Reply to decision letter reviews: #162

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/FnVnbMiRfTSF

A track-changes manuscript is provided with the file: "PCIRR-RNR-Lerner&Keltner(2001)-rep-ext-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	R2: We adjusted the PCIRR study design table for a better summary.
	R3: We added the ethics approval information and uploaded approval to OSF. We differentiated the two methods for categorizing events as ambiguous/unambiguous. We added the keywords "risk preference" "hope" "optimism" to the abstract.
Introduction	R2: We elaborated further on our choice for replication.
	R3: We added literature review about trait hope.
	R4: We added a brief explanation of the framing effect.
Methods	R1: We added an explanation of the adjustments about the Asian Disease Problem. We added the plan for conducting an additional data collection if the familiarity check leads to large exclusion. We reported the Cronbach alpha of the measures in the target article and planned for a comparison.
	R3: We added an explanation of the adjustments about the Asian Disease Problem. We added the information about deviation from Study 3 of the social factor.
	R4: We added the criteria for the high-level conclusion of the replication.

Section	Actions taken in the current manuscript
Results	R2 & R3 &R4: We added tables based on simulated data and specified the statistical tests.
	R4: We added the regression tests without trait hope.
Discussion	R2: We added a strategy on how to compare and conclude the two methods of categorizing events.
	R3: We added a future planned discussion regarding using online methods.
Supplementary materials	R1: We added the calculation process of the sensitivity analysis. We adjusted the age-sensitive exclusion criteria to also consider personal choice.
	R3: We added the calculation process of the sensitivity analysis. We specified the reversed items in the measures and the index method for the framing task.

Note. Ed = Editor, R1/R2/R3/R4 = Reviewer 1/2/3/4

Response to Editor: Prof. Chris Chambers

Four reviewers have now completed their evaluations of the Stage 1 manuscript. Overall, I would judge their reviews to be critical but cautiously optimistic -- and highly detailed and constructive. All highlight the value of this replication attempt while also noting a range of areas that would benefit from improvement or clarification in order to meet the Stage 1 criteria.

Across the four sets of evaluations, some of the major issues to address include (in no particular order): strengthening the justification for a replication (see below for my view on this point), considering the pros and cons of a pre-test, including contingent analyses in light of potential failure of statistical assumptions, clarifying conditions under which hypotheses will be deemed unsupported, strengthening the justification of the SESOI, consideration of equivalance testing to provide positive evidence of absence, clarifying the rationale for studying framing effects, confirmation of ethics approval, tackling major methodological deviations from the original study (e.g. such as as the sampling method, and strategies to address it), clarification of study materials to match the original methods, and overall clarity of presentation in key areas.

Considering the the first point above -- strengthening the justification for conducting a replication -- under the PCI RR criteria, replications require no additional justification over and above any other kind of research, and unlike some journals, we do not evaluate Stage 1 RRs on the basis of the perceived importance or value of a research question (but rather, the scientific validity of that question). For this reason, there is no risk of rejection on the grounds that a replication may be unnecessary. That said, I think it would be very much in your interest to address this point constructively in revision because it will make the presentation more compelling for readers, and the reviewer who raised it offers some useful ideas on how to do so.

This list of issues is not exhaustive and you will find that the reviews provide a wide range of additional points. On the basis of their assessments, and my own reading of the manuscript, I am happy to offer the opportunity to submit a major revision.

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit. We've addressed the reviewers' comments and revised accordingly.

Response to Reviewer #1: Dr. Kelly Wolfe

Thank you for your submission of this Stage 1 Registered Report, where you outline your plans to conduct a replication study of Lerner and Keltner (2001), with the aim to replicate the findings that those who are angry tend to seek out risk and make more optimistic judgments (similar to happy individuals' optimistic judgment), while fearful individuals tend to be risk averse and make pessimistic judgments.

I'm pleased to see more replication studies in recent years and have looked forward to reviewing yours. Overall, it looks like a well-thought-out design, with some proposed changes to optimize the study where details of the original study were unavailable or would not suit the proposed work. You have provided much detail that benefits both the reader and those interested in replicating this or another study. In my response, I have included some suggestions and comments that I believe will add to the quality of the work and/or are worth considering before commencing the data collecting of this project. In the top are comments/suggestions that require more consideration, with smaller comments/suggestions below that.

Thank you for the positive opening note, your careful review, and the detailed constructive comments.

1. Under the assumption that you contacted the authors of the original works, the outlined decisions in the manuscript seem reasonable, considering that the original paper does not mention, or make available, many details on its materials (such as the items chosen for aggregated scores).
However, a change in the study's population will provide a different sample from the original study (which had a student sample). Especially considering the COVID-19 pandemic, I see the benefit of using MTurk in terms of the ease (and safety) of collecting data, but I worry that this causes the proposed study to navigate further from its *close* replication effort. I understand that collecting a student sample of 700 is likely not possible due to the current circumstances as well as the replication study being an undergraduate project. Instead, I suggest that the authors consider imposing some further requirements to make the sample more alike to the sample of the original paper, specifically concerning age. This could be in the form of an age range, or a cut-off age (based on median or mean age of US undergraduate students). I did see that participants over 40 years old

> **are excluded at present, but that may still lead to a sample with a higher (mean) age than the original sample.**

We understand your concern about the deviations from the original study caused by a different sample. Our definition of "close replication" is based on the classification by LeBel et al. (2018).

We have conducted many other replication studies using the MTurk samples. Based on our extensive experience, we believe that we have established this sample to be reliable and consistent with other samples.

