**RECOMMENDER**

**General comment:** I have gone with the two reviews I have in, following your email, to expedite the process. The reviewers are largely positive, but have a number of points to clarify. I have a couple of points as well.

**Authors' response:** We would like to thank you for the positive reception of our manuscript and for fastening the reviewing process of our Stage 1 Registered Report.

**Comment 1:** In a Registered Report, one ties down all analytic and inferential flexibility. While in the pilot models 2 and 3 fare best, in the main study this may or may not be true. Be absolutely explicit about the criteria you will use to end up with the one model you will draw inferences from. Make sure anyone reading your planned analyses and having your raw data for the main study would end up making the same decisions. Further, you have two DVs- there is room for inferential flexibility here (what if one DV shows one thing and the other another?), as well as a familywise error rate problem. Given your pilot, I would suggest picking one of the DVs - probably absolute prediction error. Otherwise you need to correct for familywise error and specify your decision rule depending on the different pattern of possible results.

**Authors' response:** Thank you for this suggestion. We would like to point out that, in all models of this study, there is only one dependent variable in the study: running_pleasure. There are, however, two main independent variables (absolute_prediction_error and relative_prediction_error) and we agree that it might be better from a purely analytical point of view to test only one of these predictors. Nevertheless, we think that, since it is the first study to operationalize such indexes of physical exertion prediction error, it is important to test the same Linear Mixed Models twice (i.e., one model with absolute_prediction_error, a second with relative_prediction_error). It is indeed a purpose of the study to examine which of these 2 variables would be the most relevant.
Besides, the way we described the stepwise progression of models (i.e., the numbering of the models) was maybe confusing. Specifically, we actually test a null model (step 1), then a 2nd model (model step 2) then a 3rd model (step 3). Step 2 and step 3 were undertaken twice: once with absolute prediction error, and a second time with absolute prediction error as a predictor. Accordingly, through the revised version of the manuscript, we edited the numbering of the models as follows (see also revised manuscript, pages 14-15):

- Absolute_prediction_error: step 1 (null model), step 2 (the model with random intercepts and fixed slopes) and step 3 (the model with random intercepts and random slope of prediction error).
- Relative_prediction_error: step 1 (null model), step 2 (the model with random intercepts and fixed slopes) and step 3 (the model with random intercepts and random slope of prediction error).

Lastly, concerning the stepwise progression of models (between step 2 and step 3), we would like to acknowledge that the set of predictors will be kept anyway (e.g., even if they are non-significant in step 2, they will be kept in step 3), regardless of whether they are significant or not based on the rationale that all these predictors are meaningful from a theoretical point of view and would be included in the final model (even if non-significant).

**Comment 2:** The function of calculating power is to control the error rate of missing out on interesting effects. But that means one needs to calculate power with respect to roughly the smallest effect that you do not want to miss out on. The PCI RR guidelines for authors puts it "power analysis should be based on the lowest available or meaningful estimate of the effect size". As one reviewer points out, you actually use an effect size estimate for power larger than that obtained in the pilot. Yet presumably an effect considerably smaller than found in the pilot would still be interesting and one you would not want to miss out on. The issue is discussed in some detail here: https://doi.org/10.1525/collabra.28202 I would suggest finding the 80% CI for each effect in the pilot that is in your Design Table, and using the bottom limit of the interval in each case as the effect you use for calculating power. But you might have other ideas, after the reading the linked paper, for how to address the point.

**Authors' response:** We would like to thank you and the reviewers for having spotted this error. We have now conducted the power analysis based on the lowest meaningful estimate of the effect size obtained in the pilot study, which is the conditional $R^2$ of .41 from step 2 models with relative prediction error as predictor. This information has been added in the revised manuscript as follows (see also page 18, paragraph 3): "In line with recent guidelines that suggest running power analysis based on the lowest meaningful estimate of the effect size (Dienes, 2021), we ran sample size estimation analyses with a conditional $R^2$ of .40 based on the results obtained in the step 2 model of relative prediction error. 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 28 participants along 5 measurement points (i.e., a running session) is required for third-step models."

**Comment 3:** Finally, you phrase the issue as predicting pleasure from prediction error, but prediction error is not completely assessed until pleasure is. A causal diagram of what is going on could of course take many forms. No need to do anything about this now, but I presume the discussion will touch on this point - all in good time.

**Authors' response:** Thank you very much for this insightful comment. As suggested, we will address this issue in the Discussion section of the Stage 2 manuscript when ready.

## REVIEWER 1

**General comment:** Thank you for the opportunity to review this work. The proposed study addresses an important question and has the potential to contribute to the literature on exercise-related affect. The research question is scientifically justifiable, based upon existing evidence and current literature and the hypothesis is logical and plausible. Overall, this is a strong manuscript, however I do have some concerns and suggestions that I would encourage the authors to consider. These are listed below. Good luck with this research!

**Authors' response:** We would like to thank you for the positive reception of our manuscript and for the insightful and in-depth comments.

### Major comments

**Comment 1:** There are a few grammatical errors throughout – I pointed out a few examples in the specific comments below, but the final paper will need to be proofread / edited carefully.

**Authors' response:** The manuscript has been reviewed for spelling and grammar by all authors. The final paper (Stage 2 Registered report) will be edited by a professional scientific editing service.

