March 24th, 2023

Dear Dr. Veli-Matti Karhulahti:

Please find enclosed the revised version of our manuscript, *Optimizing Esports Performance Using a Synergistic Mindsets Intervention* (PCI Registered Reports #364) (word count = 19 663). The revised manuscript is 79 pages long and includes one table and three figures.

We are grateful for the thoughtful comments. These suggestions helped us to strengthen our paper. We believe that we have been able to address each comment/suggestion, and in what follows, we did our best to facilitate your reviewing process. We provide a point-by-point reply to the Reviewers' final comments. Each comment (in bold) is followed by our reply and an indication of where in the manuscript we have addressed the comment. Revised sections are highlighted in green in the manuscript file.

We thank you again for considering our work and for helping us strengthen our manuscript. We look forward to hearing from you and hope you will find this revision appropriate for IPA in *PCI Registered Reports*.

Sincerely yours,

Gratefully,
Maciej Behnke
Daniël Lakens
Kate Petrova
Patrycja Chwiłkowska
Lukasz D. Kaczmarek
Jeremy P. Jamieson
James J. Gross

Veli-Matti Karhulahti (Recommender)

Thank you for all careful revisions and responses. I have now received all reviews and the reviewers collectively agree that the work is almost ready for in-principle acceptance. There are a few minor reviewer comments that I encourage you to consider in your final revision. Note that this round we had one more expert who was unable join the review in the first round – this ensures that, at Stage 2, we will have experts who are familiar with the Stage 1 plan even if someone is unavailable next year. I leave a few brief notes of my own.

Reply: Thank you!

#1. As a follow-up to my earlier comment #3 where I referred to the gaming disorder scale as an exclusion criterion, I think it's worth giving it a bit more thought. Since participants are paid for playing games and you might learn that some meet gaming-related diagnostic criteria, it could strengthen the study to have a more explicit plan regarding participants in this hypothetical risk group.

Reply: After the discussions within the team and consultations with the clinical psychologists about recent approaches to problematic gaming in Poland, we changed our approach and clarified it in the revised version of the manuscript.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

We will exclude participants who meet diagnostic thresholds for problematic gaming. To screen participants, we will use the Gaming Disorder Test (GDT; Pontes et al., 2021; see Supplementary Materials for details), a validated psychometric test (Cudo et al., 2022; Karhulahti et al., 2021) developed to assess gaming disorder defined in the International Classification of Diseases (ICD-11; World Health Organization, 2018). The GDT will be a part of the online study registration. Any participants who meet the criteria for gaming disorder (i.e., endorsement of all four diagnostic criteria as assessed by each GDT item: marking '4: Often' or '5: Very often'; Pontes et al., 2021) will be evaluated by a clinical psychologist (retained as a consultant on the project). Participants who are judged to have a diagnosis of gaming disorder will be excluded from participation.

#2. I suggest a minor revision for the justification of exclusions (p. 13). The age limit is clear, but the other exclusions don't seem to follow logically: "We will recruit Polish-speaking players as the study will be run in Poland. We will recruit only male players due to their predominance (76%) among first-person shooter gamers." I believe in both cases the justification is feasibility, along the following lines: "Because including non-Polish and non-male participants would entail producing and testing different sets of group-specific research materials, the study will include only Polish male players" (just an example, feel free to rephrase as you see best or rebut).

Reply: We clarified this issue.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

We will recruit Polish-speaking players as the study will be run in Poland. We will recruit only male players due to their predominance (76%) among first-person shooter gamers (Statista, 2023). Including non-Polish and non-male participants would entail producing and testing different group-specific research materials. Furthermore, gender and language might become confounding factors to the study, which we would not be able to test adequately due to the expected small number of eligible participants from these groups.

#3. This comment doesn't need a response, but I want to leave it at Stage 1 in case it will be discussed at Stage 2. Note that in the design table column "theory that could be shown wrong" you only name the synergistic mindset model. I agree it's good to be very careful and selective about theoretical inference, but at the same time I am thinking whether the results might be theoretically informative also beyond this single model – after all, the synergistic model stands on other established theory. Preregistered meta-theoretical inference for the upcoming discussion section could be informative for the research program's development at larger scientific scale.

Reply: In the design table (where specifying a theory that could be shown wrong), we added two categories of models and theories that can be a) mainly (i.e., synergistic mindsets model), and b) partially shown wrong.

As for the models that can be partially shown wrong, we added: the biopsychosocial model of challenge and threat (Blascovich, 2008), the growth mindset model (Dweck & Yeager, 2019; Yeager & Dweck, 2020), the arousal reappraisal model (Jamieson et al., 2018), the stress-can-be-enhancing mindset model (Crum et al., 2013, 2017).