Specifically, in a number of replications, when we conducted replications on both students samples and online on Mturk, we found the findings consistent across the two samples. For example:

1. Chandrashekar, P., Yeung, S. K., Yau, K. C., Feldman, G., Ma, Y. T., Ho, C. M., Cheung, C. Y., Lui, C. H., Kulbhushan, T., Wong, C. Y., Pillai, T., Thirlwell, N. T., Leung, W. N., Li, Y. T., Tse, C., Cheng, B., & Chan, C. (2020). Agency and self-other asymmetries in perceived bias and shortcomings: Replications of the Bias Blind Spot and extensions linking to free will beliefs. DOI: 10.13140/RG.2.2.19878.16961/4. Retrieved April 21, 2022, from https://www.researchgate.net/publication/331431431_Agency_and_self-other_asymmetries_in_perceived_bias_and_shortcomings_Replications_of_the_Bias_Blind_Spot_and_extensions_linking_to_free_will_beliefs

2. Chen, J., Hui, L. S., Yu, T., Feldman, G., Zeng, S., Ching, T. L., Ng, C. H., Wu, K. W., Yuen, C. M., Lau, T. K., Cheng, B. L., & Ng, K. W. (2021). Foregone Opportunities and Choosing Not to Act: Replications of Inaction Inertia Effect. Social Psychological and Personality Science, 12(3), 333–345. https://doi.org/10.1177/1948550619900570

We consider the move from student samples to MTurk to be an advantage, especially given that the original article did not indicate that age is of any theoretical implications or that the findings are limited to a specific age, or to a specific university, or to a specific time. Whenever we repeat classic findings in a replication, we are running these in different context, time, participants demographics and background, methods of delivery, the experimenters, etc. Age is a rather visible factor, but it is not necessarily the most important one, and we cannot control for all these factors. One of the important aspects of a replication is to be able to examine the robustness of the replication, beyond age, beyond gender composition, beyond method of delivery, beyond time, and so the focus on adjusting a specific salient factor like age limits our ability to execute much needed replications.

To address this point we incorporated the discussion of the differences between the original sample and our sample into the manuscript, making it clear what the deviations are.

In addition, we will be discussing the limitations of the sample difference in drawing inferences from this replication in the Discussion section of the manuscript, following data collection in Stage 2. Please see the update to the discussion section.

Importantly, we share all of our data and code, and so anyone who is specifically interested in running further analyses on a specific included demographic, like age, could run such factors as possible moderators to examine whether there is an interaction.

Lastly, a sidenote: Yes, we did contact both authors on Nov 5, 2021 and one of them kindly responded with the following, wishing us good luck with our project:

> "I'm sorry but I don't have those materials either. We ran the studies 24 years ago, and our department moved buildings…".

2. In the manuscript, you propose altering the Asian Disease Problem (Tversky & Kahneman, 1981) to a generalized Pandemic Problem, using larger numbers in the choice options:
"*We also adjusted the original estimates of people saved or killed by 1000 times, given that under the current COVID pandemic the original estimate of 600 people killed may not be regarded by our participants to be a true pandemic related decision*".
However, I've read some studies that show that the Asian Disease Problem yields similar findings as prior to the pandemic (for example, see
https://www.sciencedirect.com/science/article/pii/S0925753520304628#s005

0). As such, I am wary of changing one of the core materials used in the original study, and would advise not to, unless there's evidence to support the proposed change.

We appreciate the note, and it is important feedback for us that we need to explain a bit more about this adaptation.

This is a very common adaptation to the original scenario. We have used the adapted version of Asian Disease Problem (i.e., 1000 times the figures) in other projects we ran on the Asian Disease before (for example, see Feldman, Wong, and Baumeister, 2016). Furthermore, there are studies using this adaptation by many other researchers such as Dylman and Champoux-Larsson (2020) and Miozzo et al. (2020). We believe the adjustments are both necessary and valid.

To clarify this point, we added this explanation to the Measures section in the manuscript.

References:

- Dylman, A. S., & Champoux-Larsson, M. F. (2020). It's (not) all Greek to me: Boundaries of the foreign language effect. *Cognition*, 196, 104148. https://doi.org/10.1016/j.cognition.2019.104148
- Feldman, G., Wong, K. F. E., & Baumeister, R. F. (2016). Bad is freer than good: Positive–negative asymmetry in attributions of free will. *Consciousness and cognition*, 42, 26-40. https://doi.org/10.1016/j.concog.2016.03.005
- Miozzo, M., Navarrete, E., Ongis, M., Mello, E., Girotto, V., & Peressotti, F. (2020). Foreign language effect in decision-making: How foreign is it?. *Cognition*, 199, 104245. https://doi.org/10.1016/j.cognition.2020.104245

> **3. On page 27, in paragraph 3, you describe the following:**
> *"In addition, we added a question checking participants' familiarity with the framing effect question, and we excluded those indicating previous experience from analyses on that question".*
> **Considering that the Asian Disease Problem is a commonly used method to assess framing effects, it is likely that participants on MTurk will recognize those type of questions from prior studies. My concern would be that having a set sample size of 700 could lead to you having to remove the data of many participants, resulting in a smaller sample size than expected. Instead, I suggest specifying that you will collect more data if you find that the required exclusions lead to a sample size smaller than you require. If I have misunderstood (i.e., you meant that the sample size is n = 700, after data subject to exclusion has been removed), please let me know, and clarify this in the manuscript.**

Thank you, this is a good point and a great suggestion. We did not expect large exclusions based on our previous replications asking participants regarding their familiarity with our target common judgment and decision making paradigms. That said, we agree that we should plan ahead for this possibility.