**Comment 2a:** Insufficient detail to enable replication. The start-to-run program is not clearly described. I understand that there is some flexibility (based on the SET) but there is no description of the program at all. I'm also curious why there is only one group session per week. If these are beginning runners more frequent meetings might help with adherence? Another consideration, if the "free run" days are entirely self-selected ("allowing participants to choose the duration, frequency, and intensity of each run", p. 8), how will you ensure progression toward the 5-mile goal?

**Comment 2b:** Related to the above point, the authors state that "participants will be able to choose how they want to follow the proposed training program (e.g., to strictly follow it, to use it as a basis for training, or choose not to follow it)" (pg. 8). I am not sure that this is the best approach. There is no standardization of the intervention in this case. Specific to the variable of interest, forecasted RPE, this could be problematic. it is possible that if one's forecasted RPE is high that they might decide to not follow the program and do something easier. Or they may make an adjustment during the workout that affects the prediction error. Page 7 indicates that runners can decide to "walk when they feel the need to do so" – I understand this from a safety and adherence perspective, and for the other reasons given in the paper (autonomy, enjoyment), but I am concerned that the difference between forecasted and retrospective RPE could be affected by this. For example, let's say today, according to the app, I am supposed to run for 4 miles; my forecasted RPE is 7. I start running and decide I don't feel like doing 4 miles today so I only do 3 miles. Or I complete 4 miles but I walk for half the distance. Now, my retrospective RPE is 5. Does this mean I had a prediction error of 2? Or does it reflect that I decided not to follow the recommended workout?

**Authors' response:** Thank you for these important comments. We fully agree that details were lacking with regard to the running program available on the Formyfit app. Nevertheless, it is important to acknowledge that, in the present study, the primary goal of the start-to-run program is to get participants to perform enough running sessions to test our hypothesis on the impact of

RPE prediction errors on running pleasure. Accordingly, the coaching group sessions and the 5 miles running event (happening in March 2023) were implemented to foster participant's repetition of running sessions across time. This aspect has been better detailed in the Methods section, as follows (see also revised manuscript, page 7, paragraph 3): "The primary goal of the start-to-run program is to provide a context that will allow participants to perform enough running sessions (minimum 5; see also the **Sample size estimation** section) to test our hypothesis on the impact of RPE prediction errors on running pleasure."

We also wanted to follow the same logic (i.e., to get participants to repeat running sessions) by adding the possibility for the participant to follow a progressive training program, which was available on the Formyfit app. However, your comments make it clear to us that this option created some high degrees of confusion. We thus decided to withdraw this option (i.e., to follow a progressive training program) from our protocol, and to extract the information on the running duration (based on participants' $VO_2$max estimation) from the Formyfit program as a general guide for the recommended running duration. This information is acknowledged in the Methods section, as follows (see also revised manuscript, page 9, paragraph 1): "Importantly, since participants will be novice or low frequent runners, the Formyfit app will recommend running duration based on participants' $VO_2$max (estimated from the SET). These recommendations will be made available to the participant on the app and could be downloaded in a document format. The participants will be able to choose whether or not they want to follow the proposed running duration."

We also totally agree with the rationale of the example detailed in your **comment 2b.** In our opinion, this example might also describe issues that could be encountered by imposing participants to follow the distance and pace imposed by a progressive training program (as in progressive training program of the Formyft app). If we take back your example, if the participant chooses to run on hilly terrain, then running 4 miles on this route will be totally different than running on flat terrain. The same logic goes if participants decide to run in a group: it will be difficult for them to strictly respect the program. These are all situations that we wanted to avoid by imposing participants to strictly follow the distance and pace imposed by a progressive training program.

Interestingly, in the pilot study, less than 20% of the 228 running sessions were undertaken with the Formyfit program (N = 41) and the majority of these sessions were done in the beginning of the program (see also the variable "session type" the database available at https://doi.org/10.17605/OSF.IO/2SB86). It is also important to note that the LMM findings were not significantly modulated when deleting these 41 observations. Moreover, in the pilot study, participants were allowed to write a commentary on the Formyfit app after each running session (the option of posting a post-session commentary has also been acknowledged in the Methods section: see page 13, paragraph 1; see also **Figure 1Biv**). Several comments from the participants relied directly on the above-mentioned arguments. For instance, one participant said that it was very difficult to keep the running speed/pace advised by the app, especially when running uphill. Some other participants detailed that it was difficult for them to cope with the "vocal coach", especially when running uphill (i.e., Formyfit includes the option of receiving automatic oral feedback for guiding the individual on the speed to be adopted at specific section of the run; e.g., "slow down", "keep the pace"). The list of commentaries from the pilot study (in French and translated in English) is available at https://doi.org/10.17605/OSF.IO/2SB86.

To sum up, the Formyfit app will be used for obtaining pre- and post-running ratings, as well as covariate measures (average running speed, running distance, degree of familiarity with the running route, individual versus group session, music vs. no music). Besides, participants will have access to the general Formyfit dashboard app featuring summary information on their running sessions (e.g., frequency, average distance, average speed, and heart rate). This later aspect has now been acknowledged to the Methods section (see page 9, paragraph 1).

