Reply to decision letter reviews: #187

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/vwLZGsOeUHoh

A track-changes manuscript is provided with the file: PCIRR-RNR-Soman 2001-Replication-Manuscript-v4-G-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	Ed, R3, R4: We proofread the manuscript and fixed the suggested typos. We made further minor edits to facilitate readability.
Introduction	Ed, R2, R3, R4: We expanded our introduction based on the feedback provided.
Methods	R1: We added information that the Soman's (2001) original sample comprised students. Ed, R2, R3: We clarified and expanded our reason for using a single link. R2: We clarified the sample upon which our power analysis is based. Ed, R2: We increased our assumed exclusion rate to 15% from 5% thereby aiming to collect data from 600 participants in order to end up with a final sample size of at least 515. Ed, R2, R3, R4: We moved the exclusion criteria to this section and expanded to provide all details for ensuring high-quality data collection. Ed, R2: We updated the categorisation of our replication according to LeBel et al. (2018) criteria. We now categorize Study 5 as being between a close and far replication.
Results	R3, R4, R5: We have clarified our motivation for conducting the additional within-subject exploratory analyses, and have removed the proposed logistic regression analysis for Study 1.

Section	Actions taken in the current manuscript
Discussion	Ed, R1, R4: We added planned discussions to contextualize our findings and added a subsection on "Limitations of replication", to discuss the differences between our replication and the original study.
Supplementary materials	Ed, R2: We updated the "Classification replication" table to reflect the differences between ours and the original. Ed, R2, R4: We moved the exclusion criteria to the main manuscript and expanded to provide all details for ensuring high-quality data collection.

Note. Ed = Editor, R1/R2/R3/R4/R5 = Reviewer 1/2/3/4/5

[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.]

Response to Editor: Prof. Chris Chambers

> **Five expert reviewers have now evaluated the Stage 1 manuscript. As you will see, the overall tone of the reviews is mixed, ranging from very positive to quite negative. However, all of the reviews are very constructive and I believe that a carefully considered revision and response can be suitable for in-principle acceptance (IPA).**

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

> **Several reviewers raise the concern of having the same participants complete all 3 studies in a single session as this may cause carry-over effects (despite counterbalancing) and could excacerbate demand characteristics.**

We believe this very point is a major advantage to our design rather than a disadvantage, 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 three 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 three effects.

In addition, this helps address concerns regarding the sample and attentiveness. When we have some failed studies and some successful studies, 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.

> **Sampling confounds or biases due to participants with prior experience of sunk cost studies, or the demographics differing from Soman 2001 (raised by Ronayne and other reviewers; this will require some specific consideration of whether and how these characteristics differ from the original study)**

We expanded our literature review, and now note that age is a variable that may potentially affect the differences between money and time effects. We added this as a planned point to return to in our General Discussion at Stage 2.

We also added a few planned notes in the General Discussion on limitations of the replication referencing the general population vs students issue.

Clarity and coherence of the analysis plan (raised by multiple reviewers)

We revised our exploratory analyses, removing the logistic regression for Study 1, yet kept the analyses for Studies 2 and 5 as those allow us to explore and potentially gain interesting insights, taking advantage of the within-subjects design.

In our revision we worked on improving clarity and coherence, by adding more details to our explanation of the planned exploratory analyses.

Considering additional factors that justify the scientific validity of the replication (as noted in Soman's review, e.g. proposed weakening of sunk cost effects over time; Peetz is the most critical on this point while also offering helpful suggestions for improvement, including consideration of additional literature)

Please see our reply to the reviewers and our improvement based on their feedback. We revised the introduction to better address the background for this replication.

Ensuring that the replication is as close as possible to the original study. Several reviewers raised concerns about procedural deviations. Some deviations will be inevitable – as always the the key is to identify those that risk violating theoretical coherence or which introduce (or resolve) methodological problems

We elaborated further on our decisions regarding adjustments, both in the manuscript and in our replies to reviewers below. We added clarifications regarding the modifications we made to the wording of the stimuli in Studies 1 and 5. We believe that according to the common replication evaluation criteria by LeBel et al. (2018) these differences still fall under the "direct replication" as these changes do not affect operationalization.

Resolving the question as to whether the replication focuses on the most important studies from Soman 2001 and is therefore optimally positioned to answer the research question (an interesting point raised by Leder)

We understand, yet this is a rather theoretical debate based on subjective evaluation. Studies 1 and 2 were the first in that paper, and so are the foundations to what came later, and Study 5 is a study that resembles and builds on Study 1 by examining an additional factor, that we thought would add additional insights. Replicating these does not contradict an empirical investigation that would look at the other studies, and one replication attempt should not invalidate the other. Our replication can help better our understanding regarding the stability, reliability, and robustness of studies on this phenomenon, and so a follow-up investigation aiming to replicate the other studies could build on our results to better finetune the priors for that attempt.

> **Clarification of methodological details such as consideration of additional exclusion criteria and replacement of excluded participants (raised by multiple reviewers)**

We now elaborate on our exclusion criteria in the main manuscript, and the steps we take to ensure high data quality. We also increased our assumed exclusion rate to 15% thereby aiming to collect data from 600 participants in order to end up with a final sample size of at least 515.

> **Quality of writing. There were differing views from reviewers (e.g. Olivola vs Soman). Personally I found the manuscript sufficiently clear to understand as a non-specialist, but it could be improved. Pass through and proofread carefully at the revision stage**

In this revision we aimed to clear any typographical and grammatical errors, while also making minor stylistic edits throughout to facilitate readability.

Response to Reviewer #1: Prof. Dilip Soman

This report is part of a larger project that aims to assess the replicability and dynamic stability of findings in decision-making research. I am a big supporter of this initiative and therefore delighted to have the opportunity to read this stage one submission. This paper replicates Soman (2001), whose finding basically showed that the sunk-cost effect which had been previously demonstrated to be relatively reliable in the domain of monetary costs, weaken and sometimes disappears when the costs are temporal in nature. I note that the studies reported in that paper were conducted in the 1997-1999 period (so now about 23-25 years ago), so I am interested to see how the results turn out now.