Our decision of sample size was determined by a combination of expected effect size, funding budget, and the sample size of the original article.

We added a statement in the Outliers and exclusions section in the manuscript, that if exclusions are larger than 20% (140 of 700) and we failed to find support for the core hypotheses, that we will embark on an additional data collection to meet the 700 after exclusions based on the exclusion ratio we observed and analyze and report the data together.

> **4. p.12, paragraph 2.**
> *"Study 2 showed that the more fearful and happy tend to be more optimistic compared, whereas the angrier tend to be more pessimistic."*
> **Fearful and angrier should be reversed in this section; in the original paper, it was found that angry was associated with optimism instead of fear.**

Thank you for catching that and pointing it out.

We corrected it by reversing "fearful" and "angrier":

> "Study 2 showed that the angrier and happier tended to be more optimistic, whereas the more fearful tended to be more pessimistic.

> **5. p.16, paragraph 1. You discuss the calculated power in this section and mention having done a sensitivity analysis. Please specify the method you used to calculate this (e.g., GPower, data simulation with pilot data, et cetera).**

Thank you for this suggestion. We added the GPower calculations screen captures in the "Power analysis of original study effect to assess required sample for replication" section in the supplementary.

> **6. p.16, paragraph 1. The sample size for this study has been increased compared to the original study, but it is unclear what informed this sample size. For example, did you include an expected percentage of excluded data based on your pilot study (e.g., say, 10% of data in the pilot was excluded so the same percentage is applied in the calculation of the study's sample size), or are you applying a standard additional percentage to your sample size calculation (not informed by your pilot, but perhaps by prior experience)? Please include this in the manuscript.**

Thank you, that is a good point. We added 70 participants to the planned data collection, to allow for 10% exclusions, and left the rest as is.

We added a brief clarification on this point in the power analysis section in the manuscript:

> We instead aimed for a sample size of 770 participants taking into account 10% exclusion with a final sample of 700, aiming to go beyond the largest study in the original (Study 2 had 601 participants) to allow for exclusions and additional analyses. Our sensitivity analysis indicated this sample would allow for the detection of $r = .12$ and Cohen's $f = 0.06$ for the interaction (4 predictors, 2 groups; both 95% power, alpha = 5%, one-tail) and considered weak effects in social psychology (Lovakov & Agadullina, 2021), and therefore reasonable as our Smallest Effect Size of Interest (SESOI).

> **7. p.21. Please report the reliability of the included measures (based on prior research), where possible.**

Good suggestion.

As the original article did not specify the items used for each measure, we re-selected the items for most measures. Therefore, the reliability calculated in the original article may not apply here.

We will calculate the Cronbach's alpha for each measure after the data collection, and we will compare this to the original's.

To address this point, we added the reporting of the target's Cronbach alpha, and aimed to briefly compare and discuss reliability of the measures. See Table 5 in the manuscript.

> **8. Supplementary materials, p.15, specific criteria 3.**
> *"Participants who reported they were aged 40 or above. The responses for the item "having a mentally gifted child" in the optimism measure will be excluded (Wong et al., 2019)."*
> **This is unclear to me. Can you explain how the age exclusion relates to this item?**

We set this exclusion criterion based on a former successful replication by our team of Weinstein (1980). Having a mentally gifted child was regarded as an age-sensitive and parenting-choice related event in the replication (Wong et al., 2019), thus excluded from participants over 40 years old and those who do not plan to be parents.

We also adjusted the demographic section of the Qualtrics to include the following questions:

> 1) "Are you currently a parent? (*Yes*/*No*)" and if they answer "*No*" then display the question "You indicated you are not currently a parent - Do you plan to become a parent at some point in the future? (*Yes*/*No*/*Not sure*)"

The responses to the item "having a mentally gifted child" will be excluded for participants who are a) over 40 years old; b) currently a parent; OR c) not planning to become a parent in the future.

> 2) After the marital question ("What is your current marital status?"), for participants who choose "*Not married*" display the question "Do you plan to get married at some point in the future?" (*Yes*/*No*/*Not sure*)

The responses to the item "marrying someone wealthy" will be excluded for participants who are a) Married or engaged; OR b) Not planning to get married in the future.

Response to Reviewer #2: Dr. Max Primbs

The authors, Lu and Feldman, present a Stage 1 Registered Replication Report for the foundational Lerner & Keltner paper by 2001., which investigated associations between dispositional fear, anger, happiness and risk optimism or risk preference. They adjust the original design to fit for an online study and propose to combine studies 1-3 of Lerner & Keltner into one design. Furthermore, they extend the original paper by adding hope as additional dispositional emotion.

Before coming to the review part I want to thank the authors for their efforts to increase the quality of psychological research and want to laud the general research programme by Dr. Feldman aimed at replicating foundational work in judgement and decision making research. I also want to laud the fact that the authors decided to submit their work as a registered report and that they recognize that direct replications need not repeat mistakes of original studies.

In summary, I recommend acceptance with revisions. I will detail major points to be addressed first and minor points and stylistic issues to be addressed after. Briefly put, the authors need to a) better justify some of the choices they make regarding study selection, sample size justification, and measure selection and b) make it clearer how possible findings will be interpreted.

Thank you for the very positive opening note, and for the constructive helpful feedback.

Major points:

1. The authors argue for their choice study by stating that there are a) no direct replications, b) they can improve on the methodology used by the original authors and c) the impact of the original study. While I agree that the original study was quite impactful and am also not aware of any direct replications, I`d ask the authors to provide stronger justification for the need of a direct replication.