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

Mainly: the synergistic mindset model (Yeager et al., 2022);
Partially: the biopsychosocial model of challenge and threat (Blascovich, 2008), the growth mindset model (Dweck & Yeager, 2019; Yeager & Dweck, 2020), the arousal reappraisal model (Jamieson et al., 2018), the stress-can-be-enhancing mindset model (Crum et al., 2013, 2017);
In case of mixed findings (not significant and not equivalent), we do not draw full theoretical implications for alternative or null hypotheses.

Lee Moore (Reviewer 1)

#1. The authors have done a good job of revising the registered report based on my feedback and suggestions (although it is worth them reading my published work more closely when describing how they will score the demand and resource evaluation data - i.e., substract evaluated demands from resources to get a score ranging from -5 to +5). While I could quibble with one or two of the authors responses, overall, the registered report is excellent and describes what will be a highly rigorous and superb piece of research in a comprehensive, accurate, and replicable way. It has been an interesting process reviewing this registered report and so thanks for the opportunity to be involved. I wish the authors all the best with the data collection and analysis phase, and I look forward to seeing the final write-up in due course.

Reply: Thank you for your kind words! We incorporated your suggestion and changed how we plan to operationalize the challenge/threat ratio using the approach presented in Moore et al., 2013; 2014, not in Moore et al., 2012.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

A ratio will be calculated by subtracting demands from resources (range: -5 to +5), with a more positive value reflecting a challenge state and a more negative value reflecting a threat state (Moore et al., 2013, 2014).

Ivan Ropovik (Reviewer 2)

Thanks to the authors for considering my suggestions. As I already expressed in my review of the first submitted version, I think that the proposed study will be informative and may serve as one of the good-practice examples in the field. Therefore, not now do I see any "disqualifying factors", to use authors' words. Being pragmatic about research has its merits. Every study has strengths and weaknesses and it is completely fine as long as the writing is clear in that manner and provided that the weaknesses of the design do not disproportionately warp the reflection of the underlying studied phenomena.

My two main worries about the reliability of the measurement and the unbalanced demand characteristics of experiment conditions remain, but I also get the other side of the coin, namely that the authors chose to optimize for a greater cumulative potential of the proposed research with respect to the existing evidence in the field.

Any one of these two possibly biasing factors (measurement error does not have to be just random noise) or their interaction can lead to false positive results, so laying out these possible weak links in the limitations section seems important to me.

Below, I offer a few follow-up comments/responses on authors' edits and replies. Regarding the issues that I do not return to, I was either satisfied or okay with the proposed revision/rebuttal. Overall, I think that there are no outstanding issues that should prevent the authors from running the study as proposed and am happy to hand over the final call on any further revisions to the discretion of the editor.

Reply: Thank you!

#1. "As requested, we have included a critical interpretation of the existing literature in the Introduction. However, we respectfully disagree that psychophysiological challenge/threat or affect regulation research provides "weakly informative designs." But, we acknowledge that you might have a different opinion on this topic."

As I said, I have no expertise in literature on reappraisal. Why I made such a bold claim? Some time ago, I was doing an internal review of a protocol of one large multi-site study studying the effect of a similar reappraisal intervention. The lead authors explicitly chose this type of affect regulation intervention because it proved to have the strongest effect in the meta-analysis by Webb et al. (2012). In my view, this is always a poor strategy, for various reasons. This is also the metaanalysis used in the present RR to get an expected effect size. For the sake of the review and protocol revision in that past study, I looked at the included studies in detail and I carried out a re-analysis using arguably more state-of-the-art methods. I also looked at studies from another systematic review of reappraisal interventions by Cohen & Ochsner (2018). What I found and documented was an array of studies having various methodological issues (sizeable proportion of experiments lacking a control group), mostly feeble manipulations on tiny samples yielding huge effects, indications of selective reporting (p-values < .01 completely lacking, e.g., six out of seven available focal tests of the claims for mixed reappraisal had a p-value ~ .04), or study-level data patterns inconsistent with expectations under both, H0 or H1. Since you are dealing with the reappraisal literature, that is why I tried to voice my concerns about the robustness of the given literature and a need for a more critical appraisal of the evidence reported in that literature. Of course, I looked only at a slice of that literature, but that slice did not spur much confidence, to say the least. If interested, I'm putting this part of my review dealing with the re-analysis of Webb et al. here:

https://docs.google.com/document/d/1Q_8134QurWIKdUEzmJRjs7SZN9aJpUiPnZxbTe_3KiA/edit?usp=sharing
Analytic output is available here: https://rpubs.com/ivanropovik/592468
Code with data are available here: https://github.com/iropovik/PSAcovid002review

No need to react to that on your part, I just wanted to back up the bold claim from my review of your RR.

