Reply to decision letter reviews: #164

We would like to thank the editor and the reviewers for their useful suggestions and below we provided 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/gEEpjXsNbTSB

A track-changes manuscript is provided with the file:
"Thaler 1999 replication & extension-main manuscript-track-changes.docx"

Summary of changes

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

Section	Actions taken in the current manuscript
General	Ed: We tried our best to improve the overall coherence of the presentation. R1: We emphasized that the current project aimed at testing different parts of the mental accounting phenomenon and the results of each Problem will be analyzed separately. R2: We added more details to enhance understandability.
Introduction	Ed: We further elaborated on the Problems. R1: We included more details on the rationalization of each Problem. R2: We elaborated more on the overall mental accounting framework.
Methods	Ed: We responded to the methodological concerns. R1: We modified the power analysis section.. R2: We randomized participants to answer 9 of the 18 Qualtrics blocks.
Results	R1: We revised the Important Note at the beginning of the section and updated the data analysis of Problem 12.
Discussion	We added a paragraph discussing the limitations of the current project.
Supplementary materials	R2: We added an explanation on our inclusion criteria for Problem selection.

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

Response to Editor: Prof. Chris Chambers

> **The majority of issues focus on the rationale and level of detail surrounding the predictions and analysis plans, methodological concerns surrounding bias control and sample quality, and overall coherence of the presentation.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit. We really appreciate the time and effort that you and the reviewers have dedicated to the manuscript. We have taken the comments on board and modified the manuscript. Hope this updated version is satisfactory.

Response to Reviewer #1: Dr. Barnabas Imre Szaszi

> **First, I would like to thank the authors for their interesting submission. In general, I believe replications (and extended replications) are an important part of the scientific endeavor, and my general impression about the submission is that the authors did a thorough work and proposed a study with potentially important implications that is worth pursuing. Also please note, that I'm not an expert on the topic of mental accounting, so in my review, I cannot and won't focus on the specifics of the theory and its embeddedness into the literature but rather focus on other methodological issues. Below, I will try to provide several suggestions and depict some points which, I believe, require further clarification.**

We would like to thank you for the kind words and the thoughtful comments. We believe your input is invaluable in improving our manuscript. Please kindly find the below content a detailed point-by-point response to all comments, suggestions and questions.

General comments:

> **In this project, the authors aimed to revisit the classic mental accounting phenomenon to examine the replicability of 17 experiments from Thaler (1999). I have two major points to discuss:**
> **1. In contrast to classic replications where a (group of) researchers conduct the replication of one specific study, the authors of this study propose to conduct 17 replication. In my view, it is one of the strongest part of the study as such variance can provide a better understanding of the mechanisms and limitations of the mental accounting theory.**
> **However, my impression is that the authors need to provide a more detailed explanation of how they will use the data to make an inference about the theory. Here is what I mean:**
> **The authors claim that they are going to base the interpretation of data on "LeBel et al. (2019) outcome interpretation criteria - 1) signal / no signal, 2) consistency/inconsistency, 3) larger / smaller / opposite effect, by comparing replication effect confidence intervals to the original effect size". And that is great for individuals studies. However, in the abstract, their most important sentence and the one which refers to the empirical results refers to all the findings together. "On average, we found [weak to no / weak / medium / strong] empirical support for the mental accounting hypotheses.". It is not clear from the manuscript (at least I missed it) how they are going to summarize the results of the 17 experiments to make a general inference and create the sentence suggested in the abstract. In that**

sense, the authors' proposal is greatly alike to a meta-analysis, as they want to make a general statement about a specific effect in general, based on the outcome of several experiments, but it is not detailed how they are going to do the "meta-analysis" itself just on how they are going to get the effects that go into the meta-analysis. If this is the case a more detailed elaboration of this process is needed.
Additionally, it would be important to discuss in the manuscript, that if the authors will find a meaningful variation in the effect of (empirical support for) mental accounting hypothesis, why do they think they find this variation. Is it because these experiments test a different part of the theory, or the variation may come from other methodological differences between the Problems? If the authors think that the experiments test the same general theory and not different parts of the mental accounting theory, why don't they merge all the data into one model in some way? If they test different parts of the theory, it would be great to see how what we can learn from them. (Maybe these questions are hard to answer but I think that having a clear answer on these would highly increase the scientific value of the manuscript.) Alternatively, the authors could interpret each experiment separately. In that case, both the methods and the theory for each study (Problem), should be discussed in more detail - although my sense is that this is not what they are aiming for.

Thank you for the comment and suggestions. We appreciate the opportunity to elaborate further and add information.

These collections of experiments were intended to test several different aspects of the mental accounting theory, and an attempt for a theoretical integration would be too complex and of scope much larger than what we intended here. We agree that this would be of value, and we commented on this as directions for follow-up research.

In the Abstract, we changed the summary sentence into:

> *"Out of the 17 mental accounting hypotheses, we found empirical support for X with effect sizes ranging from X.XX [X.XX, X.XX] to X.XX [X.XX, X.XX], and no empirical support for Y with effect sizes ranging from X.XX [X.XX, X.XX] to X.XX [X.XX, X.XX]."*

In the Results section, we address and summarize the results of each of the Problems separately. In our initial Registered Report manuscript we already included tables with summaries of both the descriptives and the statistical tests of these problems.

We will devote a section of the Discussion section to summarize the overall findings and then discuss possible insights from and future research looking into the variations in the effects.

Given that the different experiments use different methods and designs, aggregation across these is complex and goes beyond the scope of this project. That said, we are making all our data and code publicly available for anyone who would like to build on our work to gain additional insights. We added this as a suggestion for future research in our discussion's limitations section.

We have added further details for each Problem in Table 1 and Table 3.

> **2. Another general point is that the authors could do much better work at making the manuscript more streamlined and understandable for the readers. As you will see in the details of the more specific points below, there are several things that are hard to understand.**

This is indeed a challenge, especially in this specific project in which we cover so many different problems. On the one side we are committed to full transparent reporting, and on the other side we strive for conciseness and clarity. We appreciate the suggestions for helping us strike a better balance at that and made adjustments accordingly, yet decided not to introduce major structural changes.