I agree with the authors' narrative on the need for conducting the replication. In addition to the points that the authors make [about the importance of the phenomena and citations], there are additional reasons that warrant a replication effort.

a) I believe that much has changed in the world since the time the original studies were conducted in terms of how people evaluate time versus money. For instance, waiting time back in the late 1990s when people did not have access to smartphones and were not connected to the Internet on the go was clearly more aversive than it is today.

b) It is also generally accepted that we live in much more of a time-constrained society today than we did in the past.

c) While we can question its scientific basis, the growth of the "no regrets, don't look back" philosophy towards life might certainly have implications for how people consider past sunk costs more generally.

Therefore, I have personally been very interested in the question of how some of the older demonstrations that relate to the properties of time as cost change if at all in today's world as a function of time. I would suggest the authors include some discussion on the aspect of dynamic stability of these effects as part of the justification for the replication.

I very much enjoyed reading this really well-written and well-structured stage one registered report (like the registered report, I write my review in a past tense recognizing that the actual data / analysis will look different). Based on all of the stated criteria, I believe that this registered report meets all of the necessary components. I especially appreciate the clarity with

> **which the sampling, analysis, and procedures were described. I was also thankful to the authors for the use of tables and visuals that made it easy for the reader to follow along.**

We appreciate the positive supportive constructive note, especially given that it is coming from the author of the target article for replication. In our experience so far, positive feedback from original authors has been quite rare, so this does mean a lot to us.

> **1) Table 2 is fantastic because it clearly allowed us to compare between the original and the replication. In terms of the gender and other demographic details, I do agree that the original manuscript did not disclose this information; however, we do know that all participants in that paper were undergraduate students so it might help to include that information.**

Yes, thank you. We previously mentioned this in the Supplementary in a table "Original vs replication methodological comparison".

We added an additional row in Table 2 in the main manuscript to indicate the different sample source.

> **2) There are also a couple of differences in the procedure for Study 5 that might be worth highlighting.**
> **First in the original paper, participants were students who were enrolled in a particular class.**

We included a planned subsection in the General Discussion addressing in advance "Limitations of our replication and directions for future research", where we plan to address this point in detail:

> "Our replication had limitations, and we needed to make several adjustments to the target's design to accommodate our sample and method of delivery. First, participants in the original study were students who were enrolled in a particular class, whereas participants in our replication were sampled from the general population. This makes it possible that the student sample was systematically different in some respect, compared to the general population. "

We will discuss the implications further after the data collection analysis.

> **Second, the manner in which opportunity cost was manipulated was slightly different in the original as compared with the replication (the exact words were different).**

Yes, we agree, and were planning to address this in Stage 2 discussion. We now added this to our new "Limitations of our replication and directions for future research" subsection in the "General Discussion":

> "Second, we made adjustments to the opportunity cost manipulation."

We will discuss the implications further after the data collection analysis.

> **Third, in the replication study, education was delivered by means of additional paragraphs that informed participants about economic approaches to time whereas in the original study, the manipulation was executed through differences in when the data were collected - for some participants, the study was done prior to a classroom discussion on the economic value of time versus for others it was done after the classroom discussion. It would be helpful to highlight these differences in the report.**

We added this to our new "Limitations of our replication and directions for future research" subsection in the "General Discussion":

> "Third, in the original, the education intervention was implemented by manipulating when the study was conducted – either before a classroom discussion about the economic value of time (control condition) or after (education condition) – whereas in our replication, the intervention was implemented by having participants read information on the screen and complete comprehension checks. These changes were necessary given the change in the medium, yet it may have affected the results."

> **3) I particularly like the summary of results section that clearly lays out the differences between findings from the original studies versus the replication and I also appreciated Figure 1, which visually communicated the same information succinctly. I also appreciated the additional analysis and robustness checks.**

Thank you for your encouraging comments.

> **One important aspect of both Soman (2001) and therefore this replication relates to the reliability of the sunk cost effect with monetary costs. Soman (2001) started with the previously demonstrated sunk (money) cost effect and made a case for why the effects would be weaker for time costs. Over the past 25 years, if people have indeed embraced the "don't look back" philosophy and do not pay as much attention to sunk costs more generally, it might create a situation where the basic premise of "weaker effects for time" might not make much sense. This perhaps leads me to suggest that part of your motivation for doing this replication should also include a brief discussion on the corpus of literature showing the sunk-cost effect in monetary domains and whether that is likely to hold up today.**

We now expanded on our introduction, yet we prefer to focus on a concise intro to the effect and leave possible interpretations of what might or might not be regarding a successful/unsuccessful replication to the discussion in Stage 2.

To address this point, we added a paragraph in the General Discussion to elaborate on the suggestion point about how cultural factors might have influenced the effect more generally:

> "Our replication was conducted more than two decades after Soman (2001) was published, with changes in the way people think of both time and money that might have impacted the findings. This is partly why ongoing repeating replications are needed, to keep our knowledge about an important phenomenon up to date. "

We will discuss the implications further after the data collection analysis.

> **Other than these suggestions, I thought the manuscript was extremely well written, well organized, methodologically sound and a pleasure to read. I would like to commend the authors on their thorough work and the excellent initiative.**

Thank you again for the encouragement. It is greatly appreciated.

Response to Reviewer #2: Dr./Prof. Johanna Peetz

> **After a careful review of this paper along with the provided materials on OSF, I find myself sceptical about the value of the proposed research. The scientific validity of the research question is unclear, and the methods (as proposed) would in my opinion not provide meaningful conclusions due to questionable data quality of the proposed sample and confounds and divergence from the original study in study design. I outline my concerns in more detail below.**

Thank you very much for reviewing our work and providing us with feedback.

> **The literature review or background section of this paper is extremely sparse and is insufficient in outlining the reasons for the research. The provided minimal arguments for this replication project are a) the impact of the original paper and b) the fact that it has not been (directly) replicated yet. There are lots of similarly influential papers that have not yet been replicated – so a more fulsome explanation of 'why this one' seems necessary.**

Yes, we would have liked to see a lot more influential papers being replicated, and this replication is part of a large-scale collaborative effort to replicate many classics in the judgment and decision-making literature (https://mgto.org/pre-registered-replications/). At this point in time we believe there is already a general consensus regarding the need for replications, and "making replications mainstream", without the specific need to justify a replication of a well-cited impactful paper that has not been subjected to independent pre-registered/Registered Report direct replications.

