March 7th, 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 = 18656). The revised manuscript is 68 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 your consideration of our work and for helping us to strengthen our manuscript. We are looking forward to hearing from you and hope that 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 submitting a highly rigorous Stage 1 proposal to PCI RR and giving us the opportunity to assess it. I have now received three reviews, two with highly detailed feedback regarding various aspects of the study and one specifically verifying computational reproducibility. We had one unfortunate reviewer cancellation in the process, which delayed decision, but I believe the present reviews were worth the wait and provide highly useful comments that help making final improvements to your plan.

We fully agree with the reviewers that it's important to have this study carried out, so please utilise the rich feedback in a way that is most useful for your purposes (considering practical limits). I add a few comments as well; again, take what is valuable and skip the rest.

Reply: Thank you for your kind words!

#1. As the reviewers point out, it would be good have explicit inclusion/exclusion criteria. It is mentioned on p. 11 that inclusion requires 6h/week of CSGO, but how is

this measured (baseline item #5?) and is this the only criterion? Does this mean veterans cannot participate if they don't train/play anymore? What about age, language?

Reply: Thank you for pointing out this issue. In the baseline registration form, we will ask gamers how many hours they spend playing CS: GO during a typical week. In our revised ms, we have also clarified several additional criteria. We will invite adult, Polish-speaking male players of one of the most popular esports games: Counter-Strike: Global Offensive (CS: GO), who play at least six hours per week. We will recruit adult players because CS: GO is recommended only for +18 players (PEGI, 2023). 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 shooter gamers (Statista, 2023). Our latest study found that only 7% of Polish CS: GO players were women (Behnke et al., 2023).

We plan to recruit only active players (not veterans), as Stage 2 requires participants to play CS: GO as if they were playing during a regular week. We don't want to change how participants normally function, as would be the case for veterans starting to play. This is important to minimize the risk of boomerang effects or unintended counter-reactions to participating in the study. We clarified this issue in the description of Stage 2.

Behnke, M., Stefanczyk, M. M., Żurek, G., & Sorokowski, P. (2023). Esports Players Are Less Extroverted and Conscientious than Athletes. *Cyberpsychology, Behavior, and Social Networking*, *26*(1), 50-56. https://doi.org/10.1089/cyber.2022.0067

Changes in the manuscript: See Procedure - Stage 2 section, line numbers: 559 – 562, and Participants section, line numbers: 278 - 292.

#2. Related to the above, I am thinking whether it would be better to recruit participants based on rank or another performance indicator, rather than hours of training/play, considering performance is a hypothesis. It is possible that rank also partially explains affective experiences, so having min/max rank could help. Additionally, it feels important to control previous experience with the bot deathmatch used in the intervention; individuals who have already learned how to produce high scores in this mod could generate unwanted data (see another comment later).

Reply: This is an interesting proposition. We set the criterion of 6h/week based on our experience running previous psychophysiological studies on CS: GO and FIFA players. Recruiting 300 participants for a psychophysiological study is challenging, and it will be only harder to recruit participants from the limited gaming population. Thus, we don't want to set the threshold too high and we want to recruit experienced gamers. We believe that the threshold of 6 hours creates such an option. As we don't have data on players' ranks, we are afraid that placing a rank limit could result in a situation where we are unable to recruit a sufficient number of gamers. Thus, we decided to keep the original inclusion criteria.

In the recruitment, we will ask, " How many hours do you spend playing CS: GO against bots during the regular week?". We will use it in the exploratory analysis as the moderator.

Changes in the manuscript: See Measures section, line numbers: 776 – 777, and Participants section line numbers: 278 - 292.

#3. Still about participants: it is mentioned that individuals with significant health problems will be excluded. Does this refer to standard cut-offs with the applied scales? I'm specifically looking at gaming disorder, which you measure with a DSM-5 scale;

although I'm sceptical about the scale's clinical validity, it would be a concern to recruit (or continue having) participants who meet gaming disorder criteria at screening. Consider moving this scale to baseline registration measures as an exclusion screener.

Reply: We've never intended to exclude participants based on the cut-offs with the applied scales. In the study invitation, we state that only people without significant cardiovascular health problems or diagnosed mental disorders can participate. We clarified this issue.

Changes in the manuscript: See Exclusion section, line numbers: 415 – 420.

#4. Some existing Polish scale translations are cited. It would be good report whether you will use your own translations (when you do) or if English versions are used (if they are). The supplementary materials are informative, but it would help to have this information clearly in the manuscript.

Reply: We clarified this issue.

Changes in the manuscript: See Measures section, line numbers: 630-631, 643-644, 658-659, 671-672, 680-681, 697-698, 751, 760, 767-768, and Supplementary Materials, line numbers: 116, 123-124, 135, 156, 163, 169.

#5. Health scale from Ware Jr & Sherbourne (1992) is reported to ask physical health (supplement p. 4), but the item of the original scale measures general health. Is the modification "psychical" part of the Polish translation?

Reply: Thank you for pointing out this issue. We corrected the description of the item in Supplementary Materials.

Changes in the manuscript: See Supplementary Materials, line numbers 133-137.