Specific points:

> **The authors claim at several places in the manuscript that they conduct independent replications. I would suggest not calling these independent replications because all of the studies will be finished by the same 800 participants. If the authors want to use the word independent, I would recommend the authors to provide a more detailed explanation on why do they think it is independent.**

It is important that you brought this up, because we realized that may be confusing to some and that we needed to clarify our use of the word. The word independent was not meant that each of the replications are independent of one another, but rather that we are researchers who are acting independent from the original researchers and labs of those who originally published these findings. An independent replication is a replication by external people independent from the original author's lab. We reframed "independent replications" to "with replications by an external independent team." (subsection "Choice of article for replication: Thaler (1999)").

> **In the first pages of the manuscript, the authors claim at several places that they will conduct a close replication. At the end of the manuscript do a very nice job explaining why do claim that, but it would be great to discuss shortly why they claim that, at least refer to the LeBel et al. (2018) the first time they claim this.**

Thank you. We added the citation of LeBel et al. (2018) when we first claimed that it is a close replication.

> **It is not consistent and very hard to follow on what kind of data collection and analysis have been completed already and what is missing. Please go through the text and try to describe this in a clearer way. At some point it is written, that data collection was completed before analyses, suggesting that data collection is already completed. (Is it?) At some point you write that data collection was completed in March. Based on some parts, my understanding is that this part may only refer to 200 individuals? Or was the data only simulated by qualtrics? In other points, the text suggests that you will recruit participants from Amazon Mechanical Turk later. I'm sure there is some simple explanation but it is very hard to follow. Please clarify.**

Thank you for these comments, and we appreciate the opportunity to clarify these. Above the abstract, the method, and the results section we previously wrote:

> "**[IMPORTANT:**
> **Method and Results sections were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection.]**".

We therefore wrote the entire manuscript as a simulation of what the manuscript is meant to look like after data collection, and based on simulated data. We realize that this could be confusing, especially when we switch tenses and insert specific dates, and we amended that in the revision.

To be clear - no data collection or pre-registration has taken place yet, and these will only take place following receiving in-principle acceptance from Peer Community in Registered Reports.

The 200 participants described are random data points stimulated by Qualtrics, which we used in order to demonstrate our data analysis plan and intended to show what the results section will be like.

Our future data collection process we will use MTurk.

We further explained these issues about data collection and analysis at the beginning of the Method and Results section in the revised manuscript.

We also amended our warning to:

> "**[IMPORTANT: Method and results sections were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection. This is written in past tense yet no pre-registration or data collection have been conducted.]**"

> **The choice of study of replication sounds well-grounded, although I again have to admit that I'm not in the best position to make this call as do not know the alternative studies that could have been replicated. However, my impression is that it is strange to talk about direct replications and argue that " We chose the Thaler (1999) article based on three factors: extensive academic impact, absence of direct replications….", as the Thaler (1999) paper was a review paper and not a primary paper. As far as I know, researchers do not replicate review papers but rather the primary experiments, so the question is whether the individual studies have been replicated before. (Although note that your answer on this point might be influenced by your answer on my general comment 1).**

Very good point, thank you, we agree and appreciate the comment.

Researchers mostly focus on replicating primary experiments rather than a series of experiments covered in a review paper. In our case, we meant to say that there was no systematic attempt to replicate findings from the mental accounting literature for most of the findings reviewed in Thaler (1999), and that to our knowledge there were no published direct pre-registered well-powered replications of Thaler's own work.

We amended the framing to be more clear, and added information in Table 1 to also include the individual citations impact of each of the studies.

> **To me, the introductory part has unnecessary parts (e.g., "Trepel and colleagues (2005) even extended the mental accounting phenomenon into the field of neuroscience, where they outlined possible neural bases for Kahneman and Tversky's prospect theory". (p. 10). On the other hand, It would be helpful for the reader to have a better build intro about what we are going to learn from this study, and why do we need the study.**

We agree that more elaboration on what people are going to learn from this project will be helpful. We also deleted the reference to Trepel and colleagues' work.

We briefly addressed the point of the need for replications of impactful work that has not previously been replicated, assuming a generalized agreed understanding of the need for

replications and what we learn from replications when they succeed or fail. We added a brief reference to our need to update our knowledge of the phenomena and the experimental effect size in the context we used.

> **On page 16 you write that "Note. We are unsure about the statistical results reported in Problem 6-Condition A-5 as they seem to add up to 110%". Will you include this item as well in the analysis? My intuition would be not to use it as we cannot be sure how it affects the choices.**

Thank you for highlighting this. We realized this note was not clear enough and could be misleading. Our intention was to point out the possibility that statistical results reported were inaccurate (potentially a typo). We have now changed it into *:*

> *"The statistical results reported in Problem 6-Condition A-5 added up to 110% rather than 100%, suggesting a reporting mistake in the original article."*

We will still include this item in the proportion test analysis, but will not compare the replication effect with the original effect.

> **You write that: "We extended the replication of the experiments reviewed by also adding a test of four predictions that Thaler (1999) reflected on but have not been directly tested or shown empirical evidence for." I would frame it with a less certain language such as "we are not aware of any empirical evidence testing ...".**

We appreciate the suggestion. We have changed the sentence into:

> *"We extended the replication of the experiments reviewed by also adding a test of four predictions that Thaler (1999) reflected on. We are not aware of any empirical evidence directly testing these predictions."*

> **It is not clear to me what the authors mean by predictions of Thaler (1999). It seems to contradict to me to the fact, that in Table 3 each of the predictions is described by references to prior empirical papers. Please clarify if these predictions were tested or not in previous research.**

We appreciate the opportunity to clarify and elaborate.

The papers listed in this table are the source of the predictions and the four predictions were not previously tested to the best of our knowledge. We added a note explaining this at the end of the table for clarification.