To address this we now include an explicit section about this need, with some citations of the work on that need:

> We aimed to revisit the classic phenomenon and examine the reproducibility and replicability of the classic findings by replicating the studies and improving the design with extensions. Following the recent growing recognition of reproducibility and replicability in psychological science (Brandt et al., 2014; Open Science Collaboration, 2015; Nosek et al., 2022; Zwaan et al., 2018), we embarked on a well-powered pre-registered replication and extensions of Soman (2001).

What we wrote as a justification for why this specific target is very similar to what we wrote in the many other published replications and our team's other PCIRR manuscripts that received in-principle acceptance. Examples from recent IPA PCI-RR:

1. Revisiting and updating the risk-benefits link: Replication of Fischhoff et al. (1978) with extensions examining pandemic related factors.
[IPA] [Preprint] [OSF]
2. Revisiting stigma attributions and reactions to stigma: Replication and extensions of Weiner et al. (1988).
[IPA] [Preprint] [OSF]
3. Revisiting the links between numeracy and decision making: Replication of Peters et al. (2006) with an extension examining confidence.
[IPA] [Preprint] [OSF]
4. Revisiting mental accounting classic paradigms: Replication of the experiments reviewed in Thaler (1999).
[IPA] [Preprint] [OSF]
5. Revisiting the psychological sources of ambiguity avoidance: Replication and extensions of Curley, Yates, and Abrams (1986).
[IPA] [Preprint] [OSF]
6. Revisiting the link between true-self and morality: Replication and extensions of Newman, Bloom and Knobe (2014) Studies 1 and 2.
[IPA] [Preprint] [OSF]
7. Associations of fear, anger, happiness, and hope with risk judgments: Revisiting appraisal-tendency framework with a replication and extensions of Lerner and Keltner (2001).
[IPA] [Preprint] [OSF]
8. Revisiting diversification bias and partition dependence: Replication and extensions of Fox, Ratner, and Lieb (2005) Studies 1, 2, and 5.
[IPA] [Preprint] [OSF]

Put differently, what exactly is the scientific value of this replication? Perhaps there is reason to doubt the original effect and a direct replication would allow for falsification of an established assumption. Perhaps identifying the exact effect size of the sunk cost versus sunk time effect would be helpful to other researchers. Perhaps identifying the boundaries of the effect would be helpful. Perhaps showing that the effect can be generalized to online samples in the US more than two decades later could be helpful. As it is, the introduction specifies no concrete scientific question and does not define precise hypotheses either.

Replications do not have to be novel (by definition!) but they do have to provide a justification for replicating a specific research. Such a justification is lacking in the present paper.

While there may not be direct replications, there have been indirect replications, which this section fails to mention – such as papers replicating the sunk cost effect for money but not time in different decision contexts (Pandey & Sharma, 2019) and some that actually did show a sunk time effect in yet other contexts (Navarro & Fantino, 2009; Castillo, Plazola, Ceja, & Rosas, 2020). These should be reviewed given that they likely reflect on the research question (once one is identified).

For example, if the main purpose is to establish concrete effect sizes, past

> **indirect research might help with this just as much as this one-time high-powered Mturk direct replication – in fact, there is likely enough work out there on these questions to complete a metaanalytic review already.**
> **In sum, there is a rich literature on sunk cost effects both on time and money that developed over the past two decades since Soman's (2001) studies and the present paper should clearly situate itself within this literature.**

The question of the value of this specific replication is tied to the broader question of the value of replications, especially given its impact and lack of direct well-powered pre-registered replications. Pre-registered direct replications are core to the scientific process and help update our knowledge regarding the target phenomenon, in terms of generalizability, effect size estimates, etc. A single study in a specific context should be considered as a first step in establishing a phenomenon. In our view, replications are not related to whether or not there are doubts regarding the effect, by now there are many examples of highly impactful phenomena that were considered beyond doubt that we have repeatedly failed to replicate in several large scale Replication Registered Report collaborations: "social" priming and ego depletion are to name a few. All these had vast comprehensive meta-analyses showing support for the effects based on what we now know is a biased literature. Independent well-powered pre-registered/Registered Report replication efforts go beyond conceptual replications and meta-analyses.

There are many challenges with the existing past literature, especially given that our literature suffers from publication bias towards positive and novel findings, and is based on underpowered studies that were not pre-registered and with no materials, data, and code shared to allow for error checking and reproducibility. That said, we agree that the existing literature is suggestive and can be helpful to guide the readers regarding what the literature does show.

Our scope for this direct replication was rather narrow and focused on the empirical effort to reproduce and replicate the original findings, and so we initially kept our literature review very brief, mostly to explain how the target article was embedded in the broader literature. We do appreciate the references to related literature; and we used this comment and the provided references as an opportunity to further expand our literature review in the introduction.

> **The method section assumed a 5% exclusion rate. It appears that this is based purely on the randomly generated data set used to populate tables etc. It would make a lot more sense to base the estimated exclusion rate on known exclusion rates for online crowdsourced samples.**
> **For example, in a very recent study on inattentive Mturk responders 13% were inattentive even after a number of ex ante data quality checks were in place (Pyo & Maxfield, 2021).**

We appreciate the comment.

First, to make sure we address this, we increased the planned exclusion to 15%. We do prefer to err on the side of caution and adding 60 more participants is very reasonable and we are glad to do so.

Second, we realized we were not clear enough in how we planned to conduct the data collection on MTurk and our measures to ensure high-quality data. We now include much more details on our process and criteria. We use CloudResearch/TurkPrime with very strict criteria to ensure quality and attentiveness (e.g., their approved participants). Our own experience showed very little need for exclusions, and that is supported by some of the recent evidence by others such as Eyal et al. (2021) who found that CloudResearch-approved participants consistently pass attention and comprehensions checks with over 95% accuracy (see their Figure 8).

Reference:

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 1-20.

Additionally, the number of ex ante data quality checks in this study seems underdeveloped – there are a number of restrictions in Mturk and Prolific that can be employed to ensure a lower chance of bots that are not mentioned here (e.g., 95% hit approval).
Many resources are available outlining best practices of attention checks in MTurk (Berinsky, Margolis, & Sances, 2014; Pyo & Maxfield, 2021; Thomas & Clifford, 2017) and the current data quality checks do not follow these recommendations as far as I can tell.