Finally, on a theoretical level, we would like to acknowledge that, by letting them choosing the running trails, the frequency, and speed of their running sessions, our start-to-run program fits well with training procedures derived from the ecological dynamic approach to physical exercise (e.g., David et al., 2016 https://link.springer.com/article/10.1007/s40279-016-0511-3; Rudd et al., 2021 https://doi.org/10.1080/17408989.2021.1886271). Specifically, this approach advocates for physical exercise behaviors that consider the relationship between individuals' characteristics and functional aspects of their environment (e.g., running session undertaken under multiple contexts). This aspect is now detailed in the Methods section (page 8, paragraph 1): "This approach fits also well with training procedures derived from the ecological dynamic approach to physical exercise (e.g., David et al., 2016; Rudd et al., 2021). Specifically, this approach advocates for physical exercise behaviors that consider the relationship between individuals' characteristics and functional aspects of their environment (e.g., running sessions undertaken under multiple contexts)."

**Comment 3:** Some other methodological questions I have are… Is the program held on the same indoor track as the jog test, or somewhere else? In the weekly group sessions, how many people will run at a time? Is there a group warm-up and/or cool down? What is the role of the coaches, do they run with the participants? Are participants allowed to listen to music while they run? This will likely influence RPE.

**Authors' response:** These important aspects have been better detailed in the revised manuscript:

- Running sessions could be undertaken outdoors or indoors (on a treadmill). For the outdoors session the GPS of the Smartphone is used to estimate the running distance and average speed. When launching an indoor session, the Formyfit app records the time and participants will be informed that they will have to encode the distance manually on the app at the end of the session. (see also page 9, paragraph 1).
- For the weekly group sessions, 4 different schedules will be proposed each week with a maximum of 10 participants per group. There will be two coaches in each group session. Each group session will start with a warm-up and ends with a cool-down and stretching routine. These steps will be guided by the coaches. The coaches will run with the participants, with one coach running at the front of the group, and the other at the back. This will allow the coaches to supervise the fastest and slowest runners and give personal advice (e.g., advice on running techniques and running stance) during the running session. (see page 9, paragraph 2).
- In the free run session, participants will be allowed to listen to music if they want to. This aspect will be controlled by asking participants to report (on the app, directly after the session) whether or not they ran with music (see also **Figure 1Biii**). This binary variable will be added as a covariate in our analyses (see also model description on pages 14-15).

**Comment 4:** Has the Formyfit app been validated? What is the algorithm used by the app to design the running program? Does the app take anything else into consideration other than the results of the submax VO2 jog test? I tried to check the link provided (www.formyfit.com) but got an error message indicating "access is forbidden". However, I checked the Google app store, and (assuming I located the correct app) there are some details that are not reflected in the method – for example "each training session is accompanied by voice coaching to guide and motivate you". This seems like an important point to discuss.

**Authors' response:** These are all very important points. However, we decided that participants will not use the Formyfit progressive running program (see our response to **Comment 2**). Therefore, no reference to this program was kept (or further detailed) within the revised text. Besides, the Formyfit progressive running program (based on the $VO_2$max, age, gender and Body Mass Index) was not formally validated in a scientific study and the algorithm is kept "secret" for commercial purposes. Hence, the Formyfit was primarily used for the recording of the sessions and the possibility to include the RPE and running pleasure values.

## Specific Comments

**Comment 5:** Abstract, Line 9: Delete "on" before "the level of pleasure"; Abstract, Line 17: Replace "advance" with "indicate", or you could use "advance the idea that"; Abstract, Line 18: Delete "better" before "identifying"; Page 3, Lines 4-5: I would change this to "including health behaviors"; Page 7, Line 13: Replace "fell" with "feel"; Page 7, Line 14: Replace "technics" with "techniques";

**Authors' response:** Thank you for the thorough reviewing of our paper. These spelling and grammar issues are now fixed.

**Comment 6:** Page 3, Line 3: Sentence beginning "'It allows to effectively prepare…" is unclear.

**Authors' response:** This sentence now reads: "It allows individuals to effectively prepare for upcoming events and facilitates the enactment of goal-directed actions and the planning of behaviors, including health behaviors."

**Comment 7a:** Page 3, Lines 5 - 6: Bit of a jump here to describing this process as memory based. Needs an intervening sentence to explain.
**Comment 7b:** Page 3, Line 8: If this is 'extensively studied' is there a review article or meta-analysis that could be cited here instead of a single study?

**Authors' response:** Thank you for pointing this out. This section now reads: "Prospective thinking refers to humans' ability to mentally simulate the future (for a review, see Schacter et al., 2017). It allows individuals to effectively prepare for upcoming events and facilitates the enactment of goal-directed actions and the planning of behaviors, including health behaviors (Brevers et al., 2023; D'Argembeau et al., 2010; Schacter et al., 2017). A core feature of prospective thinking is that it enables one to flexibly retrieve and recombine past information into mental simulations related to future events (D'Argembeau et al., 2010; Schacter et al., 2017). These memory-based processes have been extensively studied with experimental tasks that involve the extraction of information about locations, objects, people, as well as more schematic and conceptual knowledge to envision general goals or events (Schacter et al., 2017). Humans can thus engage in different forms of prospection, including episodic future thinking (for example, by imagining themselves in a particular place at a specific time, bringing specific details to mind) and semantic future thinking (i.e., thinking about the future in a general, abstract manner; Demblon and D'Argembeau, 2014)."