> *"The papers listed are the sources of the predictions and none of the predictions have been tested directly to the best of our knowledge."*

> **Also, it would be needed to provide a description in the intro on what exactly these predictions are and why are they important. Please also connect this part to your reply on whether you see all these experiments (Problems) as independent tests of the same theory or different parts of the theory. Although these extensions suggest that they are additional parts of the theory, so it is not clear to me how you plan to integrate these findings into the whole picture about mental accounting theory. Again, I'm not claiming that it is problem, but that it is not clearly described in the manuscript.**

Thank you for raising these.

We added brief details regarding the predictions in Table 3. For the exact predictions we previously referred the readers to the supplementary materials section *Instructions and experimental material.*

These extensions are different parts of the theory and we will also interpret the findings separately. We now also address this issue in the Abstract too.

> "The replications and extensions examined different parts of the mental accounting theory and the results were interpreted separately."

On page 17, you write that "For each of the replication problems, we largely followed the original experimental design and only changed the questions to make them up-to-date and suitable for our targeted participants.".
It would be helpful if you could provide the summary in the supplement on where and what had been changed exactly. This would be necessary to review this manuscript without the necessity to review the 17 other papers as well. Later (now) I found this information in Table 8, which is great, but please add a reference about Table 8 when you talk about this in the manuscript earlier.

Thank you for the suggestion. We updated the reference in the revised submission:

"We summarized the changes in Table 8."

On page 21, you have Table 4 writing that it summarizes the "Differences and similarities between original studies and replication", but to me, it seems that it doesn't have this information.

The differences and similarities in this table were about the sample and the data collection process. We changed the title of the table into:

"Summary of samples in the original studies and our replication".

All the other methodological changes are detailed in Table 9 and described as deviations in the revised manuscript.

You write that you have pre-registered and provided all materials, data, code for all studies on OSF: https://osf.io/v7fbj/. However, when I click on the the Registrations tab, it says that "There have been no completed registrations of this project.". Here:https://osf.io/v7fbj/registrations. Could you help me with that? Maybe just a technical issue.
You write that "We first pre-registered the study on the Open Science Framework (OSF) and data collection was launched later in March.". Which year? See my comment also about confused description of data-collection above.

We appreciate you pointing this out. The pre-registration and data collection was not yet conducted. These sentences were meant to simulate a future manuscript and express that these processes will be done before the actual data collection.

> **"To ensure that the current replication sample has sufficient power, we calculated effect sizes and power based on the statistics reported in the original experiments. For the replication studies, Rstudio was implemented to perform power analysis, where alpha (two-sided)=0.05 and power=0.95 were used. Results of the power analysis suggested that the minimum required sample size for a power of 0.95 and alpha of 0.05 is 321 participants."**
> **Maybe I'm not well trained enough but based on what effect size did you get 321? Is it the same for all the experiments? Please clarify.**

The details about the effect size and power analysis were provided in the supplementary materials and a very detailed power analysis script was provided in the OSF.

We appreciate the opportunity to elaborate on this. The stated need for 321 participants was based on the power analysis of all problems and relying on the largest required sample size of all the problems, which in our case was for Problem 15.

We adjusted the text and added more explanations to be clearer:

> "The largest required sample size was 321 participants, indicated by the power analysis of Problem 15. Therefore, we concluded that the minimum required sample size for a power of 0.95 and alpha of 0.05 is 321 participants."

> **"...and multiplied 321 by 2.5 resulting in 800 participants, to ensure sufficient power...." It is 802 and not 800. My problem is not that you did not use a sample of 802, but that the sentence is mathematically incorrect. Please modify.**

To address feedback given by the other reviewers, we modified this part completely into:

> "Given the possibility that the original effects are overestimated, and taking into account the issues of multiple comparisons and potential exclusions, we aimed to recruit 500 participants. Given reviewer's feedback, we decided to make a change in our implementation so that each participant will be randomized into 9 of the 18 Qualtrics blocks, aiming to cut survey time by half. The implication is that the actual sample for each of the Problems would be on average about half of what we previously intended. To compensate for that, we doubled our overall sample to 1000. A sensitivity analysis indicates that we would be powered to detect effects of f = 0.17 (groups = 3, df = 1) and d = 0.29/0.36 (between, 250/166 in each condition) (both 95% power, alpha = 5%, one-tail), which are effects much weaker than any of the supported effects in the reviewed studies."

> **Another thing regarding the sample size estimation is that you write that "A sensitivity analysis indicates that a sample of 800 would allow the detection of f = 0.14 (groups = 3, df =1) and d = 0.23…an effect much weaker than any of the effects reported in the target article.".**
>
> **However, for several reasons. I'm not 100% convinced that 0.23 is not still an overestimate. 0.23 is a large effect in general in psychology. The effect sizes come from an era of science when p-hacking, cherry-picking were not even a discussed issue, and also publication bias could inflate these previous effect sizes.**

We revised our statement to be a bit more precise:

> "which are effects much weaker than any of the **supported** effects in the reviewed studies.".

Sidenote: From Cohen (1988) to the more recent and social-psychology specific Lovakov and Agadullina (2021) a large effect for Cohen's *d* is 0.65-0.8, a medium effect 0.36-0.5 and a weak effect 0.15-0.2. A Cohen's d of 0.23 is considered a weak effect (35th percentile in the entire literature). The newly targeted 0.29/0.36 is considered a weak to medium effect.

The effects reviewed in Thaler are considered some of strongest most robust effects in the literature, and our experience with many similar replication attempts of the judgment and

decision-making literature is that the JDM literature has high replication rates (currently 68% successful) and - on average - about equal effects to that in the original.

References:

Lovakov, A., & Agadullina, E. (2021). Empirically derived guidelines for interpreting effect size in social psychology. *European Journal of Social Psychology*.