#6. I don't want to further complicate scale selection (reviewers already address that in detail), but it also seems theoretically plausible for the synergetic mindset intervention to work specifically for individuals with low self-esteem (I believe this was part of Dweck's original reasoning). I wonder if e.g., Rosenberg's Self-Esteem Scale for exploratory or future analyses might be informative.

Reply: Keeping the balance between interesting constructs and participants' burden is the hardest part of planning the experimental study. As we prioritize minimizing participants' burden over our own interest and our plan to ask the participants to fill in the same scales in the 1-month follow-up, we aimed to limit the number of scales used in the study. Thus, we included a Single-Item Self-Esteem Scale ("I have high self-esteem"; Robins et al., 2001).

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring Global Self-Esteem: Construct Validation of a Single-Item Measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 27, 151-161.

Changes in the manuscript: See Measures section, line number: 757-762.

#7. On p. 13 it is noted that SESOI would be d = .07 among .03 and .05, but I don't see it explained why the former and not one of the latter. This has no pragmatic relevance, but you may wish to elaborate unless I've missed something.

Reply: We clarified this issue. In the revised version of the manuscript, we keep only the justification for the SESOI of d = .07. Based on the data from Behnke et al., (2020), we added the confidence intervals around the effect 95% CI [-.24, .38], Although the differences between rivals ranked one or two places were interesting, they were also confusing for readers. Note that despite the importance of specifying the smallest effect size of interest, this discussion has not been had among peers in this field. Our proposal here should be seen as a starting point for this discussion, not the final say on the matter.

Changes in the manuscript: See Sampling Plan – Expected Effect Sizes - Synergistic Mindsets & Performance section, line number: 305-313.

#8. Deathmatch AI is used (p. 21). For future work (no sense to make last-minute changes here), the Aim Botz mod could provide performance data in a more standardizable setting. Although deathmatch is more organic, scores are influenced by weapon choice, bot behavior, etc. That said, I wonder if it would be possible to further standardize deathmatch conditions, e.g., by fixing participants to one weapon.

Reply: We aimed to balance ecological validity and CS: GO performance standardization. In our opinion, the 2-minute deathmatches mimic the tournament mode where the matches are short games against the opponents. As weapon preference can be considered a strategic choice, we do not want to limit it. However, we turned off the first random weapon option to standardize the deathmatch.

Changes in the manuscript: See Procedure - Stage 1 section, line number: 550 – 552.

#9. Related to the above, it might be good to stress that the performance situation is human-AI and not human-human. This may be a highly relevant component in the production of competitive challenge/threat response. I don't know if the following study ever replicated but see, e.g., Kätsyri et al. 2013: https://doi.org/10.1093/cercor/bhs259 (perhaps to be considered more at Stage 2 discussion).

Reply: We incorporated the recommender suggestion and stressed that our performance is a human-computer interaction situation, and we will comment on it in the discussion.

Changes in the manuscript: See Procedure - Stage 1 section, line number: 551 – 552.

#10. Related to the above still, I'm also thinking to what degree gender affects this intervention. Especially in CSGO, competitive women players have always been a small minority, which has affected in how they experience competitive situations (e.g., Balakina et al. 2022: https://doi.org/10.1145/3569219.3569393). I believe the support for stereotype threat effects is currently weak at best, but I would consider e.g., including images of and quotes from top women players in the materials for women participants (i.e., have separate sets of materials for men and women). This could be a simple way to improve intervention effectiveness.

Reply: We apologize for not being clear about this issue, but we planned to recruit male CS: GO players - this is why all pictures in the intervention depict male gamers. Unfortunately, the female CS: GO stage in Poland is still sparse. Our latest study found that only 7% of CS: GO players were women (Behnke et al., 2023). As we observe changes in the field, we hope

our future studies will include female participants. We will address this issue as a limitation of our project.

Changes in the manuscript: See Participants section, line numbers: 278 - 292.

#11. I ran a face validity check with the materials through a Polish player, and one additional note (see also reviewer feedback) came up: the term "gamer" (graczy) addresses a specific subgroup of players, which has a strong identity connotation in this cultural context and tends to exclude some potential participants (in the same way as "scientists" would likely exclude, e.g., philosophers among researchers). It's totally ok if this is your target group, but just ensuring you're aware of that (as it would be very easy to use different terminology).

Reply: Interesting point. The Polish term "gamer" (gracz) relates to people playing computer games. We contacted Maria B. Garda, a researcher of Polish gaming culture and a semi-pro World of Tanks player, to discuss the topic. Through the discussion, we agreed that for the context of our study, the term 'gamer' (gracz) should be very well received by male CS: GO gamers.

#12. Regarding the baseline measures on p. 29 and supplement p. 6, I note the following. #5: considering using "playing" instead of "training", as people interpret "training" in many ways, e.g., ranked play isn't training? (especially if this item is used as an inclusion criterion) #6: Some people have multiple accounts and have played other mods of (almost identical) CS, so maybe provide an option to self-estimate total hours played CS?

Reply: We incorporated the recommender suggestion.

Changes in the manuscript: See Measures section, line number: 684, and Supplementary Materials, Items #5 & #6 in Baseline Registration, line number: 182-183.

