Reply to PCIRR Stage 2 decision letter: Thaler (1999) replication and extension

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response as. 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/xkRuLviKgZgM (https://osf.io/az4e7)

A track-changes manuscript is provided with the file: PCIRR-S2-RNR-Thaler1999-replication-extension-main-manuscript-trackchanges.docx (https://osf.io/s2rvn)

Response to Editor: Dr./Prof. Chris Chambers

> **Thank for your completed submission, which I enjoyed reading. The two reviewers from Stage 1 also returned to evaluate your Stage 2 submission, and as you will see, both reviews are positive about the completed manuscript. The reviews contain some useful suggestions for qualifying certain aspects of the interpretration (or at least justifying in a rebuttal why such qualifications are unnecessary), and for streamlining the presentation of the results. Provided you are able to address these minor concerns in a revision, final acceptance should be forthcoming without requiring further in-depth review.**

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

We appreciate the time and effort that you and the reviewers have dedicated to providing your valuable feedback on our manuscript. We adjusted the presentation of the results and further improved the interpretations. The reviewers' comments and concerns were addressed below in a point-by-point format.

Note: Given the feedback from the reviewers we realized an oversight which we improved on and addressed better in this revision. We now adjusted the directionality of the effects we reported so that a positive effect size would indicate an effect in the same direction as the hypotheses and the target article, and a negative effect size would be in the opposite direction. This has also helped in our summary of the comparison of the many effects in the target article with the effects found in our replication. We also added columns to the tables as well as a visual summary of the effects to allow for much easier comparison to aid readers in interpreting the many findings.

Response to Reviewer #1: Dr./Prof. Barnabas Imre Szaszi

> **First of all, thank you very much again for submitting your paper and for the opportunity to review it. Despite this being a Stage 2 review, and have read through the whole paper and provided minor suggestions. In sum, my read of the paper that it followed the plan accepted at Stage 1 and the general content of the paper is acceptable.**
>
> **Most of my comments concern the way the results are presented, but I think that the way the results are presented should be improved. Again, you need to note that I'm not an expert on mental accounting, so I could not properly review the validity of the statements concerning different theoretical aspects of the paper. You can find below my specific comments:**

Thank you so much for taking the time to review our paper and for providing such detailed feedback. We sincerely appreciate your efforts in offering so many insightful suggestions for improvement.

> **Abstract:**
>
> **.1. In XX "a" missing XXX Replication Registered Report with an American online Amazon Mechanical Turk sample using CloudResearch (N = 1007).**

Thank you. Changed to:

> In a Registered Report with an online U.S. sample recruited from Amazon Mechanical Turk using CloudResearch, we conducted a replication of 17 classic problems reviewed in Thaler (1999) (N = ~500 per problem; overall: N = 1007).

> **.2. "Systematic replications and extensions of review articles using a single data collection are a promising direction in revisiting seminal findings, mapping and examining untested assumptions and predictions, comparing different designs and effects, and identifying possible links, gaps, and future directions".**
>
> **I would be happy to read some specifics in the abstract. Would be great to read about the conclusions. E.g., in the PCIRR study design table one main question is the link between different mental accounting problems. Or any results that the authors think**

Thank you for highlighting this. This is something we set out to do in Stage 1, yet in Stage 2 realized that the many designs of the different problems ran makes a simple summary a major challenge.

Therefore, we decided to remove that statement from the abstract, and to remove that row from the PCIRR design table.

> **Introduction:**
>
> **.3. "The earliest empirical evidence on mental accounting behaviors dates back to Tversky and Kahneman's (1981) famous theater-ticket experiment (one of our replication problems)."**
> **Would be great to shortly explain what this experiment is, because the average reader probably wont know what it is, or alternatively just delete the sentence.**

Thank you for raising this concern. We added a brief explanation for it:

> Mental accounting has long been a heated topic in the field of behavioral economics, psychology, and judgment and decision making. The earliest empirical evidence on mental accounting behaviors dates back to Tversky and Kahneman's (1981) famous theater-ticket experiment (one of our replication problems). In that study, participants were asked whether they would be willing to pay $10 for a ticket following a loss, and the authors contrasted two conditions which manipulated whether the participants had lost a previously purchased ticket for the same show or lost an equivalent $10 bill. The results showed that people were less willing to purchase the ticket after losing a ticket compared to after losing an equivalent cash amount.
>
> Tversky and Kahneman proposed that mental accounting is a form of decision framing by which people formulate (psychological) accounts to evaluate events and options (as cited in Henderson & Peterson, 1992). People categorize funds into different mental accounts designed for different purposes. Participants likely perceive the funds required to

> "repurchase" the ticket as drawn from the mental account for ticket expenditures, which had already been used in the initial purchase. In contrast, the cash loss was not assigned to a discrete mental account. This distinction violates the long-standing economic notion of fungibility (Thaler, 1999).

> **.4. "In Thaler (1985) and our target article-Thaler (1999)," is this correct? Shouldn't this be "In Thaler (1985) and in our target article (Thaler, 1999)"**

Thank you for the suggestion. We changed to the following:

> "In Thaler (1985) and in Thaler (1999), our target article,..."

> **.5. Instead of this "The two recent review papers cited very similar research to Thaler (1999), such as Heath and Soll (1996), Tversky and Kahneman (1981), Thaler (1980), and Thaler and Johnson (1990). This further exemplifies the necessity in…"**
> **I would suggest to change the wording, because it is hard to follow the logic: "A very similar set of papers formed the basis of these previous review papers, exemplifying the necessity …."**

Thank you for the suggestion. We agree and changed to:

> "A very similar set of papers (e. g. Heath & Soll, 1996; Thaler, 1980; Thaler & Johnson, 1990; Tversky & Kahneman, 1981) served as the basis of Thaler (1999) and the two recent review papers. This further exemplifies the necessity in revisiting these classic findings and testing the reproducibility, robustness, and generalizability of these influential and pioneering works, to substantiate and strengthen the empirical foundations of the theoretical framework of mental accounting."

> **.6. "We also recognized the potential for improving on both transparency and methods." To "We also recognized the need to improve both the transparency and method of some problems reviewed by Thaler"**

We preferred to frame this as a "potential" rather than a "need". It suggests a less critical view of the original studies, which in the grand scheme of things used an approach and reporting standards that were common at the time. In the early stages of behavioral economics it was especially common to use single scenario vignette studies and focus reporting on descriptives. We however made a slight adjustment to the following:

> We also recognized the potential for updating and improving both the transparency and the methods used in some of the problems reviewed by Thaler (1999).

Methods:

.7. In table 3 your write (and at other places you refer to as well) that the sample size in the study is 1007. Which is factually correct, but it is misleading because it is compared with the sample size for different problems. But for the case of the present study, the true sample size is half of the study sample size (500 instead of 1000). I would suggest to use 500 as the reference number at each place suggesting that this is the sample who completed the each of the problems.

Thank you for raising this. Valuable and important feedback.

We adjusted the relevant sections and introduced a brief explanation in the method section to provide clarity on this matter.

> "As participants were randomized to complete 9 out of 18 Qualtrics blocks, there were approximately 500 participants for each problem."

We made similar adjustments to our abstract:

> In a Registered Report with an online U.S. sample recruited from Amazon Mechanical Turk using CloudResearch, we conducted a replication of 17 classic problems reviewed in Thaler (1999) (N = ~500 per problem; overall: N = 1007).

Results:

.8. Table 9 please put the original and the replication percentages next to each other to help the comparison of the two. (Apply the same logic for the other tables as well please).

Thank you. We adjusted Table 9 and all other tables to display the original's descriptives next to the replication's descriptives, to make it easier to compare the findings. We also added a column that provides our interpretation of the comparison between the two.

We also revised Table 16 that reports standardized effect sizes to allow for summarized comparison, and comparison using LeBel et al. (2019).

.9. In some of the table the font seems to be changing. (Even within the table).

Thank you, corrected.

> **.10. Please provide at least a summary textual description of all the results main results for each of the problems, or some way help the reader to understand the results. Now, the results part basically is a few sentences and a lot of tables without context which is very hard to follow.**

Thank you. We were originally aiming to keep things very concise and to the point, yet we see the value in relaying things in mode detail in text. That process has also helped us catch minor oversights and realize the need for further exploratory Stage 2 complementary analyses to help better clarify the effect.