**Comment 8:** Page 3, Line 16: The phrasing "this current gap of knowledge stems from the adoption of rating of perceived exertion…" doesn't really make sense, need to rephrase.

**Authors' response:** This section now reads: "Nevertheless, it is currently unclear how prospective thinking unfolds while generating future predictions about one's own bodily states, such as when anticipating the intensity of perceived exertion (i.e., the subjective intensity of effort, strain, discomfort, and/or fatigue that is experienced during physical exercise; Hutchinson, 2020; Robertson and Noble, 1997) of a forthcoming session of physical exercise. Indeed, the level of physical 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?”; “How was your workout?”; Foster et al., 2001; Haile et al., 2015; Robertson and Noble, 1997).”

**Comment 9:** Page 3, Paragraph 2: Just a note to be very clear with language here. The terms effort and exertion are often used interchangeably when they refer to different things (see e.g., Hutchinson 2021, https://doi.org/10.51224/B1013, Smirmaul 2012, https://doi.org/10.1136/bjsm.2010.071407 or Swart et al. https://doi.org/10.1136/bjsports-2011-090337. In other words, asking “what intensity of effort do you feel now” is not the same as asking for a rating of perceived exertion. In addition (and adding to the confusion), Foster’s Session RPE measure uses the verbal prompt ‘How was your workout?’.

**Authors’ response:** The translation of the term “exertion” in French is “effort” and we made a repeated error in the back translation by using “effort” to refer to “exertion”. The term “effort” has thus been replaced by “exertion” throughout the revised manuscript. In the Introduction section, we added a definition of perceived exertion and also added Foster’s Session RPE (“How was your workout?”) as an example of retrospective RPE.

**Comment 10:** Page 3 Lines 21-25: This is true, but the vast majority of this literature comes from well-trained athletes – is there anything from an exercise population?

**Authors’ response:** We added the following references on RPE studies undertaken among an exercise population: Thiel, C., Pfeifer, K. & Sudeck, G. Pacing and perceived exertion in endurance performance in exercise therapy and health sports. *Ger J Exerc Sport Res* 48, 136–144 (2018). https://doi.org/10.1007/s12662-017-0489-5

**Comment 11:** Page 4: Good section about the link between RPE and pleasure. Where you discuss prospective RPE (Lines 24 & 25), I wonder if this could this be linked to forecasted pleasure as well? See e.g., Hutchinson et al. 2023 https://doi.org/10.1123/jsep.2022-0243.
The described mismatch between anticipated and experienced effort is also in line with research on forecasting error (see Ruby et al., 2011, https://psycnet.apa.org/doi/10.1037/a0021859). It might be useful to briefly make this point here?

**Authors’ response:** Thank you for the positive feedback and for this very useful suggestion. The effect of physical exercise intensity on remembered pleasure has been acknowledged in this section as follows: “Another key observation from the literature on RPE is that increased perceived levels of exertion are negatively linked with the intensity of pleasure felt during the session of physical exercise (for a theoretical review, see Ekkekakis et al., 2011; for recent studies, see Hartman et al.., 2019; Hutchinson et al., 2020; Frazão et al., 2016). It has also been evidenced that decreasing the intensity of a resistance exercise session can elicit higher levels of experienced and retrospective pleasure toward physical exercise (e.g., Hutchinson et al., 2023). Besides, positive changes in hedonic responses during moderate intensity exercise have been linked to future physical activity (Rhodes and Kates, 2015).”

With regard to forecasting error, in our opinion, this type of experiment will provide promising perspectives for future studies on the impact of RPE prediction error on running pleasure (e.g., do people make higher RPE prediction errors in the first half of a running session, as compared to the second half? Is there a difference in the magnitude of RPE prediction error between the beginning and the end of a running session?). Hence, we propose instead to discuss these aspects in the perspective section of the Discussion of our Stage 2 Registered Report when ready.

**Comment 12:** Page 6, Section 2.2: How will participants be recruited?

**Authors' response:** Additional details on participants recruitment have been added as follows: "We will recruit our participants among UCLouvain students (except from the Faculty of Motor Sciences, in order not to interfere with the physical activity programs of the Bachelor/Master of Physical Education and Physiotherapy) who want to participate to our start-to-run study. Participants will be recruited via flyers with a QR code directing them to an online screening tool (LimeSurvey platform). The experimenters will also make announcements in the auditorium (after obtaining the agreement of the Professor in charge of the teaching unit). The online screening tool will first include an informed consent form. An email address and a phone number will be provided to allow participants to ask questions before agreeing or declining to participate in the study. The screening tool will then ask the potential participants (i.e., the ones who have agreed to take part in the study) to complete the International Physical Activity Questionnaire (IPAQ; Craig et al., 2003). Since it is a start-to-run program, we will recruit individuals corresponding to the "low" and "medium" physical activity categories of the IPAQ. To limit the health risks related to running exercise, each participant will then be asked to complete the Physical Activity Readiness Questionnaire for Everyone (PAR-Q+; Warburton et al., 2011). Only participants who receive a "green light" to the PAR-Q+ (i.e., who answered NO to all questions of the PAR-Q+) will be allowed to participate in the study. Participants who do not meet the study selection criteria will be informed and will not participate to the start-to-run program."