#13. P. 36 data availability says that all data will be made available, but I assume, e.g., all video data will not be made available? Any other exceptions?

Reply: We are planning to share the video data as well for participants who consent. Participants will decide what kind of data they want to share. They can decide to make their video recordings, physiological data, self-reports, or/and behavioral available.

We understand why the recommender raises the issue of sharing the video recordings - it might be problematic from the perspective of data anonymization. However, existing technology allows identifying individuals from physiological signals (e.g., ECG) like videos or fingerprints (Hernandez et al., 2015).

Hernandez, J., McDuff, D. J., & Picard, R. W. (2015, June). BioInsights: Extracting personal data from "Still" wearable motion sensors. In 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (pp. 1-6). IEEE.

Changes in the manuscript: See Procedure - Stage 1 section, line number: 445 – 447.

I was informed during the process that your reservation for the lab is now in April. Let's make sure you have a decision before that. I know it's a lot of feedback. I can be contacted directly at any point for all concerns and questions, and if possible, please

inform me some days before when you're about to resubmit so that I can prearrange a time to fully prioritize this and provide a rapid turnaround. This is an important study, and it's a privilege to help you with it.

Reply: Thank you for the kind words. We appreciate your help with the smooth revision process.

Lee Moore (Reviewer 1)

This registered report excellently describes a very interesting and impressive study, and meets most of the Stage 1 criteria to a high degree (i.e., logic, rationale, and plausibility of proposed hypotheses; soundness and feasibility of the methodology and analysis pipeline; and consideration of sufficient outcome-neutral conditions). That said, as I outline in my more specific feedback below, I believe some improvements are possible in relation to a couple of the criteria (i.e., scientific validity of the research question; clarity and degree of methodological detail). For instance, the registered report would benefit from a clearer and stronger theoretical underpinning (e.g., biopsychosocial model of challenge and threat), greater alignment between underlying theory and methods (e.g., choice of self-report measures of challenge and threat), and a more vivid description of, and stronger justification for, some methodological elements (e.g., cardiovascular data recording, self-report measures). Overall, I enjoyed reading this registered report and hope the authors find my feedback useful.

Reply: Thank you for your kind words!

Major feedback
#1. The introduction, and study outlined in the registered report more broadly, would benefit from a clearer and stronger theoretical underpinning. Indeed, at present, the authors briefly describe aspects of numerous theories (e.g., work of Blascovich, Lazarus, etc.), and the introduction would benefit from a clearer overview of the framework(s) most strongly informing the study (i.e., biopsychosocial model of challenge and threat).

Reply: We incorporated reviewers' suggestions following the specific feedback provided in Comments #7- #11.

#2. Methodologically, the measure of situational appraisals currently described does not align well with the conceptualisation of challenge and threat appraisals included in the introduction (i.e., balance of evaluated situational demands vs. personal coping resources). Greater coherence is therefore needed and so I would recommend the authors use an alternative measure (e.g., items adapted from the cognitive appraisal ratio combined into a demand resource evaluation score). Indeed, there have been calls in the challenge and threat literature for more homogenous methods to be used across studies (see Hase et al., 2019 for a review). Furthermore, linked to this point, I would encourage the authors to use timescales with their cardiovascular reactivity data that better align with previous research (e.g., final minute of resting baseline and last minute of pre-match period).

Reply: We incorporated reviewers' suggestions and added the challenge/threat ratio measure. For instance, one item will assess task demands ("How demanding do you expect the CS: GO match to be?"), and another item will assess personal resources ("How able are you to cope

with the demands of the CS: GO match?") (Moore et al., 2012, 2013, 2014). The scale will range from 1 (*not at all*) to 7 (*extremely*). These will be combined to gauge the ratio appraisals.

Furthermore, within the challenge and threat literature, the majority of studies use the final minute of baseline because this is the most relaxed period and avoids potential contamination by people reacting to having sensors placed on their bodies (Blascovich et al., 2004; Chalabaev et al., 2009; Jamieson et al., 2013; Mendes et al., 2007; Moore et al., 2012, 2013, 2014, 2015; Oveis et al., 2020; Peters et al., 2018; Scholl et al., 2017; Seery et al., 2010; Turner et al., 2012, 2014; Vine et al., 2013) vs. two or more minutes (Behnke et al., 2020, Behnke, Gross, et al., 2022; Hagen et al., 2016; Yeager et al., 2016, 2022). Thus, as suggested by the reviewer, we will use the final minute of resting baseline and the last minute of the pre-match period.

Changes in the manuscript: See Measures section, line number: 673 – 681, and Data Preprocessing section, line number: 795 – 799.

#3. Similarly, if the authors wish to assess an individual's propensity to appraise all potentially stressful situations as more of a challenge or threat, I recommend that they use the Appraisal of Challenge or Threat Scale developed by Tomaka et al. (2018), as it better aligns with how they are currently conceptualising challenge and threat in the introduction.

