# Reply to PCIRR decision letter #395:
# Hsee and Kunreuther (2000) 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 the recommender, 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/aRqDNkopOjHf**

**A track-changes manuscript is provided with the file: "PCIRR-S1-RNR-Hsee-Kunreuther-2000-replication-extension-mainmanuscript-trackchanges.docx" (https://osf.io/4nvtp)**

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: We revised the main manuscript and the supplementary carefully.<br>R3: We moved materials that are important for readers' interpretation of our paper from supplementary to main manuscript. |
| Introduction | Ed: We clarified the conceptual terms.<br>R3: We defined and clarified the concept "affect", "emotional attachment" and provided background for links between affect and insurance decision-making. |
| Methods | Ed: We addressed the concerns on the study design.<br>R1: We changed the stimulus materials of the camera scenario, included the criteria of replication closeness evaluation of our paper, addressed the four-in-one approach, and stated the criteria of the overall replication evaluation.<br>R2: We addressed the sample size issue, modified the inclusion of MTurk participants.<br>R3: We added a manipulation check and related analyses, applied outcome-neutral tests for DVs, and improved on the data analysis strategy. |
| Supplementary materials | R2: We corrected oversights and moved the scenario tables to the main manuscript. |

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

[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: Prof. Chris Chambers

> **I have now received three reviews of your Stage 1 submission. Overall, the reviews are encouraging and suggest that the manuscript will be suitable for Stage 1 IPA following a careful and comprehensive round of revision. The main issues raised by the reviewers are clarification of conceptual terms and a number of design considerations, including potential effects of demand characteristics, inclusion of positive controls/outcome neutral tests, adequacy of both the statistical sampling plan and inferential analysis plan, and the combination of multiple studies in a single unified data collection. All of the issues raised fall within the normal scope of a Stage 1 evaluation, so I am happy to invite a revision and response.**

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

# Reply to Reviewer #1: Dr./Prof. Rima-Maria Rahal

> **Law & Feldman propose a replication and extension of Hsee & Kunreuther (2000), to test if affection towards an item boost insurance purchasing and claim decision-making. The outlined hypotheses are testable, and will speak clearly to the claim that affect towards goods matters for insurance decisions.**
>
> **Materials:**
>
> **I found the submitted materials well prepared and very thorough, so that further replications based on these materials would be possible. Decisions about handling the materials are clearly outlined.**

Thank you for your feedback and the positive supportive opening note.

> **Nevertheless, two things stood out to me that I would find debatable: First, I would take issue with the decision to remove the sentence "The whole process will take 4 hours." from the stimulus materials. I would guess that this sentence was included in the original materials to obtain tighter experimental control over participants' beliefs of how cumbersome the process of filing an insurance claim would be. As such, this sentence would be helpful to reduce noise in the data (e.g., some participants estimating that they would have to drive for hours to reach the office of the company, while others assuming that the office is right around the corner).**

Thank you for raising this concern.

We appreciate the feedback and agree that the inclusion of the sentence "The whole process will take 4 hours" in the materials was aimed to standardize participant understanding and may help reduce data noise. We initially changed this in order to make the scenarios more comparable, but we see the value in keeping things as is, and then examining differences between that Study 2 camera scenario and the other scenarios.

We switched it back to include "The whole process will take 4 hours" across all conditions for the Study 2 camera scenario. Much appreciated!

> **Second, I am unsure about the four-in-one approach, where participants are shown all four sub-studies within subjects. It seems to me like it would be easy to guess the treatment variations at play, which may weaken the interpretability of the results. This is particularly the case because the decisions are hypothetical (as in the original paper, so this point itself is not a criticism of the replication attempt). A more conservative approach would be to present the materials as one-shot decisions, or to include a follow-up analysis that tests if there are differences in the effects elicited from the first decision and from the subsequent decisions where treatment variations may have become apparent.**

Thank you for raising this concern. We have successfully implemented this many times in our replications, and have received this feedback many times. In all those at the end we were able to convincingly show why this approach was beneficial and important.

We consider this design to have major advantages, going beyond the original's. We would want to know whether there would be carry-on effects and an impact of order combining several studies.

A unified study design embeds the original's separate studies, for the first study displayed to participants, but goes beyond that in allowing for additional insights by performing additional exploratory analyses either only examining the first displayed (which would mirror the original's) or with order as a moderator of the different effects.

