RECOMMENDER

Comment 1a: Thank you for your revision, which addresses the point about conducting the power analysis. However, I don't see a justification for the minimal effect size of interest. Can you put it in the section on "sample size estimation"? Why .10? In giving a reason for .10, it is helpful to give units so people are not confused, ".10 of what?" e.g. ".10 RPE Likert units". I find this greatly adds clarity whenever one quotes a dfference or slope; always give the units.

Authors' response: Thank you very much for your thorough review. This information has been added to the subsection "*sample size estimation*" (see also page 17):

"To estimate the sample size of the main study, we used the R package smir on the pilot data. In line with recent guidelines that suggest running power analysis based on the lowest meaningful estimate of the effect size (Dienes, 2021), we ran 1000 simulations with a one-unit change on the raw scale of RPE absolute prediction error predicting a raw slope of 0.10 units increase of running pleasure. Specifically, in order to run power analysis with the lowest meaningful estimate of the effect size, we decided to use 0.10 units increase of running pleasure as effect size of interest. In other word, we set an interesting effect a bit lower than the one obtained in the step 2 model from the pilot data (i.e., a raw slope of 0.15). Results indicated that for an alpha of 0.05, the power was .83 (95% confidence interval [.80 .85]) with 27 participants across 336 observations. Accordingly, if α is chosen at .05, with a minimum effect size of .10, and a power of .80 is desired, then a sample of 27 participants along 12 measurement points (i.e., a running session) is required for testing the step 2 LMM presented in the previous section."

Comment 1b: A similar point re: your table - what are the slopes? Raw slopes or standardized? Can you make this explicit.

Authors' response: We used raw slopes. This information now appears in Table 1.

Comment 2: While we are dealing with that issue still, let me mention one other thing. In your design table in saying what theory the results might count against you say "A failure to confirm this hypothesis would either question the occurrence of RPEbased prediction errors or indicate that our procedure is not appropriate to test this hypothesis. " I understand why there could be an asymmetry here, but the asymmetry here is not appealing: If you confirm the hypothesis you will conclude the variables were valid and the experiment well designed; if conversely the results go against the hypothesis you will doubt the experiment and not the hypothesis. A Registered Report is a good time to think hard about what would count against a theory. If you would doubt the study, what about it would you doubt? Could you add a outcome neutral test to forestall against that get-out clause? Or given the study as it is, while the most general theory might not be impugned, what slightly more specific theory would you give up? Can you state in this column the most general theory you *would* give up, or at least doubt, given negative results?

Authors' response: Thank you for this suggestion.

- In the case our results go against the hypothesis, doubts about our study have been mentioned in the section "Interpretation given to different outcomes" of Table 2 (see also pages 28-29):

"If the RPE absolute prediction error is significantly and positively associated with retrospective running pleasure, we will conclude finding evidence for our hypothesis. This will lead us to the interpretation that using prospective and retrospective RPE may be beneficial for better identifying sessions of physical exercise that lead to increased (or decreased) experience of pleasure.

In the case of nonsignificant effect of RPE absolute prediction errors on running pleasure, this will lead us to discuss how the current design and procedure of the physical exercise program (i.e., "self-selected" running

sessions) could be adapted (e.g., standardized running sessions) for observing a significant effect of RPE-based prediction error on running pleasure."

- The theory we would give up has been mentioned in the section "Theory that could be shown wrong by the outcomes" of Table 2 (see also page 28):

"Our study tests whether one main tenet of reward prediction error (i.e., when rewarding outcome is better than expected, it induces more pleasure than a reward that matches a prior expectation) translates to (a self-selected mode of) physical exercise.
A failure to confirm this hypothesis would question the relevance of applying central notions in theoretical models of reinforcement learning (e.g., reward prediction errors) to explain reward-based experiences (e.g., retrospective running pleasure) induced by daily-life conducts (e.g., physical exercise)."