> **Last, but not least, my prediction is that the design applied in the present experiment results can result in smaller effect sizes, as compared to the original studies where participants only had to complete one Problem, here they are asked to do 17. As a result, they are going to experience fatigue which can have a negative effect on the expected effect size. All in all, I do not know what the proper number is but I would not be surprised if with this sample size we still could not find an effect.**

Thank you. As we outlined in our introduction, we have had quite a bit of experience testing out such paradigms before, with successful implementations. For example, see our replication of Kahneman and Tversky (1972) in Wan and Feldman (2021) (https://osf.io/r4h6s/) and our replication of Heath, Larrick, and Wu (1999) in Au and Feldman (2021) (https://osf.io/szdfw/). Both of those were concluded as mostly successful with comparable and sometimes stronger effects than those reported in the original. In the example of Wan and Feldman (2021) we were able to identify groupings of the problems that worked better compared to others and derive theoretical insights for follow-up research. These insights would not have been possible without running all these in the same data collection.

The concern you raise points to an additional benefit of our design. If needed, if we fail to find support for some of the studies, we could examine order effects, and this would actually provide for an exploratory interesting insight as to how fatigue is associated with responding. These are exactly the type of tests that combining so many experiments into a single paradigm would allow us to test. Given that we are making all of our data available, anyone interested in such questions would be able to conduct further exploratory tests.

> **A few things are hard to understand in Table 4: What do you mean by original studies and replication? What does it mean that the current replication is only 200 individuals?**

We realized that the way we framed this may have been confusing, and appreciate the opportunity to improve.

The 200 “individuals” were random noise participants that we simulated using Qualtrics for the purpose of demonstrating to reviewers what the results would look like after data collection. The

column regarding our replication will be updated to the actual real numbers following the Registered Report in-principle acceptance, and real participants data collection.

We revised the title of this table to *"Summary of samples in the original studies and our replication"*.

We added further elaborations at the beginning of the Method section (the paragraph underlined and bolded).

> **Table 5 is also very hard to follow and interpret. I think that it would be even easier if there would be written one short paragraph about each Problem.**

Thank you for the suggestion. We tried different ways to summarize these designs, including writing out the design, and concluded that a table summary would be a clearer, more efficient summary of the designs.

That said, we thought more about how we can improve on this table, and made some adjustments that we hope would help make the designs easier to understand. We hope our changes have helped.

> **You write that "Scenarios were presented in random order and participants were randomly and evenly assigned into different conditions. This method was previously tested successfully in many of the replications and extensions conducted by our team".**
> **What do you mean the method was tested successfully? How did you measure the method as success? Maybe it would be better just to remove this sentence or detail how it increases the validity of the present replication.**

What was meant was that we have previously used this design in replications, and found that order or number of problems included did not result in replication failures. We appreciate the note and added the following clarification:

> "Our findings from similar projects using a similar design suggest that combining several experiments in a single data collection in random order did not impact likelihood of replication success."

> **I have noticed in the Qualtrics that only, native English speakers born, raised, and currently located in the US can participate in the survey. I couldn't find this in the method section. If this is the case, please add it.**

We now added this criteria in the Method section:

> "In the actual data collection, 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)"

> **A note: when I reviewed the experiment itself, I have noted that it was hard to understand the Mr. A and Mr. B Problem in the survey. (Where the participants are asked to make a choice about two events together vs separately). If you have any way to improve it's understandability, please do it.**

We now revised the instruction of the problem to this to enhance understandability:

> "Below you will find three pairs of events. In each case, the same events occur either on the same day (for Mr. A) or two weeks apart (for Mr. B).
>
> You are asked to judge whether Mr. A or Mr. B is happier, or in the case of two negative events, who is more unhappy. If you think the alternatives are emotionally equivalent, select "no difference."
>
> (Note: You are only asked to judge whether it is better to have the events separately or together)."

> **Maybe, it is because I'm not a native speaker but in the "Previous events and new payment" task, it is not clear to me what is the question. Whether I would by the ticket later or SOONER, or whether I would be a ticket at all?"**

Thank you, we agree this could be improved and appreciate the feedback.

We revised the instructions for Problem Q11A/B :

> "Each of the questions below asks you to imagine that a specific event took place at the beginning of the week and whether based on that you would make a purchase later in the week."

> **I'm not convinced that calling previous experiments "Problems" is the best solution. In my view, how you call it should also reflect how do you look at theoretically at each of the studies (if they are just "items" measuring the same construct, the wording Problem can be Ok, but if these measure different things and are separate experiments, I think it is not). Additionally, this should be consistent with other parts of the paper. E.g., in the title you refer to these as experiments.**

We understand this concern. "Problem" is the wording Thaler used in his articles, and so we decided to follow in his footsteps so as to avoid any confusion.

We agree that the inconsistent use may cause confusion. We changed references to these to be "Problem(s)" throughout.

> **You write on page 16 that "Please see Tables 4 and 5 for a summary of all problems and manipulations.". I think Table 4 is not relevant and table 5 does not include the problems.**

Thank you for catching that. Indeed, the original Table 4 is not relevant here. We now refer from Table 5 to Table 11 for the different manipulations and modified and improved Table 5.

> **You seem to use the words manipulations and experimental design (to me) in a strange and confusing way. For example, you write that "Problems 1, 2, 3, 6, 7, 8, 9, 11, 12, and 21, involved manipulations, and participants were randomly assigned to conditions separately in each of those."**
> **However, I think their within-subject design also involves manipulations, just not between subject manipulations. And within-subject experiments are also experiments.**

Thank you for highlighting this, great point.

We changed this sentence into:

> "Problems 1, 2, 3, 6, 7, 8, 9, 11, 12, and 21, involved between-subjects manipulations, and participants were randomly assigned to conditions separately in each of those."

> **You write that "In the actual data collection, we will categorize values more extreme than 3 standard deviations around the mean as outliers (Leys et al., 2019). Outliers would be classified as either error outliers or other outliers (Leys et al., 2019). For error outliers, outliers due to wrong data entry, we will check up the raw data to see if corrections can be made."**
> **I have some questions regarding these processes: How are you going to decide whether an outlier is an error or "other outlier"? Do you have any theoretical reason to exclude "other outliers", why are they not part of the natural distribution?**
> **Maybe I'm wrong, but I think that with dichotomous and not numeric answer options, the mean+-3SD is not a suitable method, as you do not have a mean, and for many, the Problems as people had to choose from two scenarios. That saying the proposed outlier exclusion method cannot be used for many of the problems.**

Great point. We amended this section and will not further classify the outliers.