There are three reasons for this: Firstly, there are many impactful papers which have not yet been directly replicated. Secondly, the original studies were for the time quite well powered, particularly Study 2. Thirdly, there are likely several conceptual replications and extensions of their work among those 4211 citations. Impact and the lack of a direct replication are not sufficient justification for why a direct replication is necessary. One potential way to justify the importance would be highlighting the theoretical implications of the different potential findings. I`d recommend

> **the authors to check Isager et al., 2021**
> **(https://psycnet.apa.org/record/2022-14587-001) for some additional**
> **guidance and discussion. Relatedly, the authors write that there are**
> **practical implications for other domains without spelling out what these**
> **implications are – I`d recommend that the authors are more**
> **straightforward about what those implications are.**

Thank you for the comments. We appreciate the nudge to help us elaborate more on our choice of the target study.

As direct replications are important to respond to the credibility crisis in the field of psychology, we believe that an article about an impactful theoretical framework with over 4000 citations is undoubtedly a good target of replication.

Per your points:

1. We agree that there are many other targets in the literature in need of replications, yet we do not think this should reflect on the importance of pursuing and/or publishing this specific replication.

2. The Isager et al., 2021 criteria might be a starting point for a discussion on how to approach the question or optimizing resources for replications or selecting among several targets, but we need to be very careful in applying this as a criteria in evaluating replications and deciding whether to accept/reject or proceed with a replication.
   We have had discussions with that team and we hold somewhat diverging views on the actual value formula and especially about applying this formula to evaluating suggested replications.
   We reflect on this article and the issues in applying it in the review process in a talk "Rethinking and discussing replication targets" (https://osf.io/znh85/; DOI 10.17605/OSF.IO/ZNH85).
   We appreciate the PCIRR the recommender Dr. Chris Chambers clarification that:
   "Considering the the first point above -- strengthening the justification for conducting a replication -- under the PCI RR criteria, replications require no additional justification over and above any other kind of research, and unlike some journals, we do not evaluate Stage 1 RRs on the basis of the perceived importance or value of a research question (but rather, the scientific validity of that question)."

3. Conceptual replications are perhaps indicative of interest, yet serve a very different purpose from direct replications, and we believe should be interpreted with caution, unless they are well-powered rigorous Registered Reports meeting best practices of open-science. There are plenty of case studies with very large bodies of literature with many conceptual replications that have failed to replicate (e.g., "social" priming research and Ego Depletion). We also reflect on this point in the talk "Rethinking and discussing replication targets" (https://osf.io/znh85/; DOI 10.17605/OSF.IO/ZNH85).

We took your point as an opportunity to elaborate further on our choice for replication and the importance of replicating this article. We also added a bit more on the practical implications.

> **2. Table 1 details the hypotheses but does not explain how the different potential findings would be interpreted in light of theory being evaluated. The table at the very beginning of the manuscript does a better job, at least mentioning for which theories these findings could be relevant.**
> **I`d ask the authors to add a table that details A) The Hypothesis B) The Analysis being used, e.g., as a regression equation and C) The interpretation of the different potential results in light of the theories being evaluated. It is important to detail what a significant finding or a non-significant finding will mean in terms of theory.**

Thank you for the suggestion on better linking the hypotheses, methods, and analyses, these are valuable. We adjusted the PCIRR study design table at the very beginning of the manuscript to sum up the hypotheses, analysis performed in the target article, our planned analyses, and brief description of the change.

As a sidenote, we feel as if the column "Theory that could be shown wrong by the outcomes" is a bit problematic for replications, and are hesitant to draw any strong conclusions from both a successful or failed replication, especially regarding the broader literature and the grander theories. Replications should be views as updating our effect-size estimates regarding a specific phenomenon using a specific context/design, and we should be very careful and humble about using this replication to inform our knowledge about more than that.

> **3. On Page 14/15 you write "In our main replication analyses, we aimed to use the categorization of the target article as is, yet to improve on the methods of the original we also opted to directly assess participants' perceived controllability and certainty of the events. With these measurements we sought to revisit the ambiguity categorization of the target article and also conduct analyses of ambiguity as the two continuous measures rather than a dichotomy of an aggregate."**
> **If I understand this correctly, the authors will use both the operationalisation of the original authors and their own improved version. What will they do if results of both operationalisations do not converge?**

Good point, thank you, we agree that this clarification is needed.

Our replication is focused on replicating what the original did, and we will therefore evaluate the replication success based on the target's methods. Regardless, we will supplement this with our own analyses based on the participants' categorization. If we fail to find support for the target's using their methods, yet succeed to find support using our methods, we will conclude this as an update to the target and/or a need to reframe the target's theory and conclusions.

We added a plan to discuss and directly compare the two methods in the discussion section, with brief strategy on how to conclude the various scenarios:

- Both failed to find support: Failed replication.
- Original's methods failed to find support and ours found support: an update to the target and/or a need to reframe the target's theory and conclusions.
- Original's methods succeeded in finding support and ours failed to find support: Successful replication, yet robustness challenge.
- Both found support: Robust phenomenon.