In addition, most importantly, this helps address concerns regarding the sample and attentiveness. When we have some failed replication studies and some successful replication studies from the same article, then in a separate design one may raise concerns that the failed experiments were due to sample/time/context, yet with a single unified design, that concern is addressed with the much more likely explanation that the failed replication are because of the differences between the studies.

There are many examples, but we will give one recent example that just completed a PCIRR Stage 2:

> Petrov, N., Chan, Y., Lau, C., Kwok, T., Chow, L., Lo, W., Song, W., & Feldman, G. (2023). Comparing time versus money in sunk cost effects: Replication Registered Report of Soman (2001). [PCIRR Stage 2 recommendation/Open peer review] [PCIRR Stage 1 recommendation/Open peer review] [Preprint] [Open materials/data/code]

In this project, we conducted direct replications of Studies 1 and 2 and a conceptual replication of Study 5 in the article Soman (2001) claiming that money sunk costs are larger than time sunk

costs. We used a similar unified design running the three in a single unified data collection, with the order of Studies 1 and 2 randomized. The final result was that Study 1 was successfully replicated, whereas in Study 2 there were sunk cost effects for both time and money, yet no differences between the two, which we summarized as a failed replication. We conducted order effect analyses and analyzed the data from the studies in which the study was displayed first, and across all these analyses the results were very similar and consistent.

In the past when we ran these separately, editors and reviewers would ask us to rerun the failed replication, with various post-hoc claims regarding the reason having to do with the sample or time/context. However, in the case of the Soman replication, the unified data collection clearly shows that the sample is attentive and careful, with one successful replication, which means that the failed replication is not due to issues with the sample or context/time. In addition, combining the two allows one to get better power for the same money invested, and additional analyses can be run to further identify participants who do not answer consistently across the two different scenarios in the two different studies. And, ofcourse, it shows that order does not impact these studies, which is also something that is important to know.

We ran many JDM replications with this design and across all the replications that implemented this approach we have yet to see any order effects, yet have been able to gain important insights regarding the phenomenon.

Additional recent examples with a unified design and diverging findings between studies:

Vonasch, A., Hung, W., Leung, W., Nguyen, A., Chan, S., Cheng, B., & Feldman, G. (2023). "Less is better" in separate evaluations versus "More is better" in joint evaluations: Mostly successful close replication and extension of Hsee (1998). *Collabra:Psychology*. [Preprint] [Open materials/data/code]

Chandrashekar, S., Adelina, N., Zeng, S., Chiu, Y., Leung, Y., Henne, P., Cheng, B., & Feldman, G. (2023). Defaults versus framing: Revisiting Default Effect and Framing Effect with replications and extensions of Johnson and Goldstein (2003) and Johnson, Bellman, and Lohse (2002). *Meta Psychology*. [Preprint] [Open materials/data/code]

Yeung, S. & ^Feldman, G. (2022). Revisiting the Temporal Pattern of Regret: Replication of Gilovich and Medvec (1994) with extensions examining responsibility. *Collabra:Psychology*, 8 (1): 37122. DOI: 10.1525/collabra.37122 [Article] [Preprint] [Open materials/data/code]

To address your point we added the following in our "data analysis strategy", to pre-register examining order effects in case we fail to find support for our hypotheses, with a compensation for alpha:

> **<u>Order effects between studies</u>**
>
> One deviation from the target article is that all participants completed all scenarios in random order. We considered this to be a stronger design with many advantages, yet one disadvantage is that answers to one scenario may bias participants' answers to following scenarios.
>
> We therefore pre-register that if we fail to find support for our hypotheses that we rerun analyses for the failed study by focusing on the participants that completed that study first, and examine order as a moderator. To compensate for multiple comparisons and increased likelihood of capitalizing on chance, we will set the alpha for the additional analyses to a stricter .005.

> **Sampling and Data Collection:**
> **An a priori power analysis is included, and generous upward correction provides confidence that the study will be well powered. It is clear that no data has been collected yet, and that data from 30 participants will be obtained to pretest the duration of the study to adjust payments. Extensive safeguards of data quality are included. I cannot foresee any ethical risks from this data collection.**
>
> **Data Analyses and Potential Results:**
> **The data analysis strategy is well prepared. However, possible interpretations given different outcomes should be stated more explicitly. This applies to both the individual hypothesis tests, where it should be clear which specific outcomes will confirm the hypotheses, and the overall evaluation of the replication attempts. I understand that the comprehensive method outlined in LeBel et al. (2018) will be used, but the registered report should be updated to clearly reflect how the specific outcomes of the replication attempts will be interpreted.**

Thank you for your valuable feedback. We appreciate the suggestion and you encouraging us to do better in clearly stating how we plan to evaluate the replication results in advance.