Reply: As we wrote in response to Recommender Comment #6, we prioritize minimizing participants' burden over our interest. Adding the Appraisal of Challenge or Threat Scale would result in additional 48 items (24 items, each asking two questions "How demanding is this event to you" and "How able are you to take action to deal with it?"). Moreover, focusing on contextually-variation is favoured given our focus on experimentally manipulating appraisal processes via the synergistic mindsets intervention (see Yeager et al., 2022). Thus, the individual differences in the propensity to appraise all potentially stressful situations as more of a challenge or threat is not the main focus of our study, so we decided not to include the Appraisal of Challenge or Threat Scale.

Furthermore, we choose to keep the Negative Appraisals measure to maintain consistency between our study and the initial synergistic mindset intervention study (Yeager et al., 2022).

Minor feedback

#4. The title is punchy, but could better reflect the study outlined in the registered report. For example, a key outcome, challenge and threat (or stress appraisals), is missing.

Reply: We added "Esports" to the title following reviewer 2 suggestion. We would like to keep it this way. We partly incorporated the reviewer's suggestion by adding the suggested phrases (e.g., stress appraisals) into keywords.

Changes in the manuscript: See the Title and Keywords, line numbers: 5, 46.

The abstract offers a neat summary of the study described in the registered report, including background, methods, and hypotheses. Areas for improvement include:

#5. Greater conceptual clarity via more accurate terminology. For instance, challenge and threat are appraisals of motivated performance situations that can be inferred via cardiovascular responses and are theorised to have downstream effects on affective responses (e.g., emotions), performance outcomes, etc. Thus, referring to 'challenge versus threat affective responses' is not conceptually accurate. This is a recurring issue that could be resolved throughout the registered report.

Reply: We respectfully disagree. We conceptualize affect and affect regulation using the Integrative Affect-Regulation Framework (Gross, 2015; Troy et al., 2022).

> *Affect: any response to an internal or external stimulus involving a valuation; stress and emotion can both be considered subsets of affect.*
>
> *Affect regulation: strategies used to alter affect, including attempts to change subjective experience, cognition, behavior, physiology, or the environment; coping and emotion regulation are subsets of affect regulation.*

We believe challenge and threat appraisals of motivated performance situations are affective responses. Although stress responses and emotions are often viewed as separate phenomena, they both involve appraisals and whole-body reactions to psychologically relevant situations (Blascovich, 2008; Epel et al., 2018; Gross, 2015; Lazarus, 1993; Troy et al., 2022). We plan to elaborate on this issue in the Discussion section.

#6. More methodological information could be provided. For example, it might be clearer precisely when measurements of challenge and threat will be taken.

Reply: We incorporated the reviewer's suggestion. The current version of the abstract without the results is 246 words long. Thus, although we agree that additional methodological information could be provided, we want to save space for the results and conclusions.

Changes in the manuscript: See Abstract, line number: 41 – 42.

The introduction excellently covers relevant and recent literature (e.g., Yeager et al., 2022) to 'set-up' the study described in the registered report. Areas for improvement include:

#7. Greater flow between paragraphs and subsections (e.g., opening paragraph into the subsection entitled 'how appraisals influence performance').

Reply: We clarified a few issues in the introduction, hopefully improving the flow between paragraphs and subsections

Changes in the manuscript: See Introduction section, line number: 49-57, 62, 66-68, 72-81, 87-94, 101-102, 104, 123-124, 177-179, 191-201, 213-230.

#8. More theoretical content. For instance, key components of pertinent theory are missing, such as clear definitions of concepts such as 'motivated performance situations', 'task engagement', 'cardiac output', and 'total peripheral resistance' which are central to the biopsychosocial model of challenge and threat.

Reply: We added the definitions in the revised version of the introduction.

Changes in the manuscript: See Introduction section, line number: 51-53, 87-94, 104.

#9. Extra criticality. For example, while challenge and threat appraisals are thought to lead to different emotional responses (see the predictions of the theory of challenge and threat states in athletes by Jones et al.), the empirical evidence supporting this assertion, at least in sport psychology literature, has been relatively mixed.

Reply: We are aware of mixed support from sport psychology literature for the relationship between emotions and challenge and threat appraisals (e.g., Nicholls et al., 2012; Turner et al.., 2012, 2013). However, we are operating off a direct interpretation of the BPS model of challenge and threat that conceptualize challenge and threat as affective states themselves. Certainly, specific emotional response tendencies can elicit physiological responses consistent with challenge and threat, but distinct emotional experiences can elicit similar responses. For instance, anger – a negative approach state – can lead to changes in CV functioning that can appear similar to other approach states, such as excitement.

However, we expect that by targeting mindsets - constellations of beliefs, which are temporally stable appraisals – we will be able to influence the gamers' affective responses before and during the performance. As previous challenge/threat studies in sports were based on 1-2 paragraph interventions, we consider the scope of the intensive synergistic mindset intervention to be a crucial step forward that will help us examine the associations between performance-related appraisals and affective responses. We will discuss this issue in Discussion section

#10. Figure 1 presents challenge and threat appraisal in a differ way conceptually to the text (i.e., harm and control vs. situational demands and personal coping resources), and so greater conceptual clarity and alignment is needed. Indeed, as noted above, a clearer overview of underpinning theory is needed (e.g., Blascovich's BPSM).