> **4. On Page 16 the authors explain that "Our sensitivity analysis indicated this sample would allow for the detection of r = .12 and Cohen's f2= 0.015 (4 predictors; both 95% power, alpha = 5%, one-tail), considered weak effects in social psychology (Lovakov & Agadullina, 2021), and therefore reasonable as our Smallest Effect Size of Interest (SESOI)".**
> **In my opinion, such a justification is not sufficient, because effect sizes cannot be interpreted out of contexts and claims that effects matter must be accompanied by empirical evidence for this claim. By determining a SESOI, the authors are effectively arguing that this is the smallest effect size that matters. My question would be why? One way to make an argument would be to explain how a particular effect size corresponds to real life outcomes. For example, if there is a correlation of *r* = .1 between anger and risk preference, what does that mean for the likelihood to have a car crash? One could also think about what effect sizes would be meaningful for or in light of the theory. The authors can find some guidance and discussion in Primbs et al., 2022 (Note: My own work), Anvari et al., 2022, and Anvari & Lakens, 2021.**

We found it difficult to understand whether the bottom line we should take from this comment is that you think we are too well-powered or not powered enough, or maybe a completely different point entirely. Clearer guidelines and constructive examples would have been helpful. We looked at the citations and we did not find a way to tie those into our replication effort, nor were we able to find any instances of replications making use of this paradigm.

We can see the value in such an approach, and we understand the need for a debate on the determination of SESOI, yet we do not think that such a discussion would add much here, given that our aim here is to replicate and repeat what has been done before. The justification in replications is to conduct a fair test to detect the original's findings, and we believe that we have met that criteria, in allowing the detection of what are considered weak effects in the literature.

In this specific context, we also feel that we need to be very careful when making claims about how effect sizes in a study like this may correspond to real life outcomes, and this goes far beyond the scope of what a replication is aimed to cover.

We are happy to address this comment further given clear editorial guidelines, and we would ask for citations and examples from the literature, preferably from published Replication Registered Reports, of how this was applied for replications.

> **5. For several of the trait questionnaires on page 21ff you write that you were unable to determine the exact items used. I`d ask to contact the original authors and check if they can share this information (if you did not already do this of course). Jennifer Lerner is still active: https://www.hks.harvard.edu/faculty/jennifer-lerner**

Thank you for the suggestion. We have tried to contact the original authors (see reply to Reviewer 1 above) but were unable to obtain the materials or any information about the study.

We made adjustments based on our understanding of the study and what we think would be the most appropriate inclusive test.

> **6. I believe the authors miss a strategy for analysing and interpreting non-significant results. I`d recommend to check out equivalence testing – if the aim is to replicate a finding, it is paramount that also null results can be interpreted properly.**

Thank you, we are aware of equivalence testing and of Bayesian analyses to quantify support for the null, and these can be helpful when we aim to interpret findings regarding null hypotheses. It was not clear to us what specifically you were suggesting the equivalence testing here for, given that we do not have null hypotheses, and are mostly focused on aiming to reject the null for finding support for the original's.

We will regard this direction as exploratory, and we will be making all our data and code available for anyone who would be interested in performing additional analyses.

> **Minor points:**

> **1. The authors wrote on page 10 that there is no direct replication and state on page 11 that they will conduct a close replication. I later saw your Table detailing how they arrive at the fact that this is a close replication, but for the reader on page 11 this is a confusing difference.**

Thank you for pointing this out. We added information regarding the information of defining close replication in page 11.

2. Lerner & Keltner used American Undergraduate Students. The authors seem to use a different sample: Please justify why these samples are comparable.

Thank you. Other reviewers have also raised this concern. Please see our reply to Reviewer #1's Comment 1.

3. Page 17: "We also employed the Qualtrics fraud and spam prevention measures: reCAPTCHA, prevent multiple submission, prevent ballot stuffing, bot detection, security scanmonitor, relevantID, etc.". Please name all measures you use.

Thank you. We specified all the measures we used. As indicated, we will be running pretests to ensure all goes smoothly, and if not - we will make adjustments. We will update this after data collection should anything be changed or added.

4. I applaud the authors for ensuring a fair payment of participants with the procedures lined out on page 17.

Thank you for your kind words.

5. Table 4 is difficult to understand and contains information that is presented elsewhere in the manuscript. I`d suggest removing or restructuring the table. For example, the item phrasing is included only for some of the variables but not for all, and is partially discussed again later on. I think you could remove those.

We were not sure what about Table 4 was difficult to understand, we found it to be helpful in summarizing the study design for readers in one clear table. All the measures were detailed, though we did not include the specific items, given that these are detailed in Table 12. In addition, all materials were provided in the supplementary and in the attached survey.

We added a note to inform the readers these are sample items and where to find the full measures: "Please refer to Table 12 for the full list and categorization of the events."

To further improve on Table 4 we added in the DV box:

"(The dependent variables were measured for each of the IV2 ambiguity 23 items:)"

And we indicated how many items were in each of the IV2 ambiguity categories.

6. On Page 21 the authors write: "Given the high Pearson correlation reported in the original between the two scales (r = .54), we followed the original's method in combining the two scales into an aggregate score". What do the authors plan to do if they do not observe such a high correlation in their sample?

Thank you for the comment.

We added:

If a high correlation ($r >= 0.3$) is not observed, we will still follow the target article's method, but conduct a correlation analysis for the two scales separately.

7. I`d ask the authors to include regression equations in the manuscript. They remove all uncertainty about the analyses you will conduct. For example, now I`m not clear whether you will conduct separate regressions for each dispositional emotion or combine the IVs in the same analysis (see also Major Point 2)

Thank you for this suggestion and helping us to improve on this point.

Regarding the regression equations: Our understanding is that this is not common in our literature, yet we note that there should be no uncertainty about the analyses we will conduct given that we added all the code that we aim to run.