Yes, these are all standard in our data collections. We appreciate the feedback to do better in reporting all of our measures to ensure high-quality data collection.

We now include additional information about our measures to ensure high quality data collection in the "Procedure" subsection of "Method":

> "We will recruit native English speakers who were born, raised, and located in the US on Amazon Mechanical Turk using the CloudResearch/Turkprime platform (Litman et al., 2017). Based on our extensive experience of running similar judgment and decision-making replications on MTurk, to ensure high-quality data collection, we will employ the following CloudResearch options: Duplicate IP Block. Duplicate Geocode Block, Suspicious Geocode Block, Verify Worker Country Location, Enhanced Privacy, CloudResearch Approved Participants, Block Low Quality Participants, etc. We will also employ the Qualtrics' fraud and spam prevention measures: reCAPTCHA, prevent multiple submission, prevent ballotstuffing, bot detection, security scan monitor, relevantID"

After data collection we will also report all information regarding the specifics of the data collection in "Additional information about the study" subsection in the supplementary materials:

> "**[Note: Will be completed/updated after data collection]**
>
> This study was conducted on Amazon Mechanical Turk with American participants. We imposed the following settings in recruiting our participants:
>
> 1. Participants were paid $1.25 as a fixed participation reward. This amount was determined by multiplying the expected completion time (in mins.) with the minimal federal wage in the U.S. (i.e., $0.121 per minute).
> 2. The expected completion time was set at 10 minutes in advance.
> 3. The most time we allowed each worker to complete the study was 30 minutes.
> 4. We limited all workers' HIT Approval Rate to be between 95% and 100%.
> 5. We limited each worker's number of HITs approved to be between 5,000 and 100,000.
> 6. We blocked Suspicious Geocode Locations and Universal Exclude List Workers.
> 7. We blocked duplicate IP addresses and duplicate geolocation.
> 8. We enabled HyperBatch so that all eligible workers were able to participate in our HIT immediately after the survey was launched.
> 9. We restricted workers' location to be in the U.S."

> **The switch from real in-person surveys to online surveys comes with a chance of high inattention or even 'bots' producing random noise. In a good faith replication (especially one where original in-person collection is changed to online sample pools with notorious attention problems), the steps taken to make sure that participants are real and are actually reading the questions are extremely important. The present data collection plan would not make me feel confident that any potential null effect is not actually just due to poor quality, inattentive participants.**

First, we realized an oversight. To be absolutely clear, the sample will come only from a data collection of US Americans on Amazon MTurk (using CloudResearch). We had some mention of Prolific Academic in one of the tables in the submitted Supplementary document, which we now fixed.

Second, using the checks mentioned in the above comment and CloudResearch-approved participants has been shown to be robust, with successful passing of attention and comprehension checks of over 95% consistently (Eyal et al., 2021) and has also been successfully implemented in our past replication efforts. Thus, we do not believe that inattentive participants or bots are going to impact our results in any tangible way.

We have a lot of evidence to show the reliability of our target sample. We receive this comment quite often from reviewers that we are in the process of writing a manuscript aimed to address this specific issue and help others use the platform and achieve high-quality data collections. In our manuscript, we cited and referred to many of our other completed replication projects using this very approach. We will try and summarize our experience in short below.

We completed over 80 replications of classic findings in judgment and decision making using MTurk online samples (see https://mgto.org/pre-registered-replications/), and our experience has been that these samples are very reliable, at least for replications in judgment and decision making.

There is much that we can share on that but briefly:

1. Our successful replication rate is currently at 68% (+12% mixed/inconclusive), higher than most other replication rates in other domains. Even in the ones that are mixed/inconclusive or seemed to have failed we identified reasons that are not related to the samples.
2. When conducting 8 replications in two different online samples, Americans on MTurk and British on Prolific, we found the results highly consistent across the two samples.
    1. See summary tweet: https://twitter.com/giladfeldman/status/1215175786543534090?s=20
    2. Browse the reports: http://mgto.org/hkureplications2019
3. 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.
    1. Example 1: 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. Example 2: https://journals.sagepub.com/eprint/MVTW3KE2MXN2SRRKDGYE/full
4. When we ran the exact same replications on MTurk in two time periods, with a time gap of several months to two years, ensuring different participants from the same online platform, we found highly consistent results.
    1. Example 1: https://www.researchgate.net/publication/326548295_The_impact_of_past_behavior_normality_on_regret_Replication_and_extension_of_three_experiments_of_the_exceptionality_effect
    2. Example 2: https://www.researchgate.net/publication/339167597_Revisiting_status_quo_bias_Replication_of_Samuelson_and_Zeckhauser_1988

5. We have already collected and analyzed the data for most of the IPA-ed replication and extension PCI-RR mentioned above with similar unified designs, and the majority of those turned out to be successful replications. This is further reassurance that our measures for high quality data collections and our method of unifying several similar studies works well. Examples (theses with data collection completed, to be submitted as PCI-RR Stage 2 soon):
    1. Thesis: Revisiting the link between true-self and morality: Replication and extensions of Newman, Bloom, and Knobe (2014) Studies 1 and 2.
    [IPA] [OSF]
    2. Thesis: Revisiting diversification bias and partition dependence: Replication and extensions of Fox, Ratner, and Lieb (2005) Studies 1, 2, and 5.
    [IPA] [OSF]
    3. Thesis: Revisiting stigma attributions and reactions to stigma: Replication and extensions of Weiner et al. (1988).
    [IPA] [OSF]
    4. Thesis: Revisiting mental accounting classic paradigms: Replication of the experiments reviewed in Thaler (1999).
    [IPA] [OSF]

References:

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 1-20.

Bots do not seem to be an issue with CloudResearch-approved participants and with all our checks in place. Some related readings:

- Moss, A., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. Retrieved February, 4, 2019.
- Hauser, D., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2021, August 20). Evaluating CloudResearch's Approved Group as a Solution for Problematic Data Quality on MTurk. https://doi.org/10.31234/osf.io/48yxj
- Litman, L., Rosenzweig, C., & Moss, A. (2020). New Solutions Dramatically Improve Research Data Quality on MTurk.

> **The current data collection plan also does not outline whether excluded participants will be replaced and whether the power calculation refers to the final sample or simply to the number of slots posted on Mturk.**

All analyses and power calculations are based on the post exclusion sample of $N = 515$. We clarified this further in the "Power analysis" subsection to make that point more explicit.