**Comment 13:** Page 6, Section 2.2: In the interest of equity, would you consider using the additional questions from the PAR-Q+ as a secondary screening, or requesting a physician referral rather then excluding participants completely?
Page 6, Section 2.2: If it is feasible, a VO2 max test would be preferable to the submaximal estimation – this might enable you to identify HR associated with ventilatory threshold which is theoretically important in this area of study.
Page 7 Line 2: A VO2 max test is considered safe in this population. Also, I don't see the George et al. citation in the references.

**Authors' response:** Thank you for these recommendations. We used the additional questions but did not refer to this in the original version of the manuscript. We therefore adapted to new version of the manuscript (see page 6, paragraph 3): "To limit the health risks related to running exercise, each participant will be asked to complete the Physical Activity Readiness Questionnaire for Everyone (PAR-Q+; Warburton et al., 2011) in the presence of one of the two team supervisors (BdG) who has 20 years of exercise testing experience. In the first step, only the first questions of the questionnaire will be filled out. Those who answer NO to the first 7 questions of the PAR-Q+ receive a "green light" and will be immediately allowed to participate in the study. Those who will answer YES to one or more questions will have to meet the team supervisor who will go through the additional questions (pages 2 and 3 of the PAR-Q+). If participants answer YES to any of the questions, they will not be able to start the study and will be advised to see a (sports) physician."

We agree with the Reviewer that a maximal exercise test under laboratory conditions would be the ideal scenario, but we do not have the necessary budgets and time for such a large group of participants. An alternative is to use a maximal field exercise test (e.g., 20m multistage fitness test), but we prefer not to use a max test because we deal with novice runners.

George et al. citation has been added to the references.

**Comment 14a:** Page 7 Line 9: Referring to "coaches" – what specific training will they have? PE graduate students are not the same thing as running coaches, so I assume some training would be needed.
**Comment 14b:** Page 7, Lines 13-14: Referring to the "general advice about running techniques, nutrition and sport injury prevention", how will this information be standardized?

**Authors' response:** The coaches we refer to are last year Master students (5 years Master program in Movement Sciences at UCLouvain). We are confident that they can bring this task to a successful conclusion considering that these students successfully completed teaching units on exercise physiology and physical exercise training theory at UCLouvain Faculty of Movement and Rehabilitation Sciences. The Master students could also always ask questions to the 2 team leaders who are experienced training supervisors and, if necessary, to the faculty's professor of training theory.

The general information on running techniques, nutrition, and sports injury prevention will be standardized through the articles that are available on the Formyfit blog http://blog.formyfit.com/category/articlesconseils/nutrition/ that was checked by the team leaders. The strength of working with our coaches is that the advice can be individualised to the needs of every runner. This information has been added in the revised manuscript (see page 10, paragraph 1).

**Comment 15:** Page 7 Line 20: Specify what RPE literature you are referring to here. If it is the work cited at the end of the next sentence, I would not describe this as RPE literature.

**Authors' response:** The sentence "This approach is based on RPE literature." has been deleted from the revised manuscript.

**Comment 16:** Page 9, Last Line: Referring to "relative difference will inform on the magnitude…" – Doesn't absolute difference also give this? I'm not sure I understand this?

**Authors' response:** Sorry for the confusion. The usefulness of using both absolute and relative prediction errors have been better detailed: "Absolute and relative indexes of prediction errors complement each other. Specifically, given the same absolute change, the relative change is larger in magnitude if the prospective RPE value is at a higher level than if it is at a lower level. For instance, (i) with a prospective RPE of 3 and a retrospective RPE of 5, the absolute RPE prediction error = -2 and the relative RPE prediction error = -0.33; (ii) with a prospective RPE of 5 and a retrospective RPE of 7, the absolute RPE prediction error is still = -2, but the relative RPE prediction error is now -0.40."

**Comment 17:** Page 10, Section 2.4.2: Why not use the Feeling Scale? The proposed measure is positively oriented (i.e., 6 of the 7 options indicate some degree of experienced pleasure).

**Authors' response:** Thank you for this important suggestion. The main rationale for using such an item is that we aim to adopt a pleasure oriented (instead of a displeasure-pleasure) approach. In this context, we used particular modifiers that are suitably distinct from each other for participants to express differential amounts of pleasure they had experienced during their running session. This approach was based on the Methods section form Stanley and Cumming (2010; https://doi.org/10.1016/j.psychsport.2010.06.010), who also showed that this type of modifiers were easy to understand by the participants. This aspect has been better acknowledged in the Methods section (see page 11, paragraph 4).

Besides, switching from items for indexing pleasure might be problematic at this stage as we already obtained pilot data to sample size estimation with scoring on this pleasure-oriented item as a dependent variable.

Nevertheless, we totally agree that using a measure ranging from displeasure to pleasure is an insightful alternative that could potentially offer fine grained knowledge on our indexes of prediction errors. Hence, we propose instead to discuss this important aspect in the limitation/perspective sections of the Discussion of our Stage 2 Registered Report when ready.

**Comment 18:** Page 10, Section 2.5:1: Needs detail – how are speed and distance recorded? With GPS? If GPS, how will indoor runs or treadmill runs be recorded?