Regarding clarifying our analyses: To address this and other comments received by the reviewers, we focus our analyses on simple correlations, and added a very detailed correlations table in Table 8. Regressions are meant to examine all predictors together in the same model, so in the regression analyses we will be running all three predictors in the same model, first with three predictions (replication) and then with four predictions (hope extension).

We added tables based on the simulated data in the Results section and specified the predictors for each table. We believe that the tables now address all the information readers need in order to interpret our findings and how they relate to the hypotheses.

We note that we will ofcourse share all the analyses with the outputs to allow anyone to obtain whatever else information they might require regarding our analyses.

8. On Page 28 right before the last "insert table" is a free-flying "N"

Thank you for the careful review. We removed it.

Response to Reviewer #3: Dr. Agata Sobków

> **The Stage 1 manuscript describes the proposal of a study replicating effects described by Lerner and Keltner (2001).**

> **I found the study interesting and worth to be conducted. Moreover, the authors do their best to follow open science practices and perform an informative replication.**

Thank you for the positive opening note, and for your constructive feedback.

> **Nevertheless, in my opinion, in the current form, the proposal suffers several limitations that should be clarified.**

> **1. General:**

> **a. The submission was difficult to follow and, in my opinion, was a technical merge of different parts rather than a comprehensive and consistent text with a story the authors wanted to tell. If the manuscript contains so many tables and disclosures spread through the text, the reader could miss crucial details about the research problem, method, etc. I had first to read the original article by Lerner and Keltner (2001) thoroughly to understand the design of a proposed study. Of course, making science is very technical, but without these soft and narrative aspects, the manuscript would be less accessible to readers.**

We appreciate this feedback, we would have liked to do something about this to improve, yet you did not provide us with any specific constructive advice on how to improve, other than to tell us to broadly do better on something we thought was laid out well and follows the structure of our other replication projects. The right way to structure Replication Registered Reports is a matter of subjective taste, and our strategy was to emphasize transparency and comprehensiveness over narrative, which was provided by the target article and other articles in the follow-up literature.

We would be glad to follow up and improve further given clear editorial guidelines. We would then appreciate guidelines on what the issues are and how you would expect us to improve, preferably with specific examples.

b. I like the idea of extending the original study on trait hope. Indeed it complements the design in the aspects of valence and controllability/certainty.
Nevertheless, I recommend the authors elaborate more on this emotion and how it was linked to risk perception/preference in the previous research.

Thank you for the suggestion.

We added more and elaborated further on dispositional hope. We added a broad definition of hope and some initial research on hope and risk perception/preference.

Please see subsection "Extension: Dispositional hope" in the revised manuscript.

c. The manuscript often refers to the supplement, and the supplement itself is very extensive (20 pages). I think that it would be good to add in the manuscript the precise location in this supplement (e.g., "detailed description could be found in the supplementary Table SX, p. X").

Thank you, great suggestion.

We added the precise location (subsections) in the supplementary materials when referred to from the manuscript.

2. Ethics

a. Whether the authors received the agreement from their ethics committee? I can't find this information in the manuscript (nevertheless, the HKU ethics committee is mentioned in the Qualtrcs survey). I also think that participants should be informed that their anonymized data will be publically shared.

Thank you. Yes, to our understanding this is a prerequisite in PCIRR submissions. We appreciate your suggestions on doing better with our reporting on this issue.

We obtained an ethics approval EA210265 from the University of Hong Kong Human Research Ethics Committee. We added the information in the Pre-registration and open science section of the manuscript:

"The project is part of a larger replications project that received ethical approval from the University of Hong Kong Human Research Ethics Committee (EA210265)."

We also uploaded the approval to the OSF: https://osf.io/mrfk7/

Also, we added an explicit paragraph in the consent form a section about the public availability of anonymized data:

> "Deidentified/Anonymized version of the data with no confidential data will be shared publicly to allow the reproduction and reproducibility of our research."

> **3. Method:**

> **a. I generally like the idea of merging three original studies into one, but I have some concerns about a new modified framing task. I think that, before the main replication study, authors should conduct a pilot one testing whether, in this particular form of a task (e.g., within-subject, different name, different numbers in the description), the classic framing effect would be found.**

We are unsure how to address this concern, since you did not provide any details regarding what the concern is. If this refers to our adjustment to number modification of the framing task, then please see our response to Reviewer 1's point 2. In general, the so-called "Asian disease" problem has been demonstrated and showed generalizability across many study designs and with various adjustments. Examples:

> Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian Disease problem: Time constraint, need and disease type. *Judgment and decision making*, 13(6), 529. Retrieved from: http://sjdm.org/journal/18/18415c/jdm18415c.pdf

> **b. Based on a description, it is hard to find how the authors plan to deal with ambiguous and unambiguous events, e.g., how they would be scored based on certainty/controllability. Moreover, in some parts, the authors write about this variable as continuous and as dichotomous in others (e.g., in Table 4 or in hypotheses in PCIRR-Study Design Table).**

Thank you. We realized we could do better in explaining this.

There are two parts, the replication following the original's methods and their categorization, which was a clear dichotomy, and then the extensions that we added, which involved continuous measures.

Our assessment of the replication success would follow the original's dichotomy and a comparison to their analyses, and as an extension we added continuous measures, which we will analyze against all measures and compare to the dichotomy in the target and the analyses based on that dichotomy.

We clarified our replication success/failure criteria in the discussion section.

We Also made it clearer throughout the manuscript about which categorization we were referring to.