> **Further, is there a planned point of percentage of discarded data at which the study would be deemed failed? Would you consider data even if, say, 30% of respondents have to be excluded?**

We do not have a plan for such a contingency and based on our experience, the exclusions are generally minor and have rarely had impact on the findings. We will include details about how many participants were excluded and also an analysis of pre- and post-exclusion results, as detailed in our supplementary ("Comparisons and deviations" section).

> **The planned design of running all three studies on the same sample of participants is not true to the original study designs. Since participants are being paid according to the time it takes to do the survey, it would require the same resources to run these three studies separately, so the reason for this divergence from the original study is unclear. Note that no reason is given for this considerable change from the original procedure - not even in the table outlining original vs replication methodological comparison that includes a 'reason for change' column (Appendix B).**

We laid our logic for this change in our Procedure section in the main manuscript. To address the comment, we now expanded that section (marked in bold):

> "We combined the three studies in a single online survey. This allowed us to maximize our resources and had the added advantage that we can rule out any sample characteristics that might be driving differences in successful versus unsuccessful replications. **Additionally, a single survey allowed us to conduct additional exploratory within-subjects analyses and explore links between different studies, something that is not possible with the original's design.**"

This was also elaborated in the "Original vs replication methodological comparison" table.

> **For example, if participants thought about and responded in line with a sunk time effect in an earlier question, they might then respond consistent with earlier responses in Study 5 and would be less likely to be swayed by the education condition. Giving Study 5 always after Study 1 or 2 stacks the deck against replicating the Study 5 effect because of these consistency biases in responding.**

The combined design was meant to be able to address specifically these kinds of questions. It is difficult to tell whether this change would stack the deck against the replication, increase its chances, or have no impact. Many judgment and decision-making paradigms are much stronger in within-subject designs than in between-subject design (e.g., omission bias, action-effect), though there are paradigms that only work in between-subject designs (e.g., less is better).

We are interested in whether education is capable of doing what it was hypothesized to do: change reactions. This unified design allows us to test those against baselines, and get a more accurate understanding of what is taking place.

We will examine whether there are differences in responding in Study 5, compared to Study 1, across the education and opportunity cost conditions - see subsection “Study 1 versus Study 5: Analysis of within-subject effects” in the main manuscript.

We also accept the point that the changes made to Study 5 in both the way we conducted the manipulation and in running after a similar Study 1 means that our replication criteria should be adjusted from “very close” to “between close and far replication”. This study is somewhere between a direct and a conceptual replication. We adjusted our replication classification in the supplementary to reflect that change.

Above, we mentioned several examples of our other replications with a similar unified design that were successful with the unified design providing valuable insights.

> **Even if all studies were administered in the proposed way (Study 1 and 2 counterbalanced, then Study 5), authors should check for order effects of the counterbalancing in all analyses. In the current proposed analyses, potential order effects are not tested.**

Good suggestion, we agree.

If we fail to find support for the original, then we will add additional order effect analyses. We added that to the “Additional analyses and robustness checks” subsection.

> **I disagree with the authors’ characterization of the IV materials in S3 as being the “same” as materials in the original study (according to Appendix B, Table on replication classification). In the original study, students were listening in person to a university lecture on opportunity cost. This information is coming from a source they take seriously (a professor at the university they attend) and is of considerable length and depth. This is not in any way the ‘same’ as reading 266 words about opportunity cost in an online experiment. Even if they show comprehension of this information on 2 followup questions, the information is not likely to be processed anywhere in near the same depth.**
>
> **As a followup on point 3, authors do not even plan to exclude people who did not answer one or both of the comprehension questions correctly (this is neither mentioned explicitly nor reflected in the simulated data). So, someone who skims the brief paragraph on opportunity cost and answers at random would still be included as bona-fide participant. In my opinion,**

> **this proposed experiment could not conclude whether opportunity cost education has a moderating effect on the sunk time cost effect because there is no evidence that participants read the or processed the information. Of course, the original experiment cannot be sure participants processed the information either, but in that experiment, participants were present for an hour+ long lecture – the likelihood that something 'sticks' is much higher than for 266 words in an online survey. I'm not saying that a meaningful replication of Study 5 cannot be done online but the proposed way of doing it is in no way equivalent to the original study.**

Good point, thank you for that feedback. We agree that our classification of Study 5 should be adjusted.

We updated our table in the supplementary materials subsection "Replication classification" table to "Different" for IV and DV stimuli, and the procedure to "different" given that it is always last. Thus we concluded our replication classification as "Close/far replication" instead of "Very close replication", as per LeBel et al. (2018) criteria.

> **A minor inconsistency. In the method description, the central outcome scale is labelled 1-9 (in line with the original scale). However, in the appendix (as well as in the OSF survey) the scale anchors are changed and the scale is now labelled 4 – 0 – 4 (not even -4 to +4). I would recommend the authors stay true to the original materials (as they said they did) in all aspects of the materials.**

Thank you very much for catching this and providing us with this feedback. Much appreciated!

First, we realized there were several oversights regarding our description of the measures where we could have done better and laid out things more clearly. We now adjusted the reference to the scale, explained our deviation (more on that below), and expanded on the measures in all of the studies. We also clearly label which measures are a direct replication and which measures were added as extensions.

Second, we appreciate the feedback that we should better document and explain our deviation from the original's scale of 1 to 9 to our adjusted scale of 4/0/4. We made this adjustment to try and avoid the possibility of biasing participants towards the larger number option. We thought that the 4/0/4 is more intuitive for participants to grasp than the 1 to 9. We felt that this was a needed adjustment, yet we would gladly readjust this to the original's scale given clear editorial guidelines to do so.

This has been added to the deviations table, and is now also noted explicitly in the text. We also better explain the differences between how this is presented to participants (4/0/4) and how this is coded (1 to 9).

Response to Reviewer #3: Dr./Prof. Christopher Olivola

> **I am happy to see replications of classic effects, and the sunk-cost effect (for time) is no exception. Therefore, I commend the authors for carrying this out. That said, I do have some comments and concerns about the current plan and/or manuscript:**

Thank you for your encouraging note and the helpful feedback.