We provided an initial demonstration of our approach to interpreting the data outcomes in Table 12 of the main manuscript, based on Lebel et al. (2019), and we plan to extend that following data collection.

If we understood correctly, it sounded like your concern is how the overall replication would be summarized given different conclusions from the different studies. To address this point, we added the following to the “Evaluation criteria for replication findings” subsection in the methods section:

> We pre-register our overall strategy to conclude a successful replication if at least 75% of the studies (i.e., 3 or 4, out of 4) showed a signal in the same direction as the original study by Hsee and Kunreuther (2000), a failed replication if no studies (i.e., 0 out of 4) showed a signal in the same direction as the original, and any mixed findings with lower than 75% and above 0% (i.e., 1 or 2, out of 4) to be a mixed results replication.

## Reply to Reviewer #2: Dr./Prof. Fausto Gonzalez

> **The authors propose a replication of studies 1, 2, 4, and 5 from Hsee & Kunreuther (2000). They will perform a direct replication and some extensions.**
>
> **The proposed replication plan is generally sound, and the replication (as opposed to extension) versions of the materials match the original studies.**

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

> **If possible, the authors should increase the overall sample size. I worry the "250 per each of the four conditions" criteria may lead to underpowered studies. I acknowledge that the authors are using the effect size estimates from the original studies as a basis for their sample size. Still, the small sample sizes of the original studies may not reflect the true effect size range.**

We used the target's studies as an initial basis for our analysis, yet in our "Power and sensitivity analyses" we explained in great detail that we more than doubled the required sample size with additional margins to compensate for the possibility of exclusions. We then added a sensitivity analysis that shows the ability to detect effects of $f = 0.13$ and $d = 0.32$, far weaker than the effects reported in the target article ($d = 0.48$ to $0.82$). The studies in the target article were: Study 1 had N = 83, Study 2 had N = 89, and Study 4 had N = 46, and Study 5 had N = 98 (316 overall). Our target sample is several times that of any of these studies alone and combined.

We believe that we've compensated for the target's effects convincingly, and Dr./Prof. Rima-Maria Rahal seems to agree: "An a priori power analysis is included, and generous upward correction provides confidence that the study will be well powered."

We are happy to re-evaluate and adjust further if given clear editorial guidelines.

> **Another suggestion in the data collection phase is to add that MTurk participants meet a 95%+ HIT approval rate. Otherwise, the replication attempt looks to be in a good state.**

Yes, thank you, that is one of the many data quality measures we implement in our data collection on Amazon Mechanical Turk using CloudResearch. In the supplementary materials's "Additional information about the study" subsection we already had a placeholder section that details some of the measures we typically take for the data collections, and one of them was "We limited all workers' HIT Approval Rate to be between 95% and 100%.". Some of the other

measures we wrote in the main manuscript under “Participants”, such as “CloudResearch Approved Participants and Block Low-Quality Participants”, etc.

> **One small thing to note is that on pages 8 and 28 of the materials pdf on the OSF page, where the high affection and low affection camera scenarios are meant to be, it incorrectly repeats the painting scenario. This should be corrected on the OSF page (and likely on the survey since the pdf is an exported version of the survey).**

Thank you very much for catching that! We are grateful that you first went over our materials, and then that you were able to identify this oversight. The issue has been corrected in the revisions and resubmission, and files have been updated on the OSF. Wonderful.

## Reply to Reviewer #3: Dr./Prof. Bence Palfi

> **The manuscript aims to replicate an established phenomenon according to which emotional attachment to an object is related to insurance decisions about the object. I believe that the proposed registered replication report is relevant and very promising. It would certainly be intriguing to see if this influential effect replicates. I applaud the authors for choosing the RR format and for the level of transparency and rigour regarding their design, materials and data collection plan.**

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

> **However, I have identified some issues regarding the clarity of the concept of interest, the design and the planned analyses (or lack thereof) that I believe should be addressed before the in-principle acceptance is secured.**
>
> **Critical issues**
>
> **I like the authors` approach to focus on the replication of a main effect first, and only investigate the topic further if the main effect is established. I`ve found the introduction convincing and clear about the justification of the project, but I think the introduction lacks some clarity regarding the investigated concept (affect) and some consideration of the potential underlying mechanisms of the main effect. Defining and clarifying the concept would be crucial so that the readers can assess the validity of the hypotheses and the materials.**

Thank you for your suggestion. We were aiming to keep the scope of the replication to be mostly empirical, to follow the theory, claims, and methods of the target article without aiming to revise or correct their article. That said, we see the value in expanding further to try and give readers some context and refer them to additional readings.