Reply: Thank you! We appreciate your healthy skepticism about the reappraisal literature.

#2. Response letter: "The PCI RR community might be up to date with the novel approaches for sample estimation" … expected effect sizes based on the scientific literature are often also effects of interest to scholars in these fields (Lakens, 2022).

Offering an overview of effect sizes for the given effect from past literature is fine. Trying to provide a sample size justification (in a field where it is still rare) is great. So I'm not insisting that it is to be removed.

But I cannot help myself but seeing a "power estimation" or "sample estimation" based on previous findings or expectations in general as just bad practice that exacerbates the issue of studies having low power to detect effect sizes that would be considered theoretically relevant.

Power is a pre-data concept, just like alpha. Specifically, it is the sensitivity of a given statistical test to reliably detect a range of *hypothetical* population effects of interest.

Btw, that is the definition also put forward by Morey & Lakens (2016). This paper lists 4 misconceptions of power, and the treatment of power in the present RR is, in fact an example of misconceptions #3 (Sample size choice should be based on previous results) and #4 (Sample size choice should be based on what the effect is believed to be).

Implicitly also misconception #1 (Experiments have actual/observed/post hoc power). From the paper: "the power of a test depends not on what the effect size is, but rather on all the hypothetical values it could be. … Attempts to "estimate" the power of an experiment, or a field, are based on a misunderstanding of power and are uninformative and confusing (Goodman & Berlin, 1994; Hoenig & Heisey, 2001; O'Keefe, 2007)." (p. 17). There is also still conflation (in the manuscript and response letter) of power analysis with sample size determination (the issue of powering for primary vs secondary hypotheses).

Anyway, if the sample size determination was mainly based on resource constraints (beta = .22 would seem a strange target to many readers), I think it is completely fine (and the most honest way) to say so and just let the reader see the sensitivity of the design across a range of effect sizes (or alternatively, across Ns as you do). IMHO, that would be better than reinforcing misconceptions about power and touring the reader through a determination of SESOI and wide array of past ESs, and arriving at the inability to formally reconciliate the two approaches and setting an arbitrary target of r = .22.

That said, all this is inconsequential w.r.t. the informativeness of the present design, so let's leave it at that.

Reply: We will address this issue in the discussion. Thank you for bringing it up. Daniel Lakens would like to add that he does not see any conflict between reporting a power analysis

based on expected effects in the literature, and his observations in Morey & Lakens, 2016. A study has no power - but that does not mean it is not worth reflecting on which power a study would have, given effects in the literature. This might not lead to conclusions about the power a study has for the true effect size, but it allows for conclusions about the power a study has, if readers assume effect sizes in the published literature are accurate. These might not be accurate (although they might be, as not all literatures are equally biased) but readers might find these reported effect sizes 'of interest'. And therefore, reporting which power a study has for these effects of interest (while acknowledging there are other effect sizes of interest) is coherent with what we wrote in Morey and Lakens, 2016 (see also Lakens, 2022).

#3. Response letter: "We added the attention check in the questionnaire sets to screen for careless responding. Now item #59 states:" "Please select "Strongly disagree" for this item to show that you are paying attention."

Just reiterating from my review, exclusion of careless responders should only be applied using pre-treatment measures, as carelessness itself may have been affected by the treatment, and exclusions based on that would induce bias into your model. You may also consider the methods for detecting careless responding patterns in the "careless" R package. One attention check is better than nothing, but there are more advanced methods.

Reply: We included the attention check as a pre-treatment measure. We only did this because we will collect the questionnaire data in the lab (beginning of lab sessions) rather than via an online survey; thus, we do not expect careless responders. However, as suggested by the reviewer, we reviewed other methods to identify careless responders (Curran, 2016). We will flag the careless responders if they:

1) select other answers than "Strongly agree" for the item: "Please select "Strongly agree" for this item to show that you are paying attention."
2) answer the whole baseline questionnaire faster than 3 minutes (2 s for an item, Huang et al. 2012, plus additional time for reading the scale description)
3) answer the whole baseline questionnaire with a string of identical responses greater than 40 items (half the length of the total scale; Curran, 2016)
4) answer to the last item of the baseline questionnaires – "In your honest opinion, should we use your data in our analyses in this study" – "No, I responded carelessly.".

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*, 299-311.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

We will also exclude participants identified as careless responders. We will identify the participant as a careless responder if he:

1) selects other answers than "Strongly agree" for the item: "Please select "Strongly agree" for this item to show that you are paying attention."

