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.

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.

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.

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.

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.

It is not consistent and very hard to follow on what kind of datacollection 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.

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).

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.

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.

You writhe 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 ...".

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.

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.

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.

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.

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.

"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.

"...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.

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.
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.

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?

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.

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 a success? Maybe it would be better just to remove this sentence or detail how it increases the validity of the present replication.

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.

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.

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?"

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.

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.

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.

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.

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.

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”.

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.

Best regards,
Barnabas Szaszi