RECOMMENDER

Comment 1: Only report those analyses in a Stage 1 that you will base inferences on. That is, if you plan to base inferences only on model 2, do not even report the possibility of model 3 at this stage. If you want to leave open that you will base conclusions on model 3, then indicate precisely how you will decide which model the conclusions will be based on. Incidentally as you plan to keep all variables in the equation, you don't need diffferent steps. Just enter all variables simultaneously. So if you are just going to use model 2, specify only the final equation with all variables in, and then base your conclusions on that. Note that pre-registering a specific analysis now does not stop you exploring other analyses later in Stage 2 when the data are in. But these other analyses will be in a separate non-pre-registered results section, and the abstract will draw conclusions only from the pre-registered analyses, and the discussion will keep pre-registered analyses centre-stage.

Authors' response: Thank you very much for these additional explanations. We now fully understand and agree with your suggestion. Accordingly, only model 2 specification is now detailed in the *Data analytical plan* section. Besides, the three-steps sequence is still detailed in the *Pilot data* section to provide rationales on why the step 2 model is selected as preregistered analyses. Specifically, the pilot data shows that (i) step 1 (the null model) provide evidence for a nested data structure that requires multilevel modeling rather than a single-level data analytic approach, and (ii) that the model with random intercept and fixed slope (i.e., step 2) is the best model (i.e., the model with random intercept and random slope, step3, does not result in a better fit).

Comment 2: You could treat your two IVs, absolute and relative prediction error likewise. Choose one now, and keep the other for exploration later. Just bear in mind your main conclusion in the abstract and discussion will be based on the pre-registered analysis. If you keep both in the pre-registration you need to decide how you will base conclusions on different possible patterns of results. The multiple testing problem I mentioned before remains. So if you keep both, use a multiple testing correction, e.g. Bonferroni. If you use Bonferroni, the decision rule can be if either or both slopes are significant, then there is an effect of prediction error on pleasure. You also need to calculate power with respect to the alpha determined by the multiple testing correct, e.g. .025 for Bonferroni. Now more on power. Power is to control the risk of missing out on an effect of interest. Thus, it must be calculated with respect to the smallest effect just of interest. This is of course hard to determine, but you mention the rough heuristic of using the smallest available estimated effect. Based on one pilot, this heuristic no longer captures the spirit of what is needed. I find thinking in terms of raw effect sizes more helpful than standardized ones. You usefully provide raw regression slopes from the pilot. No one can judge whether an R2 of 0.40 is interesting whereas one of 0.3 is getting boring. Whether a one unit change on your raw scale of absolute prediction error predicts 0.15 units of pleasure on your scale is interesting at least feels vaguely judgeable. (Presumably the slope mentioned next, 0.57, is for relative prediction error, so I think there is a typo here.) I presume you found this interesting, even though it is small, because it motivated you to plan this study. But it does feel on the edge. How about setting an interesting effect just a bit lower than this, say 0.1 pleasure units/abs PE unit, and caculating the power for this? I think this would work best by referring to the raw slopes in the power calculation section. As I mentioned earlier, a 80% CI on the slope, and using the bottom limit, would be a touch more objective (even if 80% is relatively arbitrary), because it is a procedure that could be repeatedly used in many cases.

Authors' response: Thank you again for these additional explanations. We now only refer to the absolute index of prediction error in the stage 1 preregistration manuscript. As suggested, the relative index of prediction error will be kept for exploratory analyses in the stage 2 preregistration paper. Regarding power calculations, while we totally understand the logic of your suggestion, we were not able to find how we could run a sample size estimation using a raw regression slope in the context of Linear Mixed Models (LMM). We thus decided to keep our previous sample size estimation analyses. Nevertheless, we remain open to any suggestion that could help us to run the rquested power analyses in the contect of an LMM. Besides, we now use a proportion of participants/running sessions that fits better the 6 months start-to-run program (i.e., 21 participants along 10 measurement points). The sample size estimation is now detailed as follows:

"A priori power analysis was performed using Power IN Two-level designs software which is designed to estimate standard errors of regression coefficients in hierarchical linear models for power calculations (Snijders and Bosker, 1993). We ran sample size estimation analyses with a conditional R^2 of .40 based on the results obtained in the step 2 model. Accordingly, if α is chosen at .05, an effect size of .40 is what we expect, and a power of .80 is desired, then a sample of 21 participants along 10 measurement points (i.e., a running session) is required."