> **c. Which 4 predictors were considered in power analysis? Dispositional emotions? Please be more precise. How about ambiguity?**

Thank you for the question. We made the description of power analysis in the manuscript more precise, as well as adding the calculation process in the "Power analysis of original study effect to assess required sample for replication" section in the supplementary.

To answer your specific question: Yes, we were referring to the four dispositional emotions and the two conditions (groups).

> **d. I think that Table 3 is problematics in some aspects, e.g.,**
> **- It is not clear what "year" means. In the case of Lerner and Keltner, it is the date of publication and not data collection (e.g., the manuscript was submitted on "October 4, 2000" and probably the data were collected even earlier)**
> **- Some values from the simulated dataset do not make sense e.g., authors would not receive an age range from 0-100 as they only accept 18+ years old participants.**
> **- There is also one critical difference between Study 3 and the proposed design that should be stressed: in Lerner and Keltner's study 3, participants responded to items in a face-to-face interview. The original authors thought that "having participants respond orally rather than in an anonymous self-report might reduce the tendency for happy and angry individuals to see themselves as comparatively less vulnerable to negative life events." (p. 152).**

Thank you for the detailed feedback about Table 3. We added a note to the table that "year" reflects the time of publication.

The simulated dataset was only meant as a simulation for what the results would look like after data collection, and was not meant to be realistic given that age is not a factor in any of our analyses.

We added the information regarding the target's Study 3, this is an important detail. Given our unified combined design conducted online, we will not be addressing the in-person factors, the thought shared by the authors was not tested, and our replication will inform us on the robustness of this design to anonymized online answering.

We added details regarding this point in our deviations section and in our future planned discussion limitations section in the manuscript.

e. Please add sample items in the manuscript when presenting used methods.

We had included sample items in Table 4.

4. Results:
a. The description of a result section is inconsistent. In some parts, there is a table with simulated data. In other, the authors describe that "[The regression tables of these analyses inserted here]" or just free space, e.g., "Third, we regressed the optimism estimates on the emotion dispositions separately for ambiguous and unambiguous events. For ambiguous events, we … For unambiguous events…". Such description is difficult to follow and be adequately understood.

Thank you for nudging us to do better in our planned data analysis.

We previous provided all our planned analysis in our code output, and we now more clearly refer to our planned data analysis code for full details at the beginning of the results section:

[Our data analysis will follow the data analysis code provided with the simulated data. Please see our OSF directory for Rmarkdown code and outputs. Below, we briefly summarize our data analysis plans, followed by a brief outline of the structure of the results section, which will be updated after data collection.]

Given that all our data is random noise, adding numbers to the tables does not seem to add much at this stage, and was simply meant as a demonstration of what results would look like after data collection.

In additions, we revised the results section to simplify our reporting of the analyses conducted, all summarized in the correlations and regressions Tables 9, 10, and 11.

b. Please plan a table summarizing descriptive statistics for framing task, optimism, controllability/uncertainty, etc.

Thank you for the suggestion. We added these descriptive statistics into the newly added correlations table, Table 9.

c. Please plan a comparison of whether the same situations as in the original one would be classified as ambiguous or unambiguous.

We included it in the "Exploratory analyses" in the results section.

We also added the comparison table to the "Exploratory analyses" section, see Table 12. We also use this table to write out the items, and refer to this table throughout the manuscript.

5. Supplementary materials:

a. It would be worth adding information about reverse coding when presenting items/scales in the supplementary materials.

Thank you, great suggestion. We added information about which items are reversed in the supplementary and the newly added Table 12 (marked as (R)).

b. It is also not clear how the responses in a framing task will be indexed.

We had included in the supplementary that the response would be indexed as 1 to 6. We added information in the Instructions and experimental material section in the supplementary about how the score of risk preference is calculated:

> "The risk-seeking score is calculated by summing the scores for the gain frame and the loss frame."

6. Other:

a. The authors could consider adding the following keywords: "risk preference", "hope", "optimism"

Thank you for the advice. We added these keywords to the abstract.

b. Table 1. Why "trait anger" is written in italics?

Thank you for the careful review. We removed the italics.

Response to Reviewer #4: Dr. Karolina Scigala

> **I really enjoyed reading this Registered Report. It aims to replicate three interesting and influential studies, it was written very clearly, and the methods, as well as the statistical analyses were well chosen. The authors made sure that their Registered Report is very transparent, which I applaud. Overall, I think it will be an important contribution to the field.**

Thank you for the very positive opening note, and for your comments and suggestions.

> **I do have several comments, which I outline below:**
> **Introduction**
> **1.   In the abstract, the authors state:**
> **"*Our replication [failed to find/found] support for the original findings regarding associations between dispositional emotions and two risk-relevant measures: risk preference and risk optimism [summary effects sizes and CIs].*"**
> **I suggest that the authors clarify how exactly they will decide whether they failed to find support or found support for the original findings. For instance, how many of the hypotheses need to be (dis)confirmed to state that support was (not) found?**

Excellent suggestion, thank you for nudging us to do better here. Yes, this is a complex issue, and there are no clear guidelines for how to summarize complex replications.

Broadly, our replication merged three studies of the original article and so we planned to summarize for each study how many of the hypotheses in that study were supported (with effect sizes). This would allow nuanced understanding of the findings.

We also added a clarification for the high-level conclusions that our summary would be as follows (see Table 7 in the manuscript):

- Fully successful replication: Finding support for all studies.
- Mostly successful replication: Successful replication (all hypotheses supported) of two of the three studies.
- Mixed findings: One study out of the three or two studies out of the three with some but not all hypotheses supported.
- Failed replication: Failure to find support for all three studies.