Also, we explained that the mean +- SD method is only applicable to numeric answers.

> **To me, it seemed that you do not report an effect size measure for the proportion test. If I'm right, could you do that?**
> **It would be laudable from the reader's point of you to use the same effect size measure for each of the problems. I understand, that it is not plausible, but at least creating a common effect size measure would help get a sense of the comparison of Problems and the overall results.**

This suggestion is appreciated, and we could see the appeal in trying to provide one overarching conclusion for all effects together, but this is very tricky to implement. Conversions and grouping different designs into a single comparison is very difficult to impossible to do well and may lead to confusion, misunderstandings, and under/overestimations of effects. We did try and address this by grouping the problems of similar designs and effects, in hope that this would allow readers a more simplified overview of several of these problems together.

The last point: In general, I really do not like (trust) MTurk samples. At least, I had very bad personal experiences with them. Although I note that on page 19, the authors have employed a great number of countermeasures against low-quality responses, -which is great! - still, it is possible that any lack of effect, or smaller than the original effect would arise from the fact that the data come from an MTurk sample. Beyond discussing the potential limitation of the sample, If the authors have the possibility, I would strongly encourage them to collect data from other sources/populations/means as well. My feeling is that doing so would exponentially increase the potential of the paper to become to be influential in the field as this way it could provide a much stronger feeling about the generalizability and empirical support of the mental accounting hypothesis. Otherwise any reader with similar experience to mine, and I know that I'm not alone, can easily disregard the results thinking that "this is just another MTurk' sample study"]

We were sorry to hear of your personal experiences with MTurk. You did not specify the exact concerns you have so it is difficult for us to address personal experiences and a general dislike for one of the most commonly used platforms in social psychology.

Our experience is completely different, and we have a lot of evidence to show the reliability of this 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 have 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 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

As you pointed out, we indicated that we will be using a lot of measures to ensure we have high-quality data and we are running this on a very specific tested sub-sample of MTurk using the platform CloudResearch and we employ many other quality checks. We have successfully implemented those before many times and have seen time and time again that they perform well.

> **Again, thank you for the interesting submission. I really hope that my comments could be used to improve the paper. Looking forward to reading the responses and the revised paper**

Thank you again for the careful review and insightful suggestions for our manuscript. We revised our manuscript accordingly and believe that the manuscript is better and stronger thanks to your suggestions.

Response to Reviewer #2: Dr. Féidhlim McGowan

> **This Stage 1 registered report outlines a plan to replicate many of the experiments in the Thaler (1999) review article "Mental Accounting Matters". The authors pre-register the method and analysis plan in line with best practices in open science. The power analysis is comprehensive and so too is the procedural aspect of the registered report. I commend the authors in particular for making the experiment available to pilot as part of the review process.**

Thank you very much for the positive opening note.

> **However, I have serious concerns about the precision of the hypotheses, and the apparent gap between what the replication does and the proposed interpretation of the results. My primary issue is with the intended broad scope of the project, which is not matched by an analysis plan that can harness the repeated-measures nature of the data. The reviewers note they have already narrowed the scope following feedback (removing an investigation of impulsivity as a mediating factor), but further extensive narrowing would improve the scientific value of the project. My overall assessment of the current report is that it fails to meet the Stage 1 criteria.**

Thank you for the constructive feedback to help us improve our manuscript. We did our best to address the comments raised.

> **The guidelines for registered reports states that "The Introduction section of the Stage 1 manuscript should make clear the underlying theory or application from which the question arises, leaving the reader in no doubt as to why the study is being proposed."**
> **The authors justify replicating the studies cited in Thaler (1999) based on the number of citations it has received, and the lack of previous comprehensive replications of the studies it cites.**
> **I was unconvinced by the summary of mental accounting. It seemed to draw too heavily on the abstract in Thaler (1999) with no reference to weaknesses or issues with the mental accounting framework that have come to light in the intervening decades.**

Our project is a very specific type of Registered Report - a replication. The aim of replications is to focus on the empirical reproducibility and replicability of the findings, and much less so on theory. We have included a brief overview of mental accounting, yet this is a vast literature and it is not in the scope of this project to cover those. The replication project is empirically very

complex and going into the literature and theory of mental accounting is not in the project scope, may confuse readers, and will likely take away focus from its intended goal.

In our revision, we added a bit more to our introduction to elaborate further on the mental accounting phenomenon and citations in an attempt to cover a bit more of the framework.

> **Thaler's (1999) review paper covered a long list of classic mental accounting experiments, and the current replication proposes to test 17 of them. How were this 17 chosen? Did the authors parse the paper for experiments that did not report their sample size? This would be a useful approach, even if it picked up 'experiments' that were only ever intended to illustrate an idea by way of example (I am thinking of Samuelson's coin flip thought experiment to his economist friend). There are many other experiments in the paper (For example, Wertenbroch (1996) who found that the price premium for sinful products in small packages is greater than for more mundane goods is particularly interesting and it would be interesting to see how much it generalizes.) so some rationale should be given for how this sample was selected..]**

This is an important point, thank you.

In our revision, we included our rationale for problem selection in the Supplementary:

> In the targeted article, Thaler (1999) covered a wide array of mental accounting studies. In the current project we focused on problems that were simplified in design and were suitable for administration online with our target sample of labor market.
>
> An example for excluded studies is the study by Simonson (1990). In this study, Simonson assigned students were asked to either 1) select among six snacks at each of the three class meetings held a week apart, or 2) select three snacks at the first class meeting to be consumed later every week. Such a study design cannot be adapted to online questions.

> **In the report snapshot, the RQs were summarisied as "Research question: 1) Do people engage in mental accounting activities? 2) Are there links between and a consistency among the different mental accounting behaviors? [my italics] After reading number 2, I was expecting to se a planned pooled analysis, for example using a linear mixed model. I come back to this briefly in point 2 and again in point 6.**

This is a good point, and we realized that we should have made this clearer.