2) answers the whole baseline questionnaire faster than 3 minutes (2 s for an item; Huang et al. 2012, and some extra time for reading the scales description)
3) answers the whole baseline questionnaire with a string of identical responses greater than 40 items (half the length of the total scale; Curran, 2016)
4) answers to the last item of the baseline questionnaires – "In your honest opinion, should we use your data in our analyses in this study" – "No, I responded carelessly" (Meade & Craig, 2012).

#4. Response letter: "Our software (Microsoft Forms) does not allow us to randomize the order of the items within the scales. … As some studies suggest, the order in which items are presented or listed is not associated with any significant negative consequences (Schell et al., 2013) and does not cause differences in average scores (Weinberg et al., 2018)."

There may or may not be order effects, as it is a highly idiosyncratic effect tied to the actual content of what is being measured. Within-block randomization is the way to play it safe. You can always use say three different forms. For instance, just three forms have been found to perform relativelly well compared to the full randomization in planned missingness designs, if the latter is not possible (e.g., paper-pencil data collection). The downside is a bit more complicated administration and joining of the data. I leave it upon authors' discretion if they think it is worth it.

Reply: As we did not find enough evidence to expect that the within-scale item randomization might influence the response patterns, we decided to only keep the within-block randomization.

#5. "If we find biologically impossible values, we will delete them. We will report the number of outliers for a given variable."

Not necessary. I'd rather try to be very conservative with the exclusion of outliers (or specific values) just as you propose and only remove very improbable or impossible values. What is more important than, e.g., counting the number of outliers, is to run the entire analysis twice, with the MAD > 3 outliers in, out, and thus checking if the decision to remove outliers or not has any material impact on any of the main substantive findings. Reporting even that in a short paragraph in the results would be nice, IMO.

Reply: We do not believe it is 'conservative' to leave values that are inconsistent with what biology tells us is possible in our analyses. Following the same reasoning, it would be 'conservative' to introduce incorrect extreme values in any analyses - however, that would simply be equally wrong. As scientists, we believe it is better to take reality into consideration when deciding which values to include in an analyses than to keep impossible values in an analysis. Therefore, we will ignore this comment by the reviewer in our analysis plan.

#6. "The support for the hypotheses will be provided if the models fit the data well, i.e., RMSEA < .06; SRMR < .0 08, CFI > .95, $\chi 2$ > .05 (Bentler, 1990)"

It is the p-value of chi^2 that should be > .05. Your model will fit very well if the actual chi^2 value would approximate the degrees of freedom. Having a threshold for the chi^2 value would be uncommon. There is also a typo in SRMR threshold.

Reply: Corrected.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

#7. "If the fit indices suggest model misfit, we will not be interpreting effect sizes."

A df = 37 model does not offer a hell of a lot dimensions of data space along which the model could be rejected, but from my experience, there is a pretty decent chance that the chi^2 will point to significant global misfit between the model and the data. This is a really risky little note in an RR :D
Anyway, this is a terribly strict requirement. I'm definitely not saying to disregard evidence against the exact or even approximate fit. Chi^2 test is the only formal test of a model and the best guard against misspecified models. A significant chi^2 only tells you that there may be a misspecification in the model and that you need to take a closer look. Taking a deeper look is essential in such case. I think it is reasonable to plan the following:
If the exact model-data fit hypothesis will be rejected, a set of careful diagnostic procedures to identify the possible local sources of causal misfit will be carried out (examining the matrix of residuals and modification indices). The fit would therefore be regarded as adequate if either (1) the exact fit test (chi^2 test) did not signal significant discrepancies between the data and the model or (2) if there was no larger pattern of substantial residuals (say > .1) indicating systematic local misfit.
With some reservations, a disconfirmed model can still be useful and its estimates can still have interpretational value provided that the fit of the model is not very bad – CFI, TLI way below .9, chi^2 being higher multiples of the model df.
You also need a contingency plan for model modification, if this is the case. Half data-driven, half theory-driven careful modifications are less of an evil than interpreting a badly fitting model, where the serious misspecifications propagate through the entire model (or throwing the data away).