REVIEWER 1

General comment: The authors have addressed my previous comments. I have no further comment.

Authors' response: We would like to thank the reviewer for the positive reception and for the insightful and in-depth comments on our stage 1 preregistration manuscript.

REVIEWER 2

General comment: The authors have responded well to each of the points raised in the previous review. The explanations provided are very helpful. Great work! I have no further major comments for this paper. I have just a few minor comments/edits below.

Authors' response: Thank you very much for the positive feedback and for these additional editing comments.

Minor Comments

Comment 1: Use the term 'perceived exertion' rather than 'physical exertion' whenever referring to exertion that is perceived. Physical exertion is usually measured objectively (e.g., heart rate, VO2) and therefore is something slightly different. This may seem a minor point (and I can only imagine how difficult such nuances are in a non-native language), but it is important for clarity. Below are some suggested corrections from the abstract:

"…We aim to examine this research question by using ecological momentary assessment of ~~physical~~ perceived exertion to be filled out before (anticipatory RPE) and after (retrospective RPE, retrospective pleasure) each running session of a start-to-run program. By capitalizing on the core dynamic of reward prediction errors, we hypothesize that running sessions that are experienced with a lesser level of ~~physical~~ perceived exertion than anticipated (a positive RPE-based prediction error) should be associated with a higher level of retrospective pleasure ~~during~~ following the session of physical exercise, and vice versa (higher score of retrospective RPE than prospective RPE; a negative RPE-based prediction error)."

Authors' response: Thank you for this suggestion. "Perceived exertion" is now used when referring to exertion that is perceived. In the title, we also replaced "physical exertion" by "perceived exertion".

Comment 2a: Some suggested edits for Page 3: Indeed, the level of ~~physical~~ perceived exertion is usually indexed while exercising (i.e., momentary ratings of perceived exertion [RPE]; e.g., "What intensity of exertion do you feel now?") or directly after the exercise session (i.e., retrospective RPE; e.g., "What intensity of exertion did you feel during this session?"; or "How was your workout?"; Foster et al., 2001; Haile et al., 2015; Robertson ~~and~~ & Noble, 1997).

Comment 2b: Where you added the detail about music on page 8, I think it would be useful to mention that this will be considered as a covariate. Suggested edit: "Specifically, participants will be encouraged to self-select their running frequency, intensity, and duration, in which they will be allowed to undertake these sessions alone or in groups, where they want. Participants will also be allowed to listen to music if they want to. These variables (i.e., presence of others and music listening) will be recorded and included as covariates in the analysis, see section 2.6.2)."

Comment 2c: Section 2.5.1: Use "assessed" or measured rather than undertaken. Also, the word perceived is not needed in this sentence (it is included in the abbreviation). Suggestion: ***Prediction error of RPE.*** RPE will be ~~undertaken~~ assessed directly before (prospective RPE) and after (retrospective RPE) each running session on the Formyfit app (see **Figure 1**). Based on Foster and colleagues' findings (2001), participants will be asked to provide a prospective or retrospective rating of their ~~perceived~~ RPE of the overall running session.

Comment 2d: Section 2.5.2. Regarding the instrument to measure retrospective running pleasure. I think this could just be a translation issue, but the responses for 0 ("not at all") and 6 ("extremely") do not match the question for this measure. If asking "What intensity of pleasure did you feel during this session?" then perhaps the translation should be 0 ("none at all") to 6 ("extreme")? The other integers of (1 = "very little", 2 = "slightly", 3 = "moderately", 4 = "quite a bit", 5 = "very much") are good.

Authors' response: Thank you very much for these editing suggestions. We have made the requested change.