A pre-registration or a Registered-Report aims to focus on the confirmatory analyses and clear hypotheses which in a replication are focused on the comparison between the original results and the replication's findings. The links were meant as exploratory directions with no clear hypotheses or tests, and so we did not elaborate these for the Stage 1 submission.

We were planning on an exploratory pooled analysis, and in the revision included more details on our plans to pursue that.

We will further respond to this at point 6.

> **More broadly, the introduction did not convince me that the replication would make a precise contribution. For example, in the study design on page 7, the authors write the following under the column titled "Theory that could be shown wrong by the outcomes": The mental accounting theory (e.g. the framing effect, prospect theory). This was surprising to me, because it appears the authors are suggesting mental accounting and prospect theory are synonymous, or that prospect theory is somehow derived from mental accounting, which is clearly not the case.**

Yes, we agree that our use of that column was not correct. We admit to being a bit confused as to how to address that column in a replication study, especially one as complex as the one we embarked on here.

We removed that column.

We are unsure how to address your request for a "precise contribution", yet for each of the Problems we replicate we will be able to conclude whether we found support for that problem looking at a specific phenomenon or not, and that is a very precise contribution to each of those.

Overall, we would be able to summarize the number of problems in that article that we found support for. We agree that this does not reflect on a whole theory but rather only on the summary of that theory that we aimed to replicate.

We adjusted our writing to reflect this point and clarify further.

> **2. Randomisation of Questions**
>
> **The authors justify the use of randomising the order of questions by reference to a previous successful replication of the Kahneman and Tversky study on representativeness heuristic. However that replication concerned one specific cognitive shortcut for probability estimates. The current study aims to replicate different pillars of a framework for making financial decisions. No in-depth consideration is given to order effects might arise with relation to answering different questions that relate to the**

> **same pillar of the framework, for example the desire to be consistent in one's responses. Alternatively, a participant might identify a fallacy in their thinking and correct it on a subsequent question. These order effects problems will add noise to the data, and you cannot control for it because there are too many different orderings (literally millions, whatever 21! is equal to).**

We gave this design a lot of thought, and have successfully implemented this in many of our replication projects. We consider this a strength of this investigation. We replied to a similar concern raised by the other reviewer regarding the inclusion of many problems in one investigation.

We do not expect any order effects given our vast experience with similar replication projects before. Still, should there be a need for order analyses to investigate things further, there are many approaches to that which can be as simple as relying on the display order as a continuous measure. This design is exactly meant to allow for many analyses to help us gain additional insights as to when an effect is likely to replicate.

We do not think participants inferring or correcting from one problem to another is a concern. We have not seen that in our other projects, and we think that would be highly unlikely. The mental effort and understanding of the materials and phenomena required to make inferences regarding consistency, logic, and accuracy across these problems is extremely high, especially given that no feedback is provided, and our target samples' participants' goals in completing these tasks fast to receive compensation.

> **Is it really necessary to ask so many questions?**
> **For example, Problem 17 is an applied version of 16. There is little reason to test both in a randomised order. A trade-off exists between the quantity of questions you ask (in this case to probe mental accounting operations) and the quality of answers you can expect to receive. A review paper is not intended to be replicated in its entirety.**

We understand this concern and we appreciate you raising this. This can be taxing on participants, and yet our target sample do surveys like these for a living, and we've seen participants do much more demanding tasks for much longer.

Tasks like Problem 16 and Problem 17 are very short and expected to take 15-30 seconds, and so there is little to lose and much to gain from including both of those in the same randomized order setup.

This setup is meant to allow us to test whether the applied Problem 17 is consistent with the similar Problem 16, and rather than assuming a similar response pattern, we would be able to test that link.

We consider our setup appropriate and compensate participants in accordance to length.

To address your concern, we also made adjustments to the length to ease load on participants.

> **An analysis plan to do a meta-analysis (see point 6 below) to come up with a score of one's "tendency to violate fungibility", for example, would have been interesting and a clear contribution.**

In this replication, we have many experimental designs and separate summary effect-sizes, and all the effects are from the same participants, and so a meta-analytic summary would be challenging and difficult to interpret. We will attempt to summarize the results clearly for each of the designs.

> **Another issue is how each problem is presented– what is the benefit of evaluating two questions on same page? It is more likely that people will spot errors in their thinking (e.g myopic risk aversion in Problem 16) when presented side-by-side with an altered version of the same scenario.]**

We tried the best we could to emulate the setup described by Thaler and the summarized studies, which in many cases displayed many of these in the same page. Many of these Problems were presented in a form of a pen and paper in which participants saw all the Problems, and were able to compare and re-evaluate.

For within-subject designs setups like Problem 16, especially those with a similar repeating scale, the benefit of including those in a single page on the same scale is to reduce cognitive load

on the participants and allow them to use the scale in a similar way without having to re-read and readjust scale use. We consider this design an advantage.

3. Power analysis:

"To ensure that the current replication sample has sufficient power, we calculated effect sizes and power based on the statistics reported in the original experiments. For the replication studies, Rstudio was implemented to perform power analysis, where alpha (two-sided)=0.05 and power=0.95 were used. Results of the power analysis suggested that the minimum required sample size for a power of 0.95 and alpha of 0.05 is 321 participants."
I had to dig into the Supplementary to see how the power analysis was actually done. The supplementary material is very detailed and that is good! It clearly states that the smallest effect size of the 17 problems was used to calculate the sample size of 321, and this was then multiplied by 2.5 in line with the rule of thumb. Some of the problems required a sample size of less than 50 people for adequate power. A more sophisticated randomisation method that branched participants between-questions in the desired ratio way would offer big efficiency gains. This is something Qualtrics can probably do, as I know other experiment builder software like Gorilla can do it. This is not a criticism, just a suggestion that with some refinement the authors could make their research funds go a lot further.
As an aside, branching that reduced the length of the experiment would also reduce the noise in the responses. When I piloted it, it took me nearly half an hour to complete, and I found it difficult to keep imagining very different scenarios, some of which were described in very little detail (for example the portfolio decision when managing a division).