Reply: Thank you for this suggestion. We provided more details on the model fit issue in the revised manuscript.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

We will evaluate multiple fit indices as evaluating any single index might be problematic (e.g., a significant χ^2 test does not have to imply the model misfit, as the significance of the test can be affected by many factors, including clustered data, non-normal data big samples; Bergh, 2015; Geiser, 2012; Kenny, 2023). We will not interpret effect sizes if χ^2 test for model fit and all fit indices suggest model misfit. If the χ^2 test detects beyond-chance discrepancies between the model and the data (significant p-value), we will examine the possible local sources of a causal misfit by examining the matrix of residuals for correlations and modification indices. If the modification indices suggest some small model modifications that also have a theoretical foundation, we will include them in the model. Then, if χ^2 test still suggests model misfit but (a) there are no large modification indices and/or residuals, (b) all other fit indices suggest model fit, and (c) there are no Haywood cases (e.g., negative variances, standardized coefficients above 1.00), we will conclude that the theoretical model is likely to be close to the observed reality and we will interpret the effect sizes.

Finally, in the exploratory analysis, we will test other reasonable SEM models (e.g., one mediational model for affect and a second for the cardiovascular challenge).

#8. "We did not find strong enough evidence on whether these factors moderate the effects of synergistic mindset intervention on cardiovascular and performance outcomes (Yeager et al., 2022) to include them in the primary model."

Effectively, only examining effect that proved to be significant in past research goes against the nature of scientific inquiry. If you are not interested to testing a moderation hypothesis, that is completely fine, but I think it should be said directly – that you chose not to test moderation within your primary model. Period. The current justification is a bit awkward.

Reply: We clarified this issue.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

We decided not to test moderation within the primary model, as we did not find strong enough evidence on whether these factors moderate the effects of synergistic mindset intervention on cardiovascular and performance outcomes (Yeager et al., 2022).

#9. Response letter: "However, we would like to keep the option of using the overall negative and positive affective experience scores by averaging the four negative affective experiences in the exploratory analysis. Although it might not be a pure robustness check for our conclusions, in this way, we will be able to observe the difference between the most popular operationalizations of affective experience and statistically superior options.

Such contingency is completely fine and even desirable! What I was objecting only is to qualify the robustness of the results by using a psychometrically inferior model – an unweighted sum score.

Reply: Thank you for the clarification.

#10. Response letter: "If we cannot use multiverse analysis, we will run multiple models and report the results in supplementary materials. After eliminating the different operationalizations of affective experience, we counted 72 possible models (3 options for affective experience x 8 options for cardiovascular measures x 3 options for game measures). This analysis aims to describe the range of effect estimates based on all reasonable data analytical decisions."

Completely agree about the fact that planning things like multiverse analyses is not a concern at this stage. That said, what you are describing is in fact a form of multiverse analysis! You don't need any expert on that, IMO. It's fine just to run all these possible models representing different design options and report what distribution of effect sizes for a few selected focal estimates did you find. Sure, having a script running through all combinations and spitting out a nice multiverse visualization is a great feature but you can do it easily later on "by hand" too. E.g., estimating the effect size and SE (or CIs) for each model, ordering by effect size and easily plotting them using a forest plot (see a very convenient forest() function in the metafor package). The reader would then easily

see the distribution of effect sizes, what proportion their CIs cross zero, etc. Or, alternatively, even only briefly describing the distribution of effect sizes verbally would be far better than nothing.

Reply: We will keep the idea of using meta-analytic tools if we cannot use multiverse analysis. Thank you!

Thanks again for the opportunity to discuss the design of this important study with you. Good luck with the study.

Reply: Thank you! We appreciate your thought-provoking comments.

Ivana Piterová (Reviewer 3)

#1. The authors have incorporated my comment into the code, so I have no further comments on the code at this stage.

Reply: Thank you!

Jacob Keech (Reviewer 4)

#1. As I have joined the review process after a very extensive and well-articulated first round of revisions, I have very few additional comments. I have read the manuscript, materials, and response to round 1 reviews thoroughly. My overall assessment is that this is a well-designed study which has been thoroughly described in the revised Stage 1 Report. The authors have also thoughtfully responded to the comments in the first round of reviews. The application of the synergistic mindset intervention to optimizing esports performance is an innovative idea and I am sure the results of the study will have substantive theoretical and practical value.

Reply: Thank you!

#2. One minor point is that I note the concerns raised by Reviewer 2 about the control condition. I agree that the authors' decision to retain the original procedure for the control condition is reasonable. This will allow comparability, and it has been tested using a large number of participants in the prior studies testing the synergistic mindset intervention. However, on page 12 of the revised manuscript, the control condition is still referred to as a "placebo control". I recommend dropping the placebo wording as the Yeager et al. (2022) paper did not describe the control conditions as a placebo control, and as has been discussed in prior reviewer comments, it doesn't appear that there are matched expectancies across conditions.

Reply: Corrected.

Changes in the manuscript: See XXX section, line numbers: XXX - XXX.

#3. I wish the authors all the best with conducting their study, and I look forward to reading the about the results in the full paper.

Reply: Thank you!