**Review of “Revisiting mental accounting classic paradigms: Replication of the experiments reviewed in Thaler (1999)” Stage 1 Registered Report.**

**Reviewer: 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.

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[[1]](#footnote-1)), 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.

I organise my comments by topic:

1. Justification of choice of replication studies and proposed contribution
2. Justification of Randomisation and issue of order effects
3. Power analysis
4. Extensions of Thaler (1999) not novel (as described)
5. Classification as ‘Close Replication’ dubious.
6. Lack for Pre-registration for pooled analysis of results (implied in RR snapshot?)
7. Dealing with Outliers
8. Minor Inaccuracies and Omissions
9. **Justification of studies**

The [guidelines for registered reports](https://osf.io/preprints/metaarxiv/43298/) 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.

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[[2]](#footnote-2) so some rationale should be given for how this sample was selected.. 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.

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.

1. **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).

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. 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.

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.

*.*

1. **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).

1. **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.

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.

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)



1. **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.

1. **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.

**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.

**8. Minor Issues**

Pg 12: Why is Thaler 2016 cited as the source for Problem 1 in Thaler 1999?

**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).

**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.

**Problem Specific Issues**

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

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 losso followed by another loss.

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?)

1. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)