In our revision we added a definition of affect and some introduction regarding the theoretical linking between affect and insurance decision making in the main manuscript, under the section “Linking Affect and Insurance decision-making: Appraisal-Tendency Framework (ATF)”

> **While reading the manuscript it was unclear to me what exactly is meant by high vs low affection and how it can be evoked/observed. It is great that all materials are transparently reported in the appendix, however, I feel that the high vs low affect manipulations are so critical to this project that they should be reported in the main text.**

Thank you for your suggestion, we wholeheartedly agree.

To ensure the transparency of our study, we now provide the definition and the manipulation in the main text, under the session: Hsee and Kunreuther (2000): Hypotheses and findings, High Affection and Low Affection Manipulation. Please see Table 2.

To further enhance the readability and clarity of our design, we moved the scenarios from the supplementary to the main manuscript. Please see Tables 4 and 5.

Together with the design table in Table 8, we hope that this provides complete transparency about all aspects of our materials and clarifies the manipulations and measures in the main manuscript so that readers don't have to guess or go search for those in other documents.

> **Also, the 4 scenarios have quite different ways to manipulate high/low affection. I think it would be ideal if some explanation or description were added about how and why these interventions were used in the original paper.**

Thank you for your suggestion.

In our revision we now provide the description of the manipulation of each scenario in the main text, under the session: Hsee and Kuneuther (2000): Hypotheses and findings, High Affection and Low Affection Manipulation.

We also acknowledge that the manipulations in the target article differ greatly from one scenario to another, some are not ideal, and we are not entirely sure why they were constructed that way. We also see some potential in also discussing the possibility of negative affect missing from the paper. We added this as a planned discussion following Stage 2 in our Discussion section of the manuscript:

> [Planned discussion (following feedback from reviewer Dr./Prof. Bence Palfi: The affection manipulations were a bit different between the scenarios and sometimes with no clear contrast between the high and low affection conditions (e.g., liked vs. not particularly crazy, and fell in love with vs. don't have any special feeling). We will discuss ways to further examine such contrasts, and also discuss high versus low affect versus high positive affect versus high negative affect versus no affect.]

**The role of demand characteristics.**

**I understand that this is a replication attempt of a phenomenon, but I believe that understanding why the effect appears is also important. Hence, I would like to invite the authors to consider the impact of demand characteristics in the current design. When reading the materials, I had the impression that explicitly telling people how they should emotionally evaluate a specific object is suggestive of the experimenter`s expectations and may give away what the affection hypothesis is about. For instance, the participants are told that they should think about an object as being very important to them, and then they are given the opportunity to demonstrate this expected commitment by reporting that they would drive as much as it takes to claim insurance (even if in real life they would not do so). Using a between-groups design reduces the impact of demand characteristics but I think it is still plausible that the main effect is driven at least to some extent by compliance.**

We understand this concern, it is a repeating concern raised in some of these paradigms. We personally tend to think of those as being overestimated, we find it hard to think of online labor market participants as taking the time to try and spend their time (and their income) to try and consider what the researchers are aiming for and then to adapt their responses in a way that would match guessed expectations. There is supportive empirical evidence to support our assertion. The first, which you raised yourself, is that it is very difficult for anyone to guess what the research hypotheses are in a between-subject design. There are studies showing that even in a within subject design this is extremely difficult for participants to do (Aczel, Szollosi, & Bago, 2018; Lambdin, & Shaffer, 2009). Then, even if we assume that they somehow guessed the research hypothesis, it is not clear why in our target sample they would try to appease that, there is no direct interaction with the researcher, and no incentive to try and change from one's own views to try and match alleged expectations from someone who you never met or will ever meet and with answers having no direct implications on compensation. Another piece of evidence comes from funneling sections that we add at the end of the survey. In our many studies with a similar design to this one, participants rarely (if ever) are able to deduce the research question or the hypotheses, even in cases where they have previously seen the materials or have been debriefed (MTurkers take hundreds of studies, and the task of remembering and connecting specific stimuli to a specific study is impossible difficult).