**Authors' response:** This important information has been included as follows (see page 9, paragraph 1): "Running sessions could be undertaken outdoors or indoors (on a treadmill). For the outdoor session, the GPS of the Smartphone is used to estimate the running distance and average speed. When performing an indoor session, the Formyfit app records the time and participants will be informed that they will have to encode the distance manually in the app at the end of the session."

**Comment 19:** Page 10, Section 2.5:2: Regarding group impact. This detail addresses one of my earlier questions, however, will this also be accounted for in the group sessions? The text suggests it is only for the free sessions.

**Authors' response:** This aspect has been acknowledged in the Methods section of the revised manuscript (see page 12, paragraph 3): "Previous research has shown that running in groups impacts the level of pleasantness of physical exercise sessions (e.g., Xie et al., 2020). Hence, we will examine whether running with or without another person during the "free" sessions (running alone vs. running with another person vs.. running with more than one person) or the coaching session per se modulates the impact of RPE prediction error on running pleasure (variable name = running_group)."

**Comment 20:** Page 10, Section 2.5:2: Regarding Habits. How will "degree of habits" be quantified? Also, if I understand, degree of habits refers to familiarity with the route. In this case should the modulation not be the other way round? Whether the degree of habit modulates the impact of RPE prediction error on running pleasure, rather than, "whether the impact of RPE prediction error on running pleasure modulates the degree of habits linked to the running program" (p. 10).

**Authors' response:** The degree of habits indeed refers to familiarity with the route. In order to avoid confusion the measure "degree of habits" has been replaced by "degree of familiarity" (see also **Figure 1A**).

We also agree that it is the degree of familiarity that modulates the impact of PE prediction error on running pleasure, not the other way around. Accordingly, the description of this measure now reads: "We will also examine whether the impact of RPE prediction error on running pleasure modulates the degree of familiarity toward the running route. Indeed, individuals might get better at predicting their level physical exertion for familiar running trails, which can modulate the impact of RPE prediction error on running pleasure."

**Comment 21:** Page 11, Figure 1B: Referring to "retrospective RPE" – This measure is not the same as Borg's CR-10 scale. For example, the CR-10 scale has and an 'absolute maximum' and decimal numbers (e.g., 0.5 and 1.5). Also, I don't see the verbal anchors for either scale? (Bi and Bii)

**Authors' response:** We used the integers and verbal anchors (0 = "null" (in French "nulle", 1 = "very very light" ("très très légère"), 2 = "light" ("légère"), 3 = "moderate" (Modérée), 4 = "somewhat hard" ("assez dure"), 5 = "hard" ("dure"), 6 = [no verbal anchor], 7 = very hard ("très dure"), 8 = [no verbal anchor], 9 = [no verbal anchor], 10 = maximal (maximale) of the French adaptation of the CR-10 scale (Haddad et al., 2013). See also page 10, paragraph 3 of the revised manuscript. It is also important to acknowledge that we made errors in the previous description of the verbal anchor of the CR-10 scale (1 = "very light", 2 = "light", 3 = "moderate", 4 = "somewhat hard", 5 = "hard", 6 = "hard +", 7 = very hard, 8 = very very hard , 9 = extremely hard). These errors are now fixed.

**Comment 22:** Page 12: Will the model be adjusted for the described covariates (running distance, running speed, group impact, habits)? I did not see this described.

**Authors' response:** This is correct. The models described in the previous version of the manuscript were the ones that were used for testing the pilot data (which included covariate measures on "distance" and "average_speed", but not covariate measures on "group", "familiarity" and "music"). We apologize for the confusion. The data analytic plan section now accurately describes the full models that will be tested in the main study (see pages 14-15). Besides, the pilot data section now describes the models that were used in the pilot (i.e, which only included the covariate variables "distance" and "average_speed").

**Comment 23:** Page 14: Did you gather any feasibility information from the pilot study? What was the retention rate? Useability of the app, etc.?

**Authors' response:** We did gather such specific types of information. Nevertheless, as discussed in our response to **Comment 2**, participants were allowed to keep comments after each running session of the pilot study, which gave us some insight on the difficulty encountered by the participants when trying to run at the specific pace advised by the Formyfit progressive running program.

**Comment 24:** Page 14, Line 11: Replace "linear mixed model" with "LMM".

**Authors' response:** Done. Thank you.

**Comment 25:** Page 14, Line 17: I am a bit confused by the reference to the variance in individual empowerment?

**Authors' response:** Thank you for pointing out this error. This sentence now reads: "As shown in **Table 1**, the intercept variance is .37 and the within-participant variance is 1.38."

**Comment 26:** Table 1 and 2. Were the indicated covariates included in these models?

**Authors' response:** The models ran in the pilot study included covariate measures on "distance" and "average_speed", but not covariate measures on "group", "familiarity" and "music". See also our response to **Comment 22**.

**Comment 27:** Page 16, section 2.8: If I understand correctly, the pilot data indicated effect sizes of .42 and .41? Then why not use .40 (rather than .50) for the estimation? The estimation suggests that 16 participants will need to run a minimum of 5 sessions? If so, it might be clearer to write this as "…16 participants along five measurement points (i.e., five running sessions) is required …". At first, I thought it meant five measurement points within each run.