Reply: We believe that Figure 1 accurately presents our way of thinking about affective responses. We clarified the introduction to align it with Figure 1. We know we use a slightly different theoretical framework than classic Blascovich's BPSM and do not necessarily follow the usual challenge/threat conceptualizations. In this project, we aim to integrate different but mutually complementary frameworks that we hope bring a novel perspective on studying affective response in the performance situation.

Changes in the manuscript: See Introduction section, line number: 62, 66. 79-80, 101-102.

#11. Given the inclusion of stress mindsets in the study described in the registered report, I would suggest briefly summarising this literature somewhere in the introduction.

Reply: We appreciate this suggestion, but we decided not to introduce the summary of stress mindset research in the introduction. Instead, we will include this issue in the discussion section.

First, we want to keep the introduction concise. In our view, the introduction should allow the reader to understand what will be done in the study, including only necessary information, without additional interjections that might confuse and slow down the reader.

Our introduction section is already ten pages and 2500 words long, which exceeds some journals' recommendations (e.g., *Nature Human Behaviour*).

Second, the introduction aimed to introduce three important concepts, namely 1) Integration of challenge/threat literature with emotion and emotion regulation literature; 2) (Re)appraisal interventions focused on the situations and responses to the situations; 3) Synergistic mindset intervention - the move from focal reappraisal interventions to bread mindset-oriented intervention. We know it is already challenging to digest all concepts together. We do not want to complicate it by introducing two additional concepts (e.g., growth and stress-can-be-enhancing mindsets).

Third, we want to present the Synergic Mindsets approach as a coherent and complementary whole. We do not want to split the growth and stress-can-be-enhancing mindsets and present them separately. The initial study shows that the synergistic approach is the most beneficial (Yeager et al., 2022). Thus, we believe that having both growth and stress-can-be-enhancing mindsets plays a crucial role in optimizing performance, and this is how we want to present it in the introduction.

Fourth, the separate mindset research summary is presented in the Procedure - Synergistic Mindset Intervention section. We believe it is a proper place to describe the foundations of synergistic mindsets.

Changes in the manuscript: See Procedure - Synergistic Mindsets Intervention section, line number: 513-515.

The method provides an excellent and relatively detailed summary of how the authors plan to collect and analyse their data. Areas for improvement include:

#12. It is unclear why physiological markers of challenge and threat are included in the primary analyses, but self-report measures are included in exploratory (or secondary) analyses, particularly given the former are thought to objectively reflect the latter. I would encourage the authors to reflect on this and consider if self-report measures of situational appraisals should be included in the primary analyses. If so, the effect sizes reported in previous research linking challenge and threat, measured via self-report, with performance could be useful in informing the sampling plan (see Hase et al., 2019 for a review).

Reply: We incorporated the reviewers' suggestions and added the challenge/threat items before each match to better align with the extant challenge/threat research. However, we approach the self-reports as similar to manipulation checks – that is, did the synergistic mindsets intervention that aimed to change appraisals actually change the appraisals? In that way, we use the typical approach in affective science, where after the affect manipulation (e.g., presentation of an amusing video), we check whether it elicited the targeted affective outcome (e.g., amusement experience). Moreover, given limits in human interoceptive accuracy and myriad reporting biases, we aim to focus the research more strongly on measuring objective physiological indicators of challenge and threat.

Changes in the manuscript: See Manipulation Check section, line number: 856 – 859.

#13. More information is required in terms of how randomisation will be achieved. In addition, the authors should consider other criteria that are commonly used to assess the methodological quality of randomised controlled studies and what they might do to ensure their proposed study satisfies these criteria. For example, will researchers

assessing outcomes (i.e., challenge and threat) be blind to group allocation? How will missing data be kept to a minimum? Etc.

Reply: We included the participants' randomization scheme in the Supplementary Materials. We added the information that the researchers conducting preprocessing, scoring, and analysis of all signals will be blind to the condition. This is achieved by not linking participant IDs with condition assignments in raw physiological data files. We aim to keep missing data to a minimum by following standards in psychophysiological research and running the study with trained and experienced experimenters. For instance, participants will remain seated in a single location throughout gameplay. This approach helps limit artifacts associated with movement in impedance cardiography and electrocardiography signals.

Changes in the manuscript: See Procedure section, line number: 462-464, and Table S2, and Analysis Plan section, line number: 783-784.

#14. The authors might want to consider assessing, and/or controlling for, interoceptive ability or awareness as this might impact the effectiveness of some elements of the synergistic mindset intervention (e.g., components based on arousal reappraisal).

Reply: Thank you for this suggestion. We decided to include the Body Awareness Questionnaire (BAQ; Shields et al., 1989) in the questionnaire set before Stage 1.

Shields, S.A., Mallory, M.E., & Simon, A. (1989). The Body Awareness Questionnaire: Reliability and validity. Journal of Personality Assessment, 53, 802-815.

Changes in the manuscript: See Measures section, line number: 763 – 771.

#15. More information is needed on how the authors will ensure and assess task engagement in the laboratory-based performance tasks. Additionally, how will it be ensured the tasks represent personally relevant motivated performance situations? Linked to this, I would strongly encourage the authors to assess HR as well as PEP as a marker of task engagement. Indeed, they could follow the work of Seery et al. and combine HR and PEP into a single index to simplify analyses, etc.