However, we agree that this concern cannot completely be ruled out, and is worth discussing. We therefore added a planned discussion for Stage 2 in the Discussion section of the manuscript:

[Planned discussion (following feedback from reviewer Dr./Prof. Bence Palfi: Discuss the possibility of demand effects in the target's design, reasons why and why not this may be the case (see reply to decision letter), and taking indirect measures to try and reassure readers (examining the funneling section).]

References:


- Lambdin, C., & Shaffer, V. A. (2009). Are within-subjects designs transparent?. *Judgment and Decision Making*, 4(7), 554-566.
- Aczel, B., Szollosi, A., & Bago, B. (2018). The effect of transparency on framing effects in within‐subject designs. *Journal of Behavioral Decision Making*, 31(1), 25-39.


**Lack of outcome-neutral tests.**

**One of the key features of RRs is the existence of outcome-neutral tests to ensure that the collected data are good enough to test the main question of interest. This is related to the ability of the IV to evoke the intended changes (high vs low affection) and to the ability of the DVs to pick up on the differences between the conditions.**

**First, clarifying what exactly is meant by "emotional attachment" is important so that the readers can assess the validity of the intervention. For instance, if emotional attachment is related to feeling sad after losing the object, then you would expect people to report being sadder after losing a high affect than a low-affect object. This could be tested with an additional question, but this may not be necessary.**

Thank you for the very valuable suggestion.

We clarified the definition of "emotional attachment" in the main manuscript, under the session Linking Affect and Insurance decision-making: Appraisal-Tendency Framework (ATF).

We feel it important to note that we do not see the manipulation as trying to evoke emotional reactions or an attachment, but rather an understanding that in the described imagined scenario, participants processed the information regarding the emotional attachment and incorporated that into their evaluations in the answers in the dependent variables (likelihood and pay/hours).

Great idea to incorporate a manipulation check. We added a question after the dependent variables asking about emotional attachment:

> "Please indicate how emotionally connected you feel towards the painting"
>
> (0 = *Not at all connected*; 5 = *Strongly connected*)

We added planned independent samples t-test comparisons between high and low affection conditions for all scenarios in the results sections, and a planned exploratory analysis of the correlations between the emotional attachment and DVs (for low/high conditions combined, per each of the studies).

> **Second, regarding the DVs:**
>
> **some outcome-neutral tests probing floor and ceiling effects, and the sensitivity of the DVs should be applied.**
> **For instance, the maximum amount of payment/ number of hours of driving measures have artificial upper bounds so they may have ceiling effects. By sensitivity, I mean the ability of the DV to detect a difference between the conditions. For instance, this could be tested by checking if objectively different insurance claims ($200 vs £100) evoke different responses: the vase in study 4 is worth $200 whereas the rest of the items are worth $100. Do people respond in the baseline (low affect) conditions differently for the vase than for the rest of the items? If not, then the DV may not be sensitive enough to detect the impact of the affection manipulation either.**

Thank you for your comment. This is an interesting insight.

As a first step, we added the amounts ($200 versus $100) to the tables in the main manuscript and in the Qualtrics, to make it clearer to readers that there are differences between the scenarios.

We also spelled out the DVs and made it clearer that the Pay DV actually differs between these conditions, such that the range for Studies 1, 2, and 5 is from 0 to $50 and more (increments of $5) whereas for the vase Study 4 it is from from 0 to $100 and more (increments of $10).

Together with our reply to Reviewer 1 regarding the camera scenario being different in setting the number of hours (4) required for processing, we added exploratory analyses comparing the studies and a planned discussion of the two parameters:

> **Differences between the scenarios**
>
> [n case some of the studies are supported whereas others do not: We will report differences between scenarios, with a special focus on differences between the Study 4 vase that was $200 compared to Studies 1, 2, and 4 that were $100, and between the Study 2 camera that explicitly noted 4 hours claim time versus the rest that did not indicate number of hours.]
>
> [...]
>
> [Planned discussion based on exploratory analyses for differences between the studies, discussing the differences in the vase scenario being $200 and the camera scenario setting processing hours to 4, and possible implications and future directions.]

We do not anticipate that these factors would have much impact on the results, but we are open to the possibility that they would and will discuss that.