We added a brief description of the findings for each problem across all tables. We also added a summary of the effect sizes and confidence intervals in Table 16.

> **.11. I would also add a summary table which shows which of the problems were replicated / or not. (So the content of the results are OK, but the way it is presented is not).**

Thank you. We agree, this was needed. We appreciate the suggestion.

We added a column in Table 9 (descriptives) and in Table 16 (effect sizes and confidence intervals) reporting our interpretation of replication success.

> **.12. "Therefore, we only compared the direction and relative magnitude of the mental accounting effects in some of the problems where it seemed to be meaningful."**
> **Please describe your criteria for seem to be meaningful (or not meaningful) was.**

Since our previous submission, we completed a collaboration with stats scholars in our field that resulted in the "Effect size and confidence intervals" book (https://mgto.org/effectsizeguide) which has helped us make progress in calculating effects from very little provided information, to tackle more complex effects from the original studies, and to compare those to our effects.

In our current revision, we include effect size calculations for all problems, summarized in Table 16. The sentence you therefore referred to is no longer relevant, and we revised to the following:

> We summarized the findings, the comparison to the original findings, and our interpretation of our findings in comparison to the original findings in Table 16.
>
> We used the replication evaluation criteria by LeBel et al. (2019). In Stage 1 and our initial submission of Stage 2 we were sometimes unable to deduce the effects, yet in our revision of Stage 2, we were able to use a guide by Jané et al. (2024) to provide effect size for most effects in the original studies, and for all of the replication effects. In the

original problems where we did not have enough information, we simply used "signal" versus "no signal" and the direction of the effect in our replication interpretation.

.13. "This exploratory analysis was an innovative and preliminary attempt to study the connections among different subsets of the mental accounting framework. The results indicated that further explorations hold some promise."

I would take out the word innovative and leave it to the reader to make this judgment. It would be great iThaler1999JAMOVI.RDataf you could interpret the results at least to some basic extent beyond this very short and vague sentence.

Thank you. Upon reflection, we decided to remove that exploratory analysis and keep things simpler and more concise. The manuscript is already very complex with many different problems using very different study designs, and our revisions added considerable text to further explain things. We felt that our attempt to integrate those together may be distracting and shift focus from the main aim of the replication. We also moved the other exploratory analyses from the main manuscript to the supplementary materials.

Discussion:

"Among the Problems, Problems 1, 14,…" -> 'Among the problems' is not needed.

Thank you. We changed that sentence, and generally revamped to integrate those summaries into the results section, per your earlier suggestion.

"…and the results were in the expected direction. The earlier loss could not induce risk-seeking in both Problems". Is the expected direction that earlier loss could not induce risk seeking?

We worked to make it clearer in each problem what the findings were in the target article, and what we found. We moved those to a more detailed reporting of the results per each problem.

> **“ … may cost more than $5 to drive 20 minutes to the other store with the increasing cost of driving. Therefore, the inconsistency may be due to participants' awareness of driving costs rather than a lack of mental accounting effect. It is also possible that over time since it was conducted, the value of $5 in relation to transport costs has changed dramatically enough to shift participants’ preferences entirely.”**
>
> **To me this a very plausible explanation, as driving 20 minutes for sure costs at least 5 dollar these days. However, if this is true, I would not call this a non-successful replication, I would categorize this as miscellaneous as potential methodological concerns were raised after the data collection.**

Thank you for bringing this up. We adhered to the pre-registered criteria. Post-hoc interpretations after seeing the result are tricky. Anything may seem more plausible and less surprising post-hoc, but we need to be very careful in not jumping into conclusions and changing our criteria. In this case, we failed to successfully replicate the study using the original stimuli, and so now future studies can follow up and examine what adjustments are needed.