**Authors' response:** Thank you for having spotted this error. We have now run the power analysis based on conditional $R^2$ of .41 (from step 2 model of relative prediction error). This information has been added in the revised manuscript as follows: "In line with recent guidelines that suggest running power analysis based on the lowest meaningful estimate of the effect size (Dienes, 2021), we ran sample size estimation analyses with a conditional $R^2$ of .40 based on the results obtained in the step 2 model of relative prediction error. 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 28 participants along 5 measurement points (i.e., a running session) is required for third-step models."

**Comment 28:** Design Table: Perhaps it was cut off, but I do not see the last column ("theory that could be shown wrong by the outcomes") in the table.

**Authors' response:** This information was missing and has now been added to the Design Table. Thank you!

## REVIEWER 2

**General comment:** The authors present a stage 1 manuscript to investigate how error predictions of perceived exertion are related to feelings of pleasure. I commend the authors on their willingness to do a registered report, detailed methods, and their sharing of data and code. I view this is as an opportunity to try and help the authors improve their future study (planned to start in October, 2023). I have a few major comments and some minor comments.

**Authors' response:** Thank you very much for the positive reception of our manuscript and for your insightful and in-depth comments.

### Major comments

**Comment 1:** In several places in the manuscript, the authors refer to "experienced level of running pleasure" (e.g., abstract, main text). This is an important note because it is one of the primary variables. However, the authors are not measuring experienced pleasure. They are planning on measuring retrospective ratings of pleasure, with measurement taking place after the running session. This would be more appropriately referred to as remembered pleasure than experienced pleasure. This is retrospective, and retrospective evaluations do not perfectly align with moment-to-moment experienced pleasure.

**Authors' response:** Thank you for this insightful suggestion. The terms "experienced running pleasure" have been replaced by "retrospective pleasure". In the title, we also replaced "pleasure while running" with "running pleasure".

**Comment 2:** In the methods, when "Running pleasure" is introduced as a variable, the authors should describe the model of affect that they are adopting. If they conceptualize pleasure-displeasure as bipolar (which I would suggest, see Russell, 1980), then they should allow for the measurement of displeasure. Currently, they only allow for no pleasure or extreme pleasure, and do not allow runners to report levels of displeasure. I strongly encourage the authors to adopt a bipolar measure that allows for the measurement of displeasure.

**Authors' response:** Thank you for this important suggestion. The main reason for using this item is because we aim to adopt a pleasure oriented (rather than a displeasure-pleasure) approach. In this context, we used particular modifiers that are suitably distinct from each other for participants to express differential amounts of pleasure they had experienced during their running session. This approach was based on the Methods section form Stanley and Cumming (2010; https://doi.org/10.1016/j.psychsport.2010.06.010), who also showed that this type of modifiers were easy to understand by the participants. This aspect has been better acknowledged in the Methods section (see page 11, paragraph 4).

Besides, switching from items for indexing pleasure might be problematic at this stage as we already obtained pilot data to undertake sample size estimation with this pleasure-oriented item as a dependent variable.

Nevertheless, we totally agree that using a measure ranging from displeasure to pleasure is an insightful alternative that could potentially offer fine grained knowledge on our indexes of prediction errors. Hence, we propose instead to discuss this important aspect in the limitation/perspective sections of the Discussion of our Stage 2 Registered Report when ready.

**Comment 3:** On page 4, paragraph 1, the authors could also discuss the findings of Rhodes & Kates (2015). https://doi.org/10.1007/s12160-015-9704-5

**Authors' response:** Thank you for this suggestion. The following sentence has been added to this paragraph: "Another key observation from the literature on RPE is that increased perceived levels of exertion are negatively linked with the intensity of pleasure felt during the session of physical exercise (for a theoretical review, see Ekkekakis et al., 2011; for recent studies, see Hartman et al., 2019; Hutchinson et al., 2020; Frazão et al., 2016). It has also been evidenced that decreasing the intensity of a resistance exercise session can elicit higher levels of experienced and retrospective pleasure toward physical exercise (e.g., Hutchinson et al., 2023). Besides, positive changes in hedonic responses during moderate intensity exercise have been linked to future physical activity (Rhodes and Kates, 2015)."

**Comment 4:** Given the importance of affective responses experienced while exercising, why not measure RPE and affective valence during exercise too? Why not also measure anticipated affect and remembered affect as well? Is this not possible, technically?

**Authors' response:** While it could be valuable and technically possible to obtain such ratings, we decided not to include ratings of physical exertion and running pleasure during the session.

A first reason is statistical parsimony, that is, to favor a simpler model with fewer parameters to test the main research question of the present study. The second main reason is that, in our start-to-run program, we will let participants choose the running trails, the frequency, and speed of their running sessions. In this context, (i) in-session ratings might be biased by the remaining distance to cover (e.g., teleoanticipation process), and (ii) it would also be difficult to adopt a specific in-session ratings (e.g., should we include one of several in-session ratings? What about short runs versus long runs? What about ratings undertaken directly prior or after a difficult section of the run?).