To probe the floor and ceiling effects, we will include an additional exploratory analysis to compare the scenarios in the high affection and low affection group separately. To do so, we will analyze the distribution of the DVs' data by calculating the percentage of participants who scored at the minimum (floor) or maximum (ceiling) levels of the scale for each scenario. If more than 15% of the are at the extreme ends of the scale, it may indicate the presence of floor or ceiling effects. On the other hand, to assess the sensitivity of the DVs, we will perform a t-test to determine whether there is support for the differences in the responses between different groups or conditions, such as the baseline (low affect) conditions for the $200 vase and the $100 items. If there is no support, it might suggest that the DVs are not sensitive enough to detect the impact of the manipulation.

> **Analysis plan. I`m not sure why is there a need to run post-hocs when it is a 2x2 design?**
> **The main effect (high vs low affect) already tests the key question of the study. This test is analogous to an independent t-test simply comparing high vs low affect groups.**

Yes, thanks for catching that, that’s an oversight, post-hocs are only relevant here for contrasting the 4 scenarios, but not needed for the 2x2 high/low affection and compensation/claim. This has been revised accordingly.

> **I agree with the authors that the standardization in terms of what DVs are used throughout the study is a good idea. It is better to use the same measure throughout all situations. However, using all the measures from the original study introduces the problem of measuring one concept with multiple items (likelihood and payment/driving variables).**
>
> **Do I understand correctly that the authors will always use the same version (either the likelihood or the payment/driving) of the variable to test the main question that was used in the original article?**

Thank you. Yes, your understanding is correct. There will be two independent versions of comparisons, using either the hours/pay DV or the likelihood DV to test the main effect in the current study. We will not be mixing the likelihood and hours/pay DVs.

We tried to make that clearer in the revision in the results' tables.

> **If so, how will the authors interpret the alternative variables, especially if there is a conflict between the findings of the two versions? I think this issue relates to the problems raised by multiverse analyses, and how different operationalisation can lead to divergent results. I believe that it would be ideal to commit to one version of the DV for each Study/situation and check the other version as a robustness test only, and raise this issue in the discussion section (especially if there is some conflict between the DVs).**

Thank you. We now made that clearer in our Data analysis strategy (bolded):

> "We examined the main effect of two-way ANOVA tests for the replication DVs, which were stated as "Replication" in Table 3, to mirror the t-test condition comparisons in Studies 1, 2, 4, and 5 - Hours DV for Study 1, Likelihood DV for Study 2, and Pay DV for Studies 4 and 5. Replication findings evaluations will be according to these DVs."

> **I think there is also a high level of heterogeneity across the interventions. Different scenarios have very different high affect interventions so it can easily happen that they produce conflicting results. How will these findings be interpreted?**

We agree that this an issue, also raised by the other reviewers (see our reply above), which is why we added the following:

> We pre-register our overall strategy to conclude a successful replication if at least 75% of the studies (i.e., 3 or 4, out of 4) showed a signal in the same direction as the original study by Hsee and Kunreuther (2000), a failed replication if no studies (i.e., 0 out of 4)

showed a signal in the same direction as the original, and any mixed findings with lower than 75% and above 0% (i.e., 1 or 2, out of 4) to be a mixed results replication.

> **Statistical inferences. The authors may find it difficult to interpret some of their results given that they are non-significant. I recommend the inclusion of Bayesian analyses (the Bayes factor) so that the authors can distinguish between inconclusive results and clear evidence for the null. I think this is especially important for replications where the predictions of the alternative hypotheses can be easily determined based on the original study (e.g., you can use the original effect size or a discounted version of it, such as 2/3 of the original effect size), and it feels important to be able to distinguish data insensitivity from true null findings (for an example, see Bago et al., 2022 in Nature Human Behaviour). Bayes factors can be included conditionally (in case a test is non-significant) or they can be run for every single statistical test. JASP (https://jasp-stats.org/) offers a simple way to run Bayesian ANOVAs and t-tests.**

Thank you, these are very valuable recommendations. We agree, and will include Bayesian analyses to try and quantify the null when we fail to find support for rejecting the null hypothesis.

We added the following to the methods section:

> **Bayesian analyses**
>
> We pre-register that in case we fail to find support for the hypothesis for any of the studies, that we will run a complementary Bayesian analysis for that study (without outlier exclusions) using a prior of 0.707 to quantify support for the null.

And under exploratory analyses in the results section:

> **Complementary Bayesian analyses (for failed replication hypotheses)**
>
> [Please note that the Complementary Bayesian analyses is only to be completed in Stage 2 following data collection in case of failed support for replication hypotheses]