> **First off, the current manuscript is poorly written. In particular, I see a lot of grammar errors that could (and should) have been checked and corrected (e.g., using grammar checks in Word). I sincerely hope the authors will make an effort to proof-read their word before they submit it for review.**

In our revision, we went over the manuscript again and tried to address grammar/spelling better.

> **The authors should cite and discuss other papers (besides Soman, 2001) that have previously tested (and found) sunk-cost effects for time (e.g., Bornstein & Chapman, 1995; Bornstein, Emler, & Chapman, 1999; Frisch, 1993; Navarro & Fantino, 2009; Olivola, 2018; Strough et al., 2008). This is important, since these other papers also speak to whether (and to what extent) there is a sunk-cost for time. In fact, I would suggest the authors provide a table that summarizes these other papers and, for each one, what they found (e.g., whether they found a sunk-cost effect of time).**

Thank you for suggesting relevant literature. We used some of these as the basis for expanding our literature review in the introduction. We do note that the intended scope for this replication is rather narrow, and we did not want to shift focus away from that with a review of the vast literature. Please also see our reply to the other reviewers on this point.

> **Having the same participants complete all 3 studies in a single session is problematic, as it may cause spillover effects, amplify demand effects, etc. The authors should consider randomly assigning participants to one of the 3 studies (not all 3). Or, at the very least, the main analyses should only focus on the first study that each participant is assigned to (and subsequent analyses can look at all 3 studies within-participant).**

Thank you, we appreciate this feedback and suggestion. Similar comments were made by the other reviewers, please see our detailed reply above.

In addition, to address your suggestion we added the following to our summary of the results:

> **Between subject studies and order effects (exploratory)**
>
> If we fail to find support for the target's findings, we will conduct additional exploratory analyses examining order effects and controlling for order.
>
> If we fail to find support for the target's findings, we will conduct additional exploratory analyses examining Studies 1 and 2 only when they were the first study presented to participants. This would address possible confounds between the studies, resembling running two separate studies.

> **Another concern, which may lead to a failure to replicate the effect, is that experienced MTurk participants may have been exposed (and some repeatedly) to sunk-cost studies, and this may hinder the effect. The authors should therefore consider limiting the study to MTurk participants who have had relatively little experience (e.g., fewer than 100 MTurk studies completed).**

We have an item in the Funnelling section (last in the survey) which asks participants "Have you ever seen the materials used in this study or similar before?", which we previously included in the supplementary materials in the "Materials used" subsection, and indicated those who responded "Yes" as part of the exclusions criteria that was in the supplementary.

We realized that this was difficult to follow and appreciate the feedback to make this clearer. We moved the exclusions criteria to the main manuscript, and provided additional information about our funneling section in the procedure subsection in Method.

> **On p. 20, the authors write: "In order to address H1, Soman (2001) conducted multiple chi-square tests. Specifically, in Study 2, he showed that in the money condition, the chi-square test found difference between sunk cost and no sunk cost conditions, whereas the same difference was not found for the time condition. A different way to approach H1 is to ask whether the likelihood of picking the option associated with sunk costs (theater performance in Study 1 and rocket engine in Study 2) is different across conditions. To address this question, we conducted a logistic regression analysis for Studies 1 and 2 for both the original and the replication data."**
> **=> I don't understand the distinction that the authors are trying to draw, here. Chi-Square tests also evaluate whether likelihoods vary across conditions, so the authors are mistaken if they suggest otherwise. I suspect they meant something else, but that it did not come across clearly in their writing.**

We agree that both chi-squares and logistic regression (LR) assess likelihoods, but chi-squares cannot test interaction effects. We are using LR, as per recommendations from R5 (Johannes Leder), which allows us to test an interaction effect in Study 2.

With that in mind, we agree with the reviewer that our writing was not clear in the original manuscript. Our goal was to communicate the importance of testing the interaction, which was the goal of our LR analysis. We have now revised the sentence in an attempt to make our point clearer:

> "A different way to approach H1 is to ask whether the likelihood of picking the option associated with sunk costs (rocket engine in Study 2) is different not only between levels of a single independent variable (sunk cost presence or sunk type) but also whether there was an interaction between the two variables."

To that effect, we also removed our LR analysis on Study 1. We agree, the chi-square analysis is sufficient.

Response to Reviewer #4: Dr./Prof. David Ronayne

> **The authors aim to replicate 3 studies from Soman (2001).**
> **The sunk cost effect is important. Distinctions between different types of sunk costs are important. Replicating studies related to that, e.g., those of Soman (2001), are therefore worthwhile contributions.**

Thank you for the positive opening note.

> **This is a replication paper; rationale and plausibility are clear. Below I suggest some wording changes from those presented in Table 1.**
> **Hypothesis 1: "More generally" does not make sense - domain is not a generalization of size... Perhaps splitting it into separate hypotheses would make more sense.**
> **Hypothesis 2b: I would not write "Rational" - you expect to be dealing with subjects who exhibit the sunk cost effect, at least when money is sunk (making their choices inconsistent with some textbook "rational" actor).**

Thank you. We stripped Hypothesis 1 down to the essential information in Table 1:

> "The sunk-cost effect is weaker in the domain of temporal costs than in the domain of monetary costs."

We also removed references to "Difficulty" and "Rational" next to Hypotheses 2a and 2b.

> **Order effects. Did Soman randomize the order of studies 1 and 2, as you will? If yes, did they find an effect? Please test for an order effect.**

Yes, we received similar advice from the other reviewers, and therefore added planned order related analyses in case we fail to find support for the original's findings.

> **Please report the average completion time and the lump sum offered for completion, not only the goal of 7.25/hour. Also, 7.25 is the federal minimum, but it differs by state (https://en.wikipedia.org/wiki/List_of_US_states_by_minimum_wage). I imagine it will affect your sample demographics, e.g., education level or employment experience, which could correlate with your outcome measures of interest. If you can, I would pay more. If you cannot, it would at least seem worth discussing.**

Thank you, appreciated. Yes, we do not restrict access based on state, but based on country. We do understand that this means that those from higher pay minimum wage states will not take part, or that participants from higher pay minimum wage states somehow decide to take part regardless.

We added the following to the planned limitations and future directions subsection in the General Discussion:

> [Based on feedback from peer review: potentially discuss how sample demographics, such as education level or employment experience, and our pay based on a federal minimum wage rather than by state, may correlate with outcome measures.]