We therefore added the following:

> Problem 2 examined people’s perceptions of the value of time. The majority of the participants were unwilling to drive 20 minutes to save $5, regardless of the price. In hindsight, this makes sense, and raises an important dilemma in conducting replication studies - whether or not to update and adjust the prices to current days. As one of the participants in the feedback section pointed out, it may cost more than $5 to drive 20 minutes to the other store with the changing costs of driving and the devaluing of money since the problem was first presented in the 1980s. What would have been plausible to some participants in the 1980s, may no longer be plausible to most of current time participants. Therefore, the inconsistent findings may not be due to lack of an effect, but rather because of the changing circumstances. The differences in the findings could be due to the passage of time, and could also be due to our sample’s demographics compared to the original’s, or because of the unified design. The exact reason for the different findings is down to speculation. Future replications could further examine adjustments to the scenarios, or to assess different moderating factors.
>
> We felt that it is important to first conduct a direct replication using unadjusted prices because otherwise the differences may have been attributed to our changes. We at least now have the insight that the phenomenon cannot be observed using the same questions in the current context, and future studies can now have a better estimate of the likelihood of replicating the study without adjustments.

> **"Taking into account the replication success of Problem 3, these together revealed that the sunk cost effect might be context-based." What do you mean by context-based? I have a strong opinion that in psychology everything is context based so this statement is very vague.**

We revised to the following:

> Problems 18, 19, and 20 all targeted the sunk cost effect. In Problem 18, more participants chose to weather the snowstorm and make an effort to go to the sports game when they paid for the ticket as compared to when they received the ticket as a gift. However, in Problem 19, the large majority of the participants said they will not continue playing tennis after suffering an elbow injury, despite the expensive $300 membership fee. In Problem 20, participants agreed that they tend to keep the uncomfortable shoes longer when the price is higher, yet with no indication for the prediction that they would also try to wear them more. Taking into account the replication success of Problem 3, these together revealed that the sunk cost effect might be context-based, such that sunk cost effects may not manifest when there is anticipated physical pain.

> **".. sizes, and to try and map links between the different studies. We provided one such initial analysis focusing… "**
> **See my related comments above. The conclusion is missing and the point why this is useful.**

Thank you for the feedback. We decided to remove these analyses. There is already too much going on in this manuscript, the studies employ designs that are very different and difficult to compare and interpret, and so we feel this is best left for future follow-ups.

> **"We recruited a much larger and more diverse sample than the original studies"… I would only say that "we recruited a larger sample." I am not very much convinced that the sample is more diverse.**

Thank you. We removed the reference to "diverse" and revised in the limitations section to:

> Our participants were exclusively from the US and recruited using an online platform, which is a limitation to generalizability (Simons et al., 2017). Follow-up studies may aim to rerun the same problems using non-US samples to explore the cross-cultural reliability of the mental accounting phenomenon. For instance, a follow-up mass collaboration project conducted by Priolo et al. (2023) was a promising attempt in examining the robustness of mental accounting across cultural contexts. In addition, we note that the data collection for this project was conducted during the COVID-19 pandemic. Though we found support for most studies, our participants may show different risk-seeking behaviors compared to non-pandemic periods. For example, Yue et al. (2020) argued that households altered their risk preference and became more risk-averse due to the pan

Response to Reviewer #2: Dr./Prof. Féidhlim McGowan

> **Thank you for the opportunity to engage with your research again. My Stage 2 review is below. I am also attaching a pdf with some minor comments, mostly on wording but also on some specific interpretations.**

Thank you for your thoughtful comments and suggestions.

> **The primary purpose of a Stage 2 review is to ensure that the pre-approved method and analysis plan was adhered to, and to assess that the interpretation of the results. Where fundamental flaws were missed at Stage 1, they can be highlighted in Stage 2; this is not relevant to this manuscript, where the reviewers' concerns at Stage 1 were taken onboard by the authors and appropriate changes made.**
>
> ***Method***
>
> **The authors provided a tracked-changes document. There are barely any differences from the IPA document, and where these differences appear they are incidental. For example, the additional changes to the Problem 10 design are minor and justified.**
>
> ***Results***
>
> **The pre-approved analysis plan has been followed. I will not comment on the significance of specific Problem replications because that is not the purpose of a Stage 2 review. The analysis procedure is sound and the description of results is comprehensive and transparent. Interpretation in the results section is kept to a minimum and deferred to the discussion.**
>
> **In terms of departures, The IPA stated that LeBel's interpretation criteria would be applied. In the end, the authors noted "we only compared the direction and relative magnitude of the mental accounting effects in some of the problems where it seemed to be meaningful." A partial application of LeBel's interpretation criteria is a sensible approach.**

Thank you. In this revision, we were able to deduce and calculate more effect sizes and confidence intervals of the original studies, and so even more effects are now compared using the LeBel et al. (2019). All are summarized in Table 16.

