 to 11 instead of calculating the numeracy from the random number. The dataset of score was meant only for the purpose of this simulation. Using this simulated variable we now report both results of original statistical analysis (e.g., ANOVA) and extension statistical analysis (i.e., correlation). Please see the updated results in our revised results section.

Supplementary Materials

p.4: "Study 1" please eliminate "mixed within-subject" (this ANOVA should have 2 between-subject factors; applies to Table 1 on p.5; even

> **though in Peters et al it is indicated as a "repeated measures" ANOVA); please eliminate or replace "with five students"; correct typo "numerach"**

To clarify things a bit, we reframed the "mixed within-subject ANOVA" to "2 between x 5 within mixed-ANOVA". We believe this is what the authors meant when they referred to "in a repeated measures ANOVA… " (p. 408).

> **p.17: pls specify the topic of the "funnelling" questions.**

The funneling questions were included in the exported Qualtrics, and are now specified in the manuscript: seriousness towards the survey, study purpose conjecture, and feedback.

> **p.32: pls clarify if participants will be able to answer from a mobile phone/whether you plan to include participants answering from a mobile phone.**

We did not plan on blocking mobile answering.

We added the statement in the supplementary:

> We did not disallow participation using any specific devices.

> **Pls clarify if, after 8 minutes have passed, participants will still be able to complete or will be logged out automatically.**

We added the clarification that participants will have 30 minutes to complete the survey in the "Additional information about the study" section.

“The expected completion time was set at 5 minutes in advance”: Does this mean you allowed 5 extra minutes compared to the expected completion time? What age were the 30 pilot participants? How realistic is the expected completion time for younger/older participants? (age range: 0-100)

The “Additional information about the study” section in the supplementary was meant to be updated following data collection. We added a note indicating that.

We indicated in the “Participants” section in the main manuscript that we were expecting 5-8 minutes, and will be pretesting this and adjusting our estimates and pay accordingly.

We will only conduct the pre-test and data collection following Registered Report Stage 1 in-principle acceptance, and they will be part of the target sample population.

Response to Reviewer #2: Prof. Daniel Ansari

> **I very much enjoyed reading this very-well written and clearly organized Stage 1 Registered Report entitled:" Revisiting the links between numeracy and decision making: replication of Peters et al. (2006) with an extension of examining confidence."**
>
> **As far as I can tell this Stage 1 Registered Report meets all the necessary components of a Stage 1 Registered Report. The sampling and analyses plans are both very clear. I also thought that the justification/rationale for conducting the replication study was very clear. It was also helpful to have the results from the randomized dataset as this helps to understand what the Stage 2 manuscript will look like following the actual data collection.**
>
> **Given my overall positive impression of this Stage 1 Registered Report manuscript, I only have a few concerns and suggestions for improvement.**

Thank you very much for the positive and supportive opening note.

> **1. While I fully agree with the authors that the hypotheses are best investigated using correlations and regressions rather than dichonomizing on the basis of numeracy, I would have liked to have seen more clear justification for this approach. The authors do not directly tell the reader why a continous approach is a better approach to handling this kind of data and research questions. I think it would be instructive if the authors provided more justification for this and pointed out the limitations/problems of a dichotomous approach and, by extension, the advantages of treating numeracy as a continous variable.**

Thank you. We provided extra justification for discussing the weaknesses of dichotomizing and the advantages of complementing these with analyses of continuous measures. We added a dedicates section in the introduction - “Extension: Numeracy as a continuous measure”.

> **2. While the internal reliability of the two numeracy scales are reported, no such data is reported for the 4 manipulations. Is this a concern? Would we not want to know the reliability of both the independent and dependent variables? If the manipulations are unreliable this might explain any null results that might be uncovered?**

As replicators, we take those as is. In a replication, we aim to repeat the methods of the original, and a re-validation of the target’s measures goes far beyond the scope of a replication.

All we can comment here is that it seems that their dependent measures were mostly choices/single items and that their manipulations seem common judgment and decision-making paradigms.

> **Related to this I was wondering how suitable the measures obtained from these manipulations are for individual differences questions, because they are derived from experimental paradigms. It has been established that measures derived from experimental research are often not well-suited for individual differences research (see: https://link.springer.com/article/10.3758/s13428-017-0935-1).**

This seems like a concern questioning the entire numeracy individual-differences literature, and addressing this question goes far beyond the scope of our manuscript.

Our aim with replications is to simply repeat what the original article did, with minor needed adjustments and possible extensions. As replicators, we take it as is.

To address this comment, we will add this to our discussion, considering it as a potential limitation and illustrating possible improvements.