In addition we now clarify in the methods section (underlined added):

> The assignment pay was calculated based on the federal wage of 7.25USD/hour (though we did not restrict participation based on state-level minimum wage).

> **Was Soman's sample undergraduate students? MTurkers are a different crowd and their different demographics may be a driver of the results you find. You cannot do a detailed comparison, as you report Soman did not disclose detailed demographic information, but there are systematic differences between college students and Mturkers e.g., age, experience, incentives, etc. I encourage a discussion of this potential source of differences, and analysis of how your subjects' demographics are associated to the treatments (see next points).**

We added the following the "Limitations of our replication and directions for future research" subsection of the General Discussion:

> "Our replication had limitations, and we needed to make several adjustments to the target's design to accommodate our sample and method of delivery. First, participants in the original study were students who were enrolled in a particular class, whereas participants in our replication were sampled from the general population. This makes it possible that the student sample was systematically different in some respect, compared to the general population. "

> **Logistic regression. You already conducted Chi-squared tests for Studies 1 and 2 and you have predicted proportions from the raw data. Why run logistic regressions with only the treatment (dummy?) variables on the RHS? The value I see in regression analysis would be to see if there were some interesting covariates of sunk cost behavior not picked up in Soman, that may explain your data e.g., subject demographics. Please revise or justify why the regressions you propose add value.**

Based on your and Dr./Prof. Christopher Olivola's (R3) feedback we improved on our reporting of the logistic regression (LR) analysis. We are using LR, as per recommendations from Dr./Prof. Johannes Leder (R5), to allow for the testing of an interaction effect in Study 2.

We revised our manuscript to reflect this:

> "A different way to approach H1 is to ask whether the likelihood of picking the option associated with sunk costs (rocket engine in Study 2) is different not only between levels of a single independent variable (sunk cost presence or sunk type) but also whether there was an interaction between the two variables."

We also removed our LR analysis on Study 1, as the chi-square analysis seemed sufficient.

> **Can you explicitly confirm whether the within-subject (design and) analysis was also done by Soman? (I guess it was not - but if yes, a comparison is needed)**

Soman (2001) did not run the studies together and therefore there was no within-subject analysis. We placed this analysis under a section "Additional analyses and robustness checks" in order to make it explicit that this is different from the replication.

> **Overall, except potentially the last point above, this is a pure replication paper. That is of course a great goal in itself. But do you want to explain your and/or Soman's results? If nothing else than by discussing how they may vary with demographics or other variables? I did not see any analysis aimed at this, yet it would seem straightforward to add (e.g., via multivariate regression - see point above) and potentially enlightening.**

We aimed this as a replication paper, and we do go beyond the exact replication by adding additional analyses and extensions. Our scope is well-defined and complex enough, and so we do not plan to make any additional exploratory analyses to examine demographics or other variables, this would complicate the investigation and shift the focus from the main aim needlessly.

We are making our materials, data, and code publicly available, so anyone interested in exploring further, is very welcome to do so.

> **Please list the details of the Qualtrics implementation somewhere, e.g., availability of a "back" button, time limits, forced responses, etc. The idea being that someone could fully replicate your work with all the same options selected in the software.**

We fully support others being able to fully replicate our work. To that effect we already shared an exported .docx, .pdf, and .qsf of our Qualtrics survey. These documents contain all relevant details that one might need to replicate our work. The .qsf is easiest to use but can only be used within Qualtrics (anyone can sign up for a free account, that allows navigating the QSF), while .docx and .pdf can be used by anyone. These are contained in the online OSF repository under the "Qualtrics survey" folder.

> **Please make your data and analysis code available ex post.**

All materials, data, and code will be made available, this is standard practice in all our projects (https://osf.io/5z4a8/).

The "Open Science Declaration" section in the studies overview reads:

> "This replication is submitted as a Registered Report (Chambers & Tzavella, 2022; Nosek & Lakens, 2014; Scheel et al., 2021; Wiseman et al., 2019).
>
> We will pre-register the experiment on the Open Science Framework (OSF) and data collection will be launched shortly after pre-registration. Pre-registrations and all materials used in these experiments are available in the supplementary materials. We provided all materials, data, code, and pre-registration on: https://osf.io/pm264/.
>
> All measures, manipulations, exclusions conducted for this investigation will be reported, all studies will be pre-registered with power analyses, and data collection will be completed before analyses. We reported results after exclusions below, and in the supplementary materials, we detailed a comparison between pre- and post-exclusion findings as well as any deviations from the pre-registered plan ("Comparisons and deviations" subsection), with additional disclosures ("Open science disclosures" subsection)."

> **The exclusion criteria seem good, but I think some other good ones are not listed. First, a criterion based on how quickly the main pages were submitted. Too fast means it was infeasible that they read the text. Second, on a page following the main text (and assuming there is no ability to go back) it would be good to ask questions to check subjects paid attention (questions they would only know the answer to if they read the text). (You ask some questions in Study 5 which require subjects to understand some conceptual info, but those are part of the education treatment.)**

We appreciate these suggestions, which helped us realize that we should do better in explaining the measures we took to ensure high-quality responding.

Before the participants embark on the studies they must indicate their consent, qualifications, and agreement, and we deliberately randomize the choices in these questions which requires participants attention. Those who fail to indicate "Yes" to these questions are asked to return the task. This helps ensure attentiveness.

We now explain this in greater detail in the "Procedure" section:

> Participants first provided consent, after which they read an outline for the studies and three questions confirmed participants qualifications as being American, their understanding of the study procedures, and their agreement to pay close attention (Yes/No/Not sure presented in random order, and participants not answering Yes were asked to return the task).

We added additional information about our recruitment criteria in how we ensure high-quality data collection, please see our reply to Dr./Prof. Johanna Peetz above.

The subsection "Additional information about the study" in the supplementary also details all the criteria used in the recruitment (which will be updated after data collection if there were any changes), please see our reply to Dr./Prof. Johanna Peetz above.

> **Related, you expect only 5% of their sample to be excluded by their criteria (p11). My sense is that that may be conservative. I believe the present study would be significantly boosted by stronger exclusion criteria and a correspondingly larger initial sample.**

Similar comments were made by other reviewers, thank you for the feedback. We now make clearer our measures for ensuring high-quality data collection, our exclusion criteria, and we updated our estimates for the exclusions to 15%, please see our reply to Dr./Prof. Johanna Peetz above.

