Reply to decision letter reviews: #164 R&R

Response to Editor: Prof. Chris Chambers

> **Thank you for your careful and thorough revision and response. One of the reviewers from Stage 1 was available to evaluate the revised manuscript within a short timeframe. As you will see, the reviewer is largely satisified, with just a few remaining clarifications to make in a minor revision. We should then be in a position to offer Stage 1 in-principle acceptance (IPA) without further in-depth review.**
> **For the reviewer who was unable to assess the manuscript within 14 days, I have considered your revisions myself and found them sufficient to proceed. In the event that IPA is awarded, I will invite that reviewer back in due course to evaluate the Stage 2 manuscript.**

Thank you for the review obtained, and the invitation for a second revise and resubmit. We appreciate your and the reviewer's time and effort. Below we respond to the issues raised by the reviewer.

We believe that we were able to answer the questions and address the concerns without requiring further changes to the manuscript. We look forward to proceeding to Stage 2 with in-principle acceptance from PCIRR. Thank you for the opportunity to take part in this initiative.

Response to Reviewer #1: Dr. Barnabas Imre Szaszi

> **I would like again to thank the authors to take the time to review the manuscript based on the comments. In general, I believe the authors sufficiently addressed most of my concerns.**

Thank you for your second review. We very much appreciate you going over all our revisions and following-up with questions.

> **Below I only focus on the comments where I feel that further changes would improve the quality and impact of the proposed study.**

> **1. You suggested the following improved sentence to be included in the abstract:**
> **"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]."**
> **Although I understand the reasoning behind the sentence, to me it is somewhat strange to dichotomize the results and the effect size ranges that way. Is it interesting what are the exact ranges for the different groups? I think rather the number of hypotheses with support and no support, and the general effect size range is interesting, but independently. However, I have no strong opinion on that.**

Yes, we understand. The tricky bit is that your comment refers to the abstract, which is very limited in length and content and does not allow for complex nuanced reporting. We could not find a way to better summarize 17 hypotheses and 21 studies in the abstract. In the abstract we could only give a summarized range, and the reader will need to go in and read the manuscript for a more in-depth breakdown of the different hypotheses and effects, we see no way of escaping that.

We would have appreciated some advice on how to improve here, we could not think of any. We ask for your and the editor's understanding that we chose to leave the abstract as is. We would be happy to revisit this in Stage 2 given clear editorial advice on the expectations of how to improve this, and we would please ask for some citations showing us similar abstracts tackling the need to summarize 17 hypotheses within one short abstract.

> **2. You write the following sidenote response regarding how big a 0.23 effect size is.**
> **"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."**
> **It does not concern the acceptability of the present paper, but I cannot stop myself from saying that I still disagree and leaning toward the effect size interpretation of https://journals.sagepub.com/doi/full/10.1177/2515245919847202, that is 0.29/0.36 is rather a medium to large effect size.**

Apologies, but we believe there is some confusion and misunderstanding here, specifically about the differences between correlational Pearson's *r* effects and the Cohen's *d* effects.

Please refer to our collaborative effect size guide on https://mgto.org/effectsizeguide in subsection "Benchmarks" where we summarize all the benchmarks we know of, including the reference to "Funder and Ozer (2019)" in our footnote 5. We are very familiar with the article you cited, and it refers to *r* correlations effects, and not Cohen's *d* experimental effects.

In this replication of Thaler's review we focused on experimental effects with Cohen's *d* and *f* effects. Therefore, according to all benchmarks we know of in social psychology that are summarized in our guide, these ***d*** = 0.29/0.36 are considered below or equal to medium/typical effects in the general social psychology literature.

However, this does not seem like a crucial issue of any relevance to our manuscript. In our previous revision we removed any reference to categorization of effect sizes, and the only mention of effects was in reference to the that the power analysis is aimed at detecting effects that are much weaker than the effects deduced from the reported statistics in the target article.

> **3. To me the last part of this sentence seems to be confusing:**
> **"Our second goal was to examine several predictions made by Thaler regarding mental accounting behaviors that the review did not cover empirical tests for."**

Apologies, but we needed more information here. We would have appreciated some information regarding the possible source of confusion.

In his review, Thaler made some generalized predictions without supplementing that with providing empirical evidence from the literature to support those. This means that there was no empirical test indicated for those hypotheses. We added our extensions aiming at testing those untested predictions in Thaler's review.

4. You write that " In a pre-registered study with an American online Amazon Mechanical Turk sample (N = 1000)". Is that n=1000 the number of people starting the experiment OR participants providing a full response to all questions OR the target number after exclusions?

In our reference to exclusions we wrote in the supplementary:

> In the actual data collection, we will focus on our analyses on the full sample. However, as a supplementary analysis and to examine any potential issues, we will also determine further findings reports with exclusions. In any case, we will report exclusions in detail with results for the full sample and results following exclusions (in either the manuscript or the supplementary).

Therefore, this indicates the full sample collected. We also wrote that we will only analyze the responses from participants who completed the survey.

Sidenote: Given our experience with the target sample, there are typically very few exclusions, MTurk participants are English proficient and attentiveness. In our experience with JDM replication so far exclusions have rarely led to any major changes in the interpretation of the results. We do not see any reason why this project would be any different here, yet we are open to the possibility, and will therefore be comparing pre and post exclusions in the supplementary, with a brief summary in the main manuscript.

Again, thank you for the interesting submission!

Thank you for your time and feedback. We appreciate it very much.

We look forward to working with you and the team again in Stage 2 after data collection.