This is very valuable feedback, we really appreciate it.

Thank you for taking our survey and giving us some indication of how long it took you. Half an hour is much longer than we intended, and our pretest estimates were for about half that time.

Based on your feedback, we changed our design to the following:

> "Given the possibility that the original effects are overestimated, and taking into account the issues of multiple comparisons and potential exclusions, we aimed to recruit 500 participants. Given reviewer's feedback, we decided to make a change in our implementation so that each participant will be randomized into 9 of the 18 Qualtrics blocks, aiming to cut survey time by half. The implication is that the actual sample for each of the Problems would be on average about half of what we previously intended. To

compensate for that, we doubled our overall sample to 1000. A sensitivity analysis indicates that we would be powered to detect effects of f = 0.17 (groups = 3, df = 1) and d = 0.29/0.36 (between, 250/166 in each condition) (both 95% power, alpha = 5%, one-tail), which are effects much weaker than any of the supported effects in the reviewed studies."

4. Extensions not Novel (as described)

The abstract states "Extending the replication, we provided an initial test of four predictions not previously empirically tested that were described in Thaler's (1999) paper as predictions."
Taking up this point on page 8, they state: "Our second goal was to examine several predictions made by Thaler regarding mental accounting behaviours that have not previously been put to a rigorous empirical test." One of these predictions is that framing the cost of services per day will be more attractive than a per year framing. However, this question has been extensively investigated in the marketing literature. Some of this work was contemporaneous with Thaler (1999), for example the Gourville (1998) pennies-a-day effect, which has received hundreds of citations. While novelty is not required, it is important to keep up to date with the state-of-the-art.

Terrific feedback, thank you for pointing us to the relevant literature. It is possible that we overlooked some follow-up literature.

We included Gourville's (1998) work as a citation to clarify that this concept has previously been examined in the marketing field.

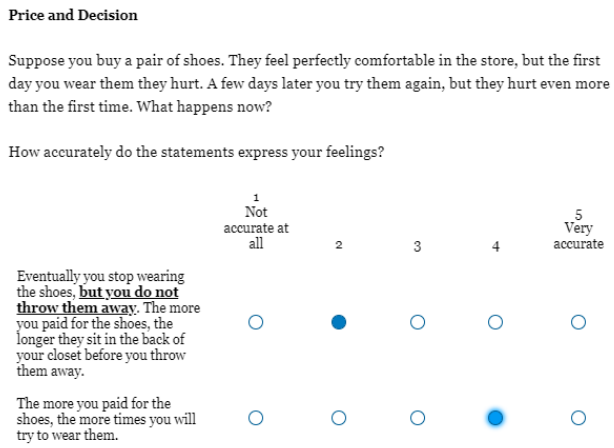
We also were more careful and humble in referring to the existence of follow-ups using "as far as we know".

Also, the design of this question on year vs. daily price framing (Problem 21) is poor. The third condition, in which participants rate the attractiveness of the per-day frame and the per-year frame on the same page, should be in a 2x2 design with two different services (for example streaming TV and music). Then the equivalence would be less obvious or could even be removed altogether while allowing the effect of the frame itself to be isolated.

The purpose of adding Condition C was simply to turn the between-subject design of Conditions A and B into a within-subject design allowing us a comparison of the two effects. The purpose of this extension is exactly to test whether a joint display would have a weaker/similar/stronger effect compared to separate conditions contrast. This mirrors some of the JDM literature in

phenomena like “less is better” testing differences for joint versus separate or within versus between.

> **Another proposed extension (Problem 20) relates to testing whether the half-life of the sunk cost fallacy depends on the price of the item. I attach a screenshot of the text for the question below. Looking back to Thaler (1999), it becomes clear that the text of the question is taken directly from Thaler’s description. This is not a test of the prediction. Instead it is a test of how much the participants agree with the prediction, which is something different entirely. An actual test would involve independent manipulation of the price of the shoes followed by an elicitation of judgments about how many times a participant would try to wear them (or reckons the buyer would try) and how long they would wait before throwing them out. (for example)**



Price and Decision

Suppose you buy a pair of shoes. They feel perfectly comfortable in the store, but the first day you wear them they hurt. A few days later you try them again, but they hurt even more than the first time. What happens now?

How accurately do the statements express your feelings?

	1 Not accurate at all	2	3	4	5 Very accurate
Eventually you stop wearing the shoes, **but you do not throw them away**. The more you paid for the shoes, the longer they sit in the back of your closet before you throw them away.	○	●	○	○	○
The more you paid for the shoes, the more times you will try to wear them.	○	○	○	●	○

Yes, we agree this needs to be clarified. This is a test of lay-intuitions regarding self-behaviors in the specific situations that Thaler described.

We added a note at the end of Table 3 indicating as much:

> “For Problem 20, we aimed to examine how much participants identify with Thaler’s prediction.”

> **5. Classification as close replication**
>
> **The authors cite the LeBel (2019) framework and use its checklist to classify the proposed replication effort as a “very close replication”. This may be technically correct within the Lebel framework, but it does not seem accurate more generally. This seems unwarranted to give one label of**

> **"very close replication" to 17 experiments, given the variations within the experiments. The experiments the authors tend to replicate were generally single-shot experiments, often conducted with students in a pen-and-paper environment. These settings allowed ample time to describe the hypothetical scenario and allow participants to immerse themselves in it, then responds accordingly. Some of the studies involved real-stakes (for example Problem 14). I would advise the authors to consider the closeness of replication at the problem-level. It would be more plausible to say we plan to closely replicate experiments A and B and a far replication of C and D. The difference in interpretation may stem from how broadly one interprets the category "Contextual factors". The authors have said 'same', i.e. no difference in context between the original and this replication but in my opinion, a series of 21 experimental questions, covering a broad range of financial decision-making scenarios, is simply a different context to being asked a single question (or a few questions) over which you have time to deliberate.**

Thank you for catching that. Contextual variables should have been labeled "Different", and according to LeBel et al contextual variables do not impact evaluating replication closeness given that all replications are always a different context. We corrected that in our Table 9.