> **A minor suggestion is that the authors develop a Figure to illustrate the change in effect size between the replication of each Problem and its original counterpart. For readability, these could be ordered from most positive change to most negative, like in a butterfly bar chart.**

Thank you for the suggestion.

We tried different ways to address this request, yet, we found that all attempts were less than ideal and potentially confusing. The biggest issue is that the different studies had very different designs, from one-sample vignettes, to within-subject designs, to between-subject designs, some with interactions or comparing several measures. Therefore, there are many different standardized effect sizes that do not easily translate to one another - Cramer's v, Cohen's g, Cohen's h, Cohen's d, Cohen's dz, and eta-squared.

We hope that the very comprehensive Table 16 would be a decent alternative to your requested figure.

> **In terms of additional analysis – and feel free to ignore this suggestion – it would be interesting to see at the individual level how many participants could be described as 'mental accounters' i.e. their responses to at least 5 of the 9 Problems were aligned with Thaler (1999). In this study design, 9 problems were randomly chosen from a set of 18 for each participant, meaning there are 48620 possible combinations. This variation is helpful in attributing any 'mental accounting' tendency to the individual rather than there being particular sequences of questions that might amplify or nullify a tendency to display mental accounting traits. In fact, it is likely that few of the 1007 participants completed the same subset of questions.**

Thank you for this suggestion.

In our revision, we decided to not proceed with the analysis linking between the different problems, given the complexity of their different designs. We felt as though whatever analysis we tried the results were of very limited value and interpretability, and that these can possibly cause confusion and take away focus from the main point of revisiting the original studies. We therefore leave this to follow-up future studies:

> In this project, we aimed to systematically revisit experiments testing different accounts of the mental accounting framework reviewed by Thaler (1999). We focused on the empirical aspects of the singular problems, and did not go further to try and discuss implications for mental accounting theory as a whole. Therefore, the results of our replications for each of the problems should be interpreted separately and cautiously. We also did not address the mental accounting tendency at the individual level. Following the suggestion from one of the reviewers, we encourage future research to delve deeper in

this regard to undertake broader theoretical integrations and examine individual tendencies.

Discussion

The characterisation of the replication as being mostly successful is accurate in my mind. Also, the authors show admirable restraint in bounding the discussion of implications.

One suggestion is to add some specifics as to what avenues of research these results point to as holding promise. For instance, I was struck by the higher proportion of participants who selected the 'indifferent' response option. Indifference is the 'rational' prediction, so it would be theoretically interesting to pursue the causes of this change, if it occurs across different samples and contexts. However, the authors rightly warn against reading too much into specific differences between this replication and the original study, and perhaps I have just strayed into doing exactly that.

Overall, this research is a valuable contribution to the body of knowledge on mental accounting, and it has been conducted in adherence to best practice in open science.

Thank you. We feel as if there is already too much going on in this manuscript, and that we should remain cautious in our interpretation. As indicated above, we added a call for future studies to build on our findings and tie this back to broader decision-making theory.

"Also, it is unlikely that the differences in participant recruitment will have an impact on the results."

"these" and "had"

We removed that subjective comment, and only discuss the differences in sample as a limitation and a direction for future research.

"Problems 16 and 17 were powerful illustrations of the myopic loss aversion effect. Participants were more willing to take risks when there was a package of 100 bets (Problem 16), or a portfolio of 25 investments (Problem 17). When the risky episodes are bracketed together, people do not evaluate the events in isolation."

I would tone down the strength of this paragraph. A set of 100 coin tosses is also different because the probability of losing money overall is nearly zero, unlikely the single coin toss where it is 50%, even if the expected return on both sets is the same.