> **I recommend the authors check the following references, especially the last one which looks to some extent at the distinction between time and money**

in a sunk cost context.
Augenblick, 2016, "The sunk-cost fallacy in penny auctions"
Olivola, 2018, "The interpersonal sunk-cost effect"
Ronayne, Sgroi, and Tuckwell, 2021, "Evaluating the Sunk Cost Effect"

Thank you for suggesting relevant additional literature. We used it to expand our literature review in the introduction.

I believe it would be better to qualify the introductory definition of the sunk cost effect as relating to *irreversible* or *unrecoverable* investments of resources, e.g., line 1 of Abstract.

We now added "unrecoverable" in our definition of sunk costs in the Abstract:

"The sunk cost effect is the tendency for an individual's decision-making to be biased based on unrecoverable previous investments of resources."

p7 "found that the sunk cost effect was ... not [present] for time": you cannot prove a null. Please re-phrase e.g., "no evidence for an effect of..."

Thank you for this feedback, much appreciated, we agree completely.

We rewrote the sentence to the following:

"Soman's (2001) core finding was that the strength of the sunk cost effect was weaker for time than for money. He further showed that the facilitation of money-like accounting for sunk time costs by highlighting opportunity costs or by educating about an economic approach to time strengthens the sunk time cost effect."

Two examples of unclear writing below from early in the manuscript. I shall refrain from further comments about the writing, but recommend you have the final manuscript proofread.
a. I do not understand the phrase at the end of the first sentence of the first paragraph of the Introduction"given that with larger sunk costs are stronger tendencies to further escalate". I would avoid the speculation over the ("vicious cycle of") consequences of the SCE, and just talk about the SCE itself, except for the discussion.
b. p1 "yet evidence is sometimes inconsistent with weak effects" does not read well. There are at least two different possible meanings.
c. p7 "appeared" rather than "re-appeared"

Thank you for the helpful suggestions.

We revised the manuscript as follows:

a. We removed “vicious cycle” from the first sentence:
   “...leading to an escalating commitment to a losing course of action”
b. We removed “with weak effects”:
   “...yet evidence is sometimes inconsistent…”
c. Changed “re-appeared” to “appeared”

I am not sure that 420 citations in 21 years is a huge amount. Also, some people think Google Scholar is a poor citations counter. I think it distracting and unnecessary and would remove it.

This is a typical section that we included in all our replication submissions, including in-principle accepted preprints by PCIRR, and is a result of requests by some reviewers to demonstrate impact, commonly measured by number of citations. Based on our experience in replicating classics, 420 citations in 21 years is definitely unusual and far above average in the judgment and decision-making domain. Google Scholar, with all its weaknesses, is considered one of the most comprehensive databases, and - unlike other resources - is completely free.

If needed, we provide some citations which demonstrate that 420 citations for 21 years ranks at top 1% of journal articles across top Psychology/Cognitive journals:

- Cho, K. W., Tse, C.-S., & Neely, J. H. (2012). Citation rates for experimental psychology articles published between 1950 and 2004: Top-cited articles in behavioral cognitive psychology. Memory & Cognition, 40(7), 1132–1161. https://doi.org/10.3758/s13421-012-0214-4
- Kurilla, B. (2015). How Many Citations Does a Typical Research Paper in Psychology Receive? – Geek Psychologist. https://geekpsychologist.com/how-many-citations-does-a-typical-research-paper-in-psychology-receive/

Typos I spotted:
Authorship declaration: "is" in line 1.
p10. "have possible detected" should be "possibly".
p15. "we found was", remove "was".11
p22. You write "no support for a main effect of sunk type" and then an effect with significance p=.001…
p22. "Soman found a main effect of sunk presence

Much appreciated. All the listed typos have been fixed.

Response to Reviewer #5: Dr./Prof. Johannes Leder

> **The study seeks to replicate the hypothetical scenarios used in the experiments 1,2 and 5. I am not sure if these are the experiments that the resources should be focused on. The study seeks to replicate a sunk cost effect for time and money DECISIONS – the replication proposed now only seeks to replicate the effects for INTENTIONS. Here is a severe mismatch. Soman (2001) used experiment 6 to validate his previous findings, for this reason, experiment 6 seems to be the most crucial experiment for his argument and not study 1,2, and 5. As he states: "Experiment 6 involved real choices made by individuals who had made real investments of time. The results validated Hypotheses 1 and 2a, namely that the sunk-cost effect was not detected in the domain of temporal investments, but it reappeared when the accounting of time was facilitated."**

Thank you for the feedback. We appreciate you sharing your view, and we agree that there is value in examining real choices made by individuals. We also see value in examining intentions. Both intentions and decisions are important, and then it is a matter of priorities. Soman has shown that the phenomenon seems to extend to both intentions and decisions, and so we felt it important to first address intent before we embark on the more costly and complex real decisions. In our view, this is a necessary first step, and we would like to first establish the phenomenon demonstrated in the initial studies.

Also, effects are typically much larger and easier to detect for intent than they are for decisions, and so this increases the likelihood that we will be able to detect effects, especially if the study design involves interactions. If we are successful in replicating Studies 1, 2, and 5, we hope that this will pave the path for pursuing a replication of Study 6 and real-life behavior in future studies.

> **I appreciate the detail and care the authors have taken in simulating the data and showing the results in an adequate statistical framework (logistic regression). The analysis of the preference ratings should be done with a cumulative logit or probit regression and not ANOVA as the measure is not truly continuous -see:**
> **Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? Journal of Experimental Social Psychology, 79, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009**

Thank you for the valuable feedback and suggestions on an alternative analysis. We also see much value in running additional analyses using logistic regression for a more robust

test of Soman's predictions, which is why we incorporated this type of analysis in the manuscript, in addition to replicating the original's analyses.

In running replications it is important that we atleast try and run the same analyses as the target's, and compare the effects using the same analyses as the ones conducted back in the day. We prefer to err on the side of doing and reporting too much than doing too little.

We added a planned brief discussion of this point in our general discussion in the "Limitations of the original study: Directions for improvement" subsection.