On a theoretical level, we would like to acknowledge that, by letting them choosing the running trails, the frequency, and speed of their running sessions, our start-to-run program fits well with training procedures derived from the ecological dynamic approach to physical exercise (e.g., David et al., 2016 https://link.springer.com/article/10.1007/s40279-016-0511-3; Rudd et al., 2021 https://doi.org/10.1080/17408989.2021.1886271). Specifically, this approach advocates for physical exercise behaviors that consider the relationship between individuals' characteristics and functional aspects of their environment (e.g., running session undertaken under multiple contexts). This aspect is now detailed in the Methods section (page 8, paragraph 1): "This approach fits also well with training procedures derived from the ecological dynamic approach to physical exercise (e.g., David et al., 2016; Rudd et al., 2021). Specifically, this approach advocates for physical exercise behaviors that consider the relationship between individuals' characteristics and functional aspects of their environment (e.g., running sessions undertaken under multiple contexts)."

Again, we totally agree with the reviewer that adding in-session ratings could bring insightful knowledge on our indexes of error prediction, but while adopting with more controlled/standardized physical exercise procedure (e.g., to make participants running/walking multiple times on the same routes, and to ask them to run at a "fast", moderate or "slow" between each different session). Hence, we propose, again, to address this important aspect in the limitations/perspectives sections of the Discussion of our Stage 2 registered report when ready.

For instance, a concept such as forecasting error (e.g., Ruby et al., 2011, https://psycnet.apa.org/doi/10.1037/a0021859) could provide perspective for future studies on the impact of RPE prediction error on running pleasure (e.g., do people make higher RPE prediction errors in the first half of a running session, as compared to the second half? Is there a difference in the magnitude of RPE prediction error between the beginning and the end of a running session?).

**Comment 5:** The authors say "Importantly, this can explain why some people find their physical exercise unpleasant...". I think I understand what the authors are trying to convey, but the link between pleasure and perceived exertion (from the prior sentence) itself does not seem to explain the affective rebound. They seem like separate concepts. In other words, people seem to experience displeasure during exercise followed by an increase in pleasure after exercise, but I am not sure that this is explained by the fact that perceived exertion and pleasure seem to be negatively associated.

**Authors' response:** We agree with the reviewer that this interpretation is not necessarily accurate. This sentence has, herefore, been deleted from the paper.

**Comment 6:** The authors mention the safety of the SET. While true, this sample answered no to every question on the PAR-Q+, and should be able to safely do a maximal test. Therefore, I'm not sure safety is a good justification here.

**Authors' response:** As what was answered to Comments 14 and 15 of Reviewer 1, we agree that a maximal exercise test (under laboratory) conditions would have been the best option. Nevertheless, we still prefer to keep the submaximal exercise test. We do this for different reasons:

- National regulations and Ethical approval are very strict when it comes to maximal exercise testing. Maximal exercise testing is allowed only if a physician (MD) is present at the site of the testing, which was not possible for this study because of time and budget.
- A submaximal exercise test limits the risks of injuries or 'over-tiring' the participants who would then stop participating.
- The aim of the exercise test was not to have the best estimate of the 'real' $VO_2$max, but to be able to see the progression of their fitness level between the start and the end of the start-to-run.

## Minor Comments

**Comment 7:** The authors mention that they will report the intention-to-treat analysis. While useful, I encourage the authors to also report a per-protocol analysis. Intention-to-treat is great, but both could be maximally informative especially if dropout seems high.

**Authors' response:**. Thank you for this comment. The 'intention-to-treat principle' is only interesting to see if the fitness level changed (i.e., a training effect), but is actually not of key importance for our analysis (RPE) as those participants that do not run do not fill out the RPE and running pleasure measures. Indeed, the primary goal of the start-to-run program is to provide a context that will allow participants to perform enough running sessions (minimum 5; see also the **Sample size estimation** section) to test our hypothesis on the impact of RPE prediction errors on running pleasure (see also page 7, paragraph 3). Therefore, in order to avoid confusion, we deleted the sentence that referred to the intention-to-treat procedure.

**Comment 8:** There are some instances of RPE referring to rating of physical exertion, but in most cases it is rating of perceived exertion. Please be consistent.

**Authors' response:** Thank you. RPE refers now to the rating of perceived exertion throughout the revised manuscript.

**Comment 9:** In the description of the relative index of RPE prediction error, it seems that the parenthetical suggests the text should read "subtracting the score of retrospective RPE from prospective RPE". Please clarify.

**Authors' response:** Thank you for having spotted this mistake. We have made the requested change.

**Comment 10:** In the design table, I encourage the authors to also report the interpretation if their hypotheses are not supported.

**Authors' response:** Thank you for this suggestion. This information has now been included in the design table.

**Comment 11:** On page 4, there is an extra period after the Hartman reference; The comma after "A key tenet from the literature on reward processing," can be removed; On page 7, I think "fell" should be "feel"; On page 7, there is an extra comma in "In addition, to the weekly coaching sessions"; I think "technics" should be "techniques" (also page 7); On page 7, "greater sense of autonomy toward physical exercise, *but* also increased...". I think but can be "and".

**Authors' response:** Thank you for your thorough reviewing of our paper. These spelling and grammar issues are now fixed. In addition, the manuscript has been reviewed for spelling and grammar by a professional scientific editing service.

**Concluding comment:** Thank you for allowing me to review this project. I think it has promise, and I hope that my comments are helpful. I especially encourage the authors to strongly consider their conceptualization and measurement of affect, and whether they are interested in experienced pleasure, remembered pleasure, or both (I encourage both).

**Authors' response:** Thank you again for the positive and insightful feedback!