Reply: We believe that situations where experienced CS: GO gamers decide to participate in the tournament and compete for a prize of approximately 600$ represent personally relevant motivated performance situations.

Incorporating the reviewer's suggestion, we will use HR as the secondary marker of task engagement. However, we favor PEP as a "purer" index of sympathetic engagement, because it directly assesses the effects of the innervation of sympathetic ganglion fibers on the cardiac pacemaker – more sympathetic activation = strong contractions = shorter PEP interval. In contrast, HR can be influenced by parasympathetic processes in addition to sympathetic impacts (e.g., vagal break processes). We also favor PEP over a combined index. PEP is more parsimonious (i.e., simpler) because it is one measure instead of an index. Thus, we will not combine the HR and PEP into an index, as we believe it introduces additional noise rather than simplifying analysis.

Changes in the manuscript: See Measures section, line number: 721 – 722, and Physiological data reduction section, line number: 814 – 815.

#16. It could be made clearer to the reader precisely where the intervention content and/or materials will be stored to enable replication (e.g., OSF)?

Reply: Our intervention content and materials will be presented in supplementary materials. On p. 41, we provided a link to the supplementary materials and the Open Science Framework (OSF) website, where all data will be openly available.

Changes in the manuscript: See line numbers: 947 – 955.

#17. It seems like two-minute baseline and recovery periods are to be used around esports matches. Is this sufficient to enable cardiovascular markers to return to baseline? You commonly see 5-minute periods of recording used in prior research.

Reply: In our previous study (Behnke, Gross, et al., 2022), we observed that gamers were able to recover during a two-minute period as they displayed similar physiological levels during the pre-study baseline and 2 minutes after the matches.

	Pre-study baseline		After the match	
	M	*SD*	*M*	*SD*
HR	76.51	11.46	75.05	10.36
PEP	113.55	14.26	116.04	15.40

Furthermore, studies on physiological recovery from negative emotions and stress that measured the time needed to recover to baseline levels also showed that two minutes of recovery should be a sufficient recovery period.

- Fredrickson & Levenson, 1998 (time to recover = 40 s, *SE* = 10)
- Fredrickson et al., 2000 (35 s, SE = 9)
- Gilbert et al., 2015 (15 s, SD = 16s)
- Hannesdóttir, 2007 (28s, SD = 33)

Finally, as the ecological validity of this study is important to us, we aimed to create a procedure that will allow for dynamic multi-round tournament competition without distracting or boring players with overly long pauses (in the feedback of our previous studies, gamers complained about the 2-minutes baseline). We decided to shorten the baseline and recovery typical 5-minutes intervals while maintaining appropriate standards.

Fredrickson, B. L., & Levenson, R. W. (1998). Positive emotions speed recovery from the cardiovascular sequelae of negative emotions. *Cognition & Emotion, 12(2),* 191–220. https://doi.org/10.1080/026999398379718.

Fredrickson, B. L., Mancuso, R. A., Branigan, C., & Tugade, M. M. (2000). The undoing effect of positive emotions. *Motivation and Emotion, 24(4),* 237–258. https://doi.org/10.1023/A:1010796329158

Gilbert, K. E., Gruber, J., & Nolen-Hoeksema, S. N. (2016). I don't want to come back down: Undoing versus maintaining of reward recovery in older adolescents. Emotion, 16(2), 214–225. https://doi.org/10.1037/emo0000128.

Hannesdóttir, D. K. (2007). Reduction of fear arousal in young adults with speech anxiety through elicitation of positive emotions (Doctoral dissertation). Retrieved from, https://vtechworks.lib.vt.edu/bitsteam/handle/10919/28941/dissertation.pdf?sequences=2&isAllowed=y.

#18. How precisely will the synergistic mindsets group report adherence and progress with the intervention? More details are needed to enable replication.

Reply: As presented on p. 25, the instructions will be: "List some of the gaming situations that elicited strong emotions or stress and the way you used Rethinking to make the situation beneficial to you". Furthermore, this information is also presented in Supplementary materials - Daily measures – Stage 2 - Daily Survey - #10 [SMI Group] (line numbers 1918-1920). We hope that our detailed documentation in two languages will enable replications.

#19. From Figure 2, it seems that the emotion recall task used in stage 3 (i.e., postintervention) is not being used in stage 1 (i.e., pre-intervention). Why? I would have thought pre- to post-intervention changes in appraisals, emotions, etc., would be vital in evaluating the effectiveness of the synergistic mindset intervention.

Reply: We treat the emotion recall task as an exploratory addition to the main project. It is only included at the end of Stage 3 because we are interested in studying emotions that will occur during the tournament – high-stakes performance. In Stage 1, gamers will be introduced to laboratory procedures and the synergistic mindset intervention. They will play training matches, but we do not expect them to have the same level of engagement as during the tournament. Thus, pre- to post-intervention comparisons would concern contextually different situations and emotions.

