Dear Chris,

Thank you for your swift action on our revision and for the feedback from the reviewers. Please find our reply to the remaining queries below.

I am, however, more worried about the sample sizes for the latter three hypotheses (manipulation checks), all of which are N=11 or less. That said, since they are manipulation checks, it is in some ways your risk to take as to whether you genuinely believe these samples are sufficiently large, because failure of critical manipulation checks is one of the few grounds for rejection at Stage 2 based on outcomes (see criterion 2A here). If you feel there is risk of failure in any of these checks, I would strongly advise increasing the sample size to avoid a rare Stage 2 rejection.

We fully agree that the outcome of the manipulation checks is critical for the interpretation of the proposed study. In this respect, two questions must be addressed: (1) Which data points could demonstrate that a manipulation has failed? (2) What are the implications of a failed manipulation check for the interpretation of the results?

Concerning the first question, we now propose an equivalence testing approach for a statistical assessment of a "null effect" that is conducted after the significance test for the manipulation check has produced a non-significant result (Lakens et al., 2018). Specifically, we propose a two one-sided tests (TOST) procedure for the rejection of a smallest effect size of interest (SESOI). Effect sizes obtained from the manipulation checks in our previous studies were large, ranging between d_z = 0.8-1.1. Even if we have sampled these large effects by chance, a conservative assumption would be that the "true" effect size will most likely lie in the medium range. Therefore, we propose d_z = 0.40 as SESOI for our manipulation checks and we adjusted the sampling plan to collect valid data sets from n = 41 (valid n). The proposed sample size n= 41 has good statistical power (1-β = 0.80) for the detection of $d_z \geq$ 0.40 and excellent statistical power (1-β = 0.95) for the detection of $d_z \geq 0.52$ in one-tailed t-tests. For the equivalence tests, we will set the upper equivalence bound Δ_U to d = 0.40 and the lower equivalence bound Δ_L to d = -0.40. If the observed effect falls between these preregistered bounds, as indicated by a non-significant test result in the TOST procedure, then the null hypothesis is rejected and the decision is made that the manipulation has failed.

Concerning the second question, a failed manipulation check would seriously threaten the conclusiveness of our results. Without demonstration of PIT effects, there would be no indication that cue-dependent (aka habitual) action tendencies were instigated in the first place, and without a devaluation effect, motivational control of the associated PIT tendency is not plausible. In short, failed manipulations would seriously question the validity of the experiment, and we are not interested in the journal publication of an inconclusive study due to inappropriate study procedures. For this worst-case scenario, we therefore explicitly agree to a Stage 2 rejection at PCI. In this case, we will make the study report publicly available in our data archive (OSF), without asking a PCI-friendly journal for publication. This conditional publication rule is now also mentioned in the preregistration document (Table 2).

The concerns of Reviewer 1 are more fundamental to the acceptability of the current proposal, and I recognise they represent a strong difference of opinion between yourselves and the reviewer. In essence, the reviewer does not believe the design is capable of answering the research question, in part due to lack of an appropriate control but also a broader misalignment between the question/theory and the proposed methodology. It is difficult to see a way forward for this manuscript without resolving this disagreement in one way or another, and I am also keen to avoid overly burdening reviewers, especially when a discussion reaches a stalemate. Instead, I am going to offer you the opportunity to revise again. If you choose to simply rebut Reviewer 1's point rather than revise the design, I will seek additional specialist input to determine whether to accept or reject the proposal as submitted (with no further revision). However, if you believe you can sensibly revise the design to address the reviewer's concern once and for all, then I will invite Reviewer 1 back for a look before issuing a final decision. In either case, the next revision is pivotal and will determine whether in-principle acceptance is achievable.

Frankly speaking, we do not believe that we can provide a rebuttal that could change this reviewer's opinion. The discussion has shifted to a debate of the proper definition of "habit", which cannot be resolved empirically (see De Houwer, 2019). Therefore, we would appreciate if you could ask a new reviewer for an independent evaluation of our revised proposal. Although we disagree with the perspective of the reviewer, we want to explicitly thank him or her for taking the time to review our proposal and for the critical questions that motivated us to include a brief discussion of habit definitions in the revised research proposal (see pp 9-10; "Dual action or Controlled Action?").

References:

De Houwer, J. (2019). On how definitions of habits can complicate habit research. *Frontiers in Psychology*, *10*, 2642. https://doi.org/10.3389/fpsyg.2019.02642

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963