> **"To summarize, the hedonic editing hypothesis was only partially supported under the particular methodology and context."**
>
> **"this".**

Thank you. We overhauled our results and discussion sections, and toned that down to simply report the results in comparison to the original effects.

> **"Replication, therefore, is an important method to set limits on certain effects."**
>
> **I do not understand this sentence and my sense is it does not follow from the previous sentence. Consider revising.**

Thank you. We removed that sentence.

> **"The value of $5 in relation to transport costs has changed dramatically enough to shift participants' preferences entirely."**
>
> **This is extremely plausible. A cursory look at gas prices over time shows that it is currently over twice the price per gallon that it was in the late 1990s, and this increase is probably greater than the increase in wages.**

Thank you. We now discuss this in the section about Problem 2:

> Problem 2 examined people's perceptions of the value of time. The majority of the participants were unwilling to drive 20 minutes to save $5, regardless of the price. In hindsight, this makes sense, and raises an important dilemma in conducting replication studies - whether or not to update and adjust the prices to current days. As one of the participants in the feedback section pointed out, it may cost more than $5 to drive 20 minutes to the other store with the changing costs of driving and the devaluing of money since the problem was first presented in the 1980s. What would have been plausible to some participants in the 1980s, may no longer be plausible to most of current time participants. Therefore, the inconsistent findings may not be due to lack of an effect, but rather because of the changing circumstances. The differences in the findings could be due to the passage of time, and could also be due to our sample's demographics compared to the original's, or because of the unified design. The exact reason for the different findings is down to speculation. Future replications could further examine adjustments to the scenarios, or to assess different moderating factors.

> **“The pennies-a-day effect was validated in Problem 21. Within and across conditions, the "merely 27 cents a day" plan was rated as more appealing than the "100 US$ a year" plan. The price frames appeared to affect the comparability of the offers, where expressing the price on a per-day basis helps to lower participants’ price sensitivity (Chioveanu & Zhou, 2013).”**
>
> **Should probably cite Gourville (1998) which I think is the origin of the term pennies-a-day effect.**

Thank you, we added the citation in this section:

> We found support for the pennies-a-day effect in Problem 21 (Gourville, 1998). Price frames seemed to affect the comparability of the offers, where expressing the price on a per-day basis helps to lower participants’ price sensitivity (Chioveanu & Zhou, 2013).

> **“In addition, many review articles, especially when conducted by those whose studies it covers - like Thaler -”**
>
> **Try to reword this to make it clearer.**
>
> **“even if there was no empirical test associated with and testing those assumptions or predictions.”**
>
> **Suggest changing to "even if there was no associated empirical test", or something similar.**
>
> **‘at present’, seems to be a word missing.**

Thank you, we revised that section to the following:

> In addition to attempting a replication of original studies we added extensions testing predictions which did not have any reports of findings. Anecdotal evidence and untested assumptions and predictions are useful, as they provide ideas for future research to build on existing empirical studies. Replications and extensions of a review article can help tackle both aspects, by systematically mapping the studies reported as well as untested claims that can be empirically tested. We hope to see more systematic replications and extensions of impactful review papers, taking a similar approach to ours.

> **“We see much promise in further studies of the links among the different aspects of the mental accounting framework.”**
>
> **'links between' sounds more natural.**

Thank you. Revised to:

> We see much promise in further studies of the links between the different aspects of the mental accounting framework.

> **“We believe our reconstruction and reanalysis of classic experiments as well as our exploratory analyses could provide an inspiration and practical tools to stimulate further follow-up research to examine the mental accounting phenomenon as a whole.”**
>
> **Suggest changing to "impetus and practical guide".**

Thank you. Changed to:

> We believe our reconstruction and reanalysis of classic experiments as well as our exploratory analyses could provide an impetus and practical guide to stimulate further follow-up research to examine the mental accounting phenomenon as a whole.

> **Problem 10, Option 4: “I cannot understand this question.” Were the other answers of the people who answered 4 or 5 to this understanding question excluded? Is there an argument that they should be? Minor point.**

Thank you.

Out of the 502 participants, only 11 individuals responded with a rating of 4 or 5 for this particular option, and an exploratory analysis excluding them had no impact on the results.