#20. Scoring information (e.g., sum totals or mean values), as well as more details relating to the validity and reliability of each self-report measure, is needed. Also, for some measures (e.g., situational affect regulation), different scales will be used pre- and post-intervention. This seems a little unusual and it is currently unclear why. Indeed, stronger justifications are needed relating to the self-report measures. For example, why have the authors decided to use the stress mindset measure over other relevant questionnaires (e.g., the instrument developed by Keech et al.)?

Reply: We clarified these issues. For the primary analyses, we will use latent variables for multi-item scales. For the manipulation checks, we will use Principal Component Analysis component scores for multi-item scales (as suggested by Reviewer 2). We added the details related to the validity and reliability of each self-report measure. We unified the scale format for situational affect regulation. Finally, as our work is heavily inspired by the initial work on synergistic mindsets interventions (Yeager et al., 2022), we will use scales used in the initial project. In that way, we will be able to compare the results from both studies.

Changes in the manuscript: See Measures section, line number: 631-633, 638, 643-646, 658-661, 672, 678-681, 698-700, 751-756, 760-762, 768-771, and Manipulation Checks section, line number: 832 – 833.

#21. After identifying outliers, how will they be dealt with (e.g., excluded, winsorized, etc.)? More information is needed to enable replication.

Reply: We clarified this issue. We will identify outliers with the median absolute deviation (MAD), with a cutoff of 3, as recommended by Leys et al. (2013, 2019). We will then delete the data if the data is identified as an error. We do not expect any measurement errors and encoding errors in affective experience data and gaming data, given the self-report

measurements we will collect. Even if a data point is identified as an outlier, we will not delete it if it represents real data rather than an error, as even an extreme score most likely correctly reflects the answer a participant wanted to give. For the cardiovascular data, we will double-check the identified outliers. If we find biologically impossible values, we will delete them. We will report the number of outliers for a given variable. Finally, we will use the Mplus default estimation option (i.e., the full-information maximum likelihood) to impute the missing values.

Changes in the manuscript: See Data Preprocessing section, line number: 817 – 827.

Ivan Ropovik (Reviewer 2)

Thanks to the authors for the opportunity to read their manuscript (ms). Overall, I think that the proposed study would be informative. One of its main selling points is that it will produce a rich dataset, making it possible to examine the effect of the designed intervention on the longitudinal trend in performance and affective measures while not relying entirely on self-reports but also collecting physiological measurements. Thanks to being a RR, the present study has the potential to bring evidence that would likely be much more robust than a modal study published in this field. That said, I have also some critical takes and suggestions for improvement.

An acknowledgment upfront, I am not a social psychologist and don't have much expert knowledge about the substantive aspects targeted by the present study. In my review, I will mainly focus on the measurement, design, and analysis side of things. As my role as a reviewer is mainly to provide critical feedback, I provide it in the form of comments below, not in order by importance but rather chronologically as I read the paper. I leave it to the authors' discretion which suggestions they find sensible and choose to incorporate. I hope that the authors find at least some of the suggestions below helpful.

Reply: Thank you.

#1. In the introduction, I am missing a bit more critical interpretational viewpoint. As usual, in most research studies, the presented past research is all taken at face value. Especially in such literature, where weakly informative designs yielding very heterogeneous findings are rather the norm, I think it makes sense to identify which of the past studies presented in the intro are vitally important for informing the theoretical underpinnings of the present study and qualify the strength of their conclusions by the methodological robustness of the design they utilized.

Reply: 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.

Changes in the manuscript: See Introduction section, line number: 213 – 230.

#2. The data will not support very wide generalization, so I would suggest revising the title to sth more specific, like Optimizing *Esports* Performance Using a Synergistic Mindsets Intervention. The same with abstract.

Reply: We incorporated the reviewer's suggestion and added the "Esports" in the title. We believe that the abstract already emphasized the esports context.

Changes in the manuscript: See Title, line number: 5.

Optimizing Esports Performance Using a Synergistic Mindsets Intervention

#3. My hunch is that affective response patterns may be rather stable characteristics that will be difficult to structurally alter with a self-administered one-shot type of intervention. Getting *a* significant effect is not that hard. Finding *the* effect using a rigorous method (even though the intervention seems face valid) is far less likely, IMO. The good thing about the design of the present study is that it attempts to partly stretch the intervention over a week. Kudos to the authors that they choose a RR format to give it a try.

Reply: Thank you for your kind words. Previous affect regulation studies (challenge/threat and emotion regulation literature) usually used shorter interventions (1-2 paragraphs interventions) and were able to find significant effects. In this project, we are taking a step forward as 1) we will use a more comprehensive intervention, and 2) participants will be able to train the learned skills. Thus, we expect that our design will allow the creation (and detection) of "the" effect. Any study can, of course, be further enhanced, but we believe our study is the most advanced project examining affective responses in the motivated performance context.

#4. "The other participants will be assigned to a validated placebo intervention focused on learning about the brain (Yeager et al., 2022)." Well, this appears to be a stretch too far for me. The cited study did not carry out *any* validation of this control condition. Being used before (even though in a Nature paper) is not equal to being validated. I'd suggest removing that remark. I'll have more comments on the control condition below.

Reply: We clarified this issue. As validated intervention refers to an intervention developed to be administered among the intended respondents – our adaptation was not previously tested on gamers - we changed the phrasing to "previously tested placebo intervention."