2.   The role of framing effects (gain/loss) in this replication is unclear. There is no mention of framing effects in the Introduction, and it only appears in the Methods section. The authors should clarify what is the role of framing effects in the replication.

Thank you for this feedback. We added a brief explanation of the framing effect in the "Overview of replication and extension" section in the revised manuscript.

3.   Related to point #2, it would be helpful if the authors could clarify briefly which results presented the original paper they decided not to replicate in the current replication, and clarify why that is the case.

We had stated in the "Overview of replication and extension" in page 11 about why we wouldn't replicate Study 4:

"Study 4 involved inducing complex emotions in participants to try and determine causality, which we felt was more appropriate for a follow-up study after reconfirming the associations in Studies 1 to 3 with a replication, and then preferably executed in a well-controlled lab setting with careful attention to possible impact on the participants."

Methods
4.   The authors state that they decided not to conduct a pretest measuring the ambiguity of events, which might be problematic. A sample of MTurk workers in 2022 quite likely differs in many ways from the original sample of undergraduate students in 2001, which makes it probable that the former group will have a different perception of event ambiguity than the latter. I understand that if the authors decide to conduct a pre-test and then choose the (un)ambiguous events based on its' results, the final procedure might differ substantially from the original one. However, by not conducting the pre-test, the authors risk that participants' event ambiguity ratings will be different than in the original study, and as a result, the data might be not usable. Summarizing, I would recommend that the authors conduct a pre-test.

We specifically addressed this concern with adding our extension, embedding the pre-test within the same study design. We used a unified design for the three studies and the pre-test, which we consider a major strength that would allow us to ensure that we know the ambiguity evaluations of the same participants who we measure on optimism. If we conduct the pre-test separately, and things do not work out in the main examination, we have no way of knowing what was the reason for the failure, whether it was that their ambiguity ratings diverged from that of the pre-tests or because this is a failed replication. This way, we can assess both to draw stronger inferences.

We agree, our sample's categorization might diverge from that of the original's sample, and we will therefore be comparing our results to that of the original categorization, and given the combined design supplementing with advanced analyses examining associations between the variables and perceived ambiguity.

> **5.   In Table 3, the authors state that some descriptive statistics were not reported in the original article. Next, on page 20, the authors state that "The specific items chosen and sued from each scale were not reported in the original article, and we therefore used the scales as is based on reported items in other studies."**
> **I assume that the authors contacted the authors of the original article to clarify these issues, and did not receive a reply (?) I think it would be good if the authors clarified that.**

Yes, we reached out to the original authors, please see our reply to Reviewer 1's first comment.

We would rather not include that information in the manuscript. We do not want to put anyone on the spot or be perceived as making any statements regarding the original authors' involvement. We would have acknowledged them if they would have helped, yet we would rather not mention anything given that we were unable to secure any materials from them.

> **6.   The authors state:**
> **"*Even attentive participants may occasionally answer some attention checks wrong, and we therefore will exclude participants who answered three or more of the five attention checks incorrectly*".**
> **Participants who answer three or more attention checks incorrectly are likely very inattentive. Hence, I would suggest that the authors use a more conservative threshold and exclude participants who fail at least one attention check.**

This is a very subjective choice, we agree, and there are reasons for either decision.

We gave this much thought, and based on our experience with MTurk and student samples we felt that excluding based on a single inaccuracy is far too strict. Any participant may lose focus or answer inattentively occasionally, yet still contribute valuable data to the survey.

Mturkers are professional survey takers, and in our replications they have generally shown to be very attentive. We consider one or two a possible mistake, with three being more of a pattern. Three will catch the automated random responding, which is what we are aiming to catch here.

We added a brief clarification on that in the Outliers and exclusions section in the manuscript.

Results

7.   I suggest that the authors prepare an alternative analysis plan in case their data do not meet the statistical assumptions needed to conduct parametric tests. Such alternative plan could include non-parametric tests or data transformations.

Thank you for this suggestion. We added a correlation table in the result section that included both the Pearson's r and the Spearman's Rho. Spearman's Rho will serve as the alternative analysis in case the statistical assumptions are violated.

8.   I suggest that the authors report replication-related findings without including hope in the regression models, and then report extension-related findings with hope included in the regression models. The reason for that is that hope was not included in the original study, and therefore including it in the replication-related regression models will make the replication (unnecessarily) less alike the original study.

Thank you for the suggestion. We adjusted the planned data analyses to include both regression models with and without dispositional hope. Please see Tables 10 and 11.

9.   In case the authors do not replicate the original findings, they should pre-register analyses that will allow them to conclude that there were indeed no effects, e.g., Bayesian models or equivalence testing.

Thank you for the advice. Please refer to our reply to Reviewer 2's Major point 6.

10.   It would be helpful for the reader if the authors clarify in the manuscript which predictors were included in each model.

We revised our Results section and added tables based on our simulated data to make the analysis plan clearer.

Additional updates

For the risk optimism measure and the corresponding ambiguity measure, we removed the item "Having to take an unattractive job". It was mistakenly regarded as the modified item of "I could not find a job for 6 months", which was actually considered as unfit for the MTurk sample and removed in the Wong et al. (2019, see https://osf.io/52mes/ , page 52)'s replication of Weinstein (1980).

We updated the Qualtrics survey, manuscript, supplementary, and the R code accordingly. We also included a table in the supplementary to summarize the adjustments for the items in the risk optimism measure (see Table 4).