LeBel et al criteria focuses on IVs, DVs, constructs, stimuli, etc. With the exception of procedure, physical setting, and population, we can summarize that our design for each of the Problems is similar/same for all criteria, which using their categorization best matches the "very close replication" criteria.

6. Lack for Pre-registration for pooled analysis of results

Exploratory Analysis – pg 46

For Problem 21, if we fail to replicate the original findings, we will try log-transforming the prices and removing all answers that are 3 standard deviations above the mean (with the criteria of p < .01 to adjust for multiple analyses). Meanwhile, in the actual data collection, we aim to examine the intercorrelations among mental accounting experiments that support the original findings.
This point should be developed. One advantage of recording multiple measure of mental accounting is that you can then run models that control for individual differences. It would have been helpful to see some pre-registration of a meta-analysis of the results, for example, a mixed-effects logistic regression model that pools the responses with binary outcomes. This type of analysis would enhance the contribution of the study.

We agree that these are exciting opportunities that emerge from this design, however, we consider these directions exploratory, going beyond the replications of the Problems in Thaler (1999). We indicated an initial correlational analysis to get at that, yet given the scope and scale of this project, we would rather focus our attention and pre-registration on confirmatory predictions.

We added more details on how we intend to conduct the baseline exploratory correlational analyses in the "Exploratory analysis" section.

7. Dealing with Outliers:

"Outliers would be classified as either error outliers or other outliers (Leys et al., 2019). For error outliers, outliers due to wrong data entry, we will check up the raw data to see if corrections can be made. Explanations will be provided if outliers are removed. Please refer to the supplementary Section "Exclusion criteria" for detailed data exclusion method. "
When I referred to the data exclusion section, I could find no mention of the method for making corrections. This is very important to specify precisely. Generally, making corrections to raw data should be avoided. If it is deemed absolutely necessary in this instance, an example to assuage concerns would be useful.

Thank you for nudging us to do better in reporting these analyses and clarifying this point.

We removed the reference to categorization of the outliers, and clarified our analysis. We replaced “corrections” to the simple reference to excluding outliers.

We will not be making any corrections to raw data, and we will be reporting results for both pre and post exclusions, with a comparison in the supplementary.

> **8. Minor issues**
> **Pg 12: Why is Thaler 2016 cited as the source for Problem 1 in Thaler 1999?**

Thank you for pointing this out. We have revised the citation to Tversky and Kahneman (1986) and updated the manuscript accordingly.

> **There are problems in Table 1**
>
> **For example, for Problem 7 the source is cited as Thaler (1999) and the hypothesis is “not explicitly stated”. This is inaccurate as reading Thaler (1999) it is clear the source is Thaler (1985), and in that paper the difference in WTP is explained using transaction utility, which depends on the price the individual pays compared to some reference price (so it does come with an hypothesis).**

Thank you for pointing this out.

We changed the citation to Thaler (1985) and added the hypothesis to the table.

In Table 3.
Shafir and Thaler, 1998 (Wine Bottle) study is described as "Manipulation with two conditions testing the *fluid* value of wine." [my italics] Repeating the pun from the paper title in the one-line description is not informative. Similarly, in Problem 21 - Three conditions are manipulated to test whether small expenses are booked. Booked is accounting jargon that Thaler (1999) explains, so its subsequent use makes sense. Table 3 does not explain 'booked' and hence it should not be used.

Thank you for highlighting those and the very careful reading. We agree.

We removed the word "fluid" and changed the Problem 21 description to:

"Manipulations with three conditions testing expenses framing."

Problem 5 – unclear when happiness is evaluated? This is an ambiguity in the original study that need not be replicated.

We are unsure of the exact concern here, but we aim to repeat original studies regardless of their possible weaknesses, to determine whether we can get similar effects.

Happiness/unhappiness is evaluated after both events happened. We have revised and simplified the instructions of this Problem to make it clearer:

"Below you will find three pairs of events. In each case, the same events occur, either on the same day (for Mr. A) or two weeks apart (for Mr. B).

You are asked to judge whether Mr. A or Mr. B is happier, or in the case of two negative events, who is more unhappy. If you think the alternatives are emotionally equivalent, select "no difference."

(Note: You are only asked to judge whether it is better to have the events separately or together)."

Problem 6 – confusing because the introduction states that you will compare single loss to loss after gain. But then after first scenario (which follows this pattern) the rest compare a loss to a loss followed by another loss.

Thank you. The example in the introduction was intended to illustrate what the scenarios will be like. We adjusted the text to be clearer:

"Consider the following two events: (A) you lose $x. (B) you lose $x after gaining/losing $y. We are interested in the emotional impact of the loss of $x in both cases.

> Are you more upset about the loss of money when it occurs alone (A), or when it occurs directly after a prior gain/loss (B)?
> Below are several questions of this type. In each case please compare the incremental effect of the event described. If you feel there is no difference you may check that, but please express a preference if you have one.
> For each of the following pairs of events, please indicate which of the two hurts more:"
>
> **Problem 10 – Thaler (1999) did not give the "I don't understand" option. Why is it being added here? (And why not add it to other questions too?)**

Yes, thank you. We should have made this clearer.

We added the "I don't understand" option here because our unofficial pretesting showed that this Problem was difficult to comprehend for people who are not familiar with wine. By adding this option, we ensure that participants will not just choose a random option when they cannot understand the question. We did not have that concern in other experiments, just those relating to drinking wine.

Additional clarifications

We have made the following changes in Qualtrics:

1. For Problem 10, we added a comma in the sentence. "*At the time that you acquire this wine, which statement more accurately captures your feelings?"*

2. For Problem 11-$20 condition, we changed a typo.

3. For Problem 12, we changed the question into *"How much would you be willing to pay* ***for the regular ticket*** *to avoid waiting for 45 minutes?"* We also updated the data analysis for this Problem.

4. For the demographic question on drinking alcohol, we removed the bold to ensure a consistent format.