However, we would like to note that we are using a control intervention used in six studies (total N = 4,291; Yeager et al., 2022). The content of the control condition used in the Nature paper was predominantly from the control condition from a prior national growth mindset experiment in the USA (Yeager et al. 2019, also Nature paper; N = 12 490), which creation was described and tested in another manuscript (Yeager et al., 2016; N = 3 676). Thus, we have strong confidence in using the control intervention presented in the manuscript. We commented more on this issue in response to reviewers' comment #33.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... & Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364-369.

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., ... & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of educational psychology*, *108*(3), 374.

Changes in the manuscript: See Present Study section, line number: 248, and Procedure – Stage 1 section, line number: 534-540.

#5. Again, rather loose use of a validity claim. "In sum, our study will provide a unique combination of high internal and external validity levels". I think get what you mean (using a controlled experiment & real-world outcome?), but I am not sure about the "high" part (more on that later – I see issues with the measurement and comparativeness of the controls). Anyway, instead of this slightly hyped-up language, I'd stick to being more descriptive.

Reply: In the revised manuscript, we toned down the language.

Changes in the manuscript: See Present Study section, line number: 273-274.

#6. The targeted population could have been described in more detail. It does not even mention that the sample will be entirely Polish (true?). Fully ok, but these details need to be acknowledged. How are the participants going to be sampled/recruited? Predominantly, what kind of people will they be? Students? General population? Is >6 hours of playing time the only inclusion criterion? Is there an expectation regarding the sample composition?

Reply: As noted in response to recommender #1 comment, we clarified the targeted population. We will invite adult, Polish-speaking male players of one of the most popular esports games: Counter-Strike: Global Offensive (CS: GO), who play at least six hours per week. We will recruit adult players because CS: GO is recommended only for +18 players (PEGI, 2023). 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 shooter gamers (Statista, 2023). Our latest study found that only 7% of Polish CS: GO players were women (Behnke et al., 2023).

We will recruit gamers from the general population via a Facebook advertisement targeted at CS: GO gamers and mailing lists among university students in Poznan. We have no expectations regarding the sample composition.

Behnke, M., Stefanczyk, M. M., Żurek, G., & Sorokowski, P. (2023). Esports Players Are Less Extroverted and Conscientious than Athletes. *Cyberpsychology, Behavior, and Social Networking*, *26*(1), 50-56. https://doi.org/10.1089/cyber.2022.0067

Changes in the manuscript: See Participants section, line numbers: 278 - 292.

#7. Part Sampling Plan/Expected Effect Sizes, the first paragraph could be structured more clearly. It currently mixes substantive assertions with generic stats meta-talk.

Reply: We clarified this paragraph.

Changes in the manuscript: See Sampling Plan section, line number: 295 – 297.

#8. I think that the Expected Effect Sizes part is a conceptually relatively weakly informative part of the research design justification. I understand that the authors wanted to provide a solid ground for the sample size determination but I think it misses the point. To elaborate, I see the following. I like the determination of SESOI that is grounded in some reality. That part is fine. But I don't get the need for "expected ES".

Trying to inform the design of a study based on (probably) non-systematic picking among relatively idiosyncratic, heterogeneous effects from published (thus subject to publication bias) literature is unfruitful, IMO. The meta-analysis by Webb et al. (2012) is, IMO, not helpful too, for that matter (more on that later). Even if there was no bias in the literature, considering "expected ES" is incompatible with the frequentist notion of power, a *pre-data*, theoretical concept (just like α), i.e., a sensitivity of a given statistical test to reliably detect a range of hypothetical population effects of interest (see Morey & Lakens, 2017). Instead of such long, numbers- and references-ladden part, I, as the reader, would prefer to see the power curve (given the specific model) for the hypothetical range of effect sizes. That range would, of course, also include the SESOI. A figure would say more than thousand words, not forcing the reader to appraise the informativeness of the present design at fixed points.

Reply: We added the power curve for our model and clarified this issue in the manuscript. Furthermore, we decided to keep both approaches (SESOI and Expected ES) in the manuscript. The PCI RR community might be up to date with the novel approaches for sample estimation (e.g., SESOI), but we expect that many readers of our manuscript will be familiar only with more traditional approaches (e.g., a sample size justification based on expected effect sizes). And even though it is better to plan a sample size based on the smallest effect size of interest (Lakens, 2022), expected effect sizes based on the scientific literature are often also effects of interest to scholars in these fields (Lakens, 2022). Of course, it is true, literatures can be biased, but this possibility was taken into account in our description. Anyone is free to completely ignore the existing literature as biased and irrelevant - but we believe our target audience will be interested in how our study design related to effect sizes that can be expected based on meta-analyses. Thus, we aimed to present a complementary approach to sampling estimation based on SESOI and expected ES. We will address problems with expected effect sizes in biased literature in the limitation section.

Changes in the manuscript: See Sampling Plan section, line number: 295-297, 306-308, 311, 314-333, 351-353.

#9. The unfruitfulness of the SESOI & Expected ES combo can IMO be seen in the Sample Size Determination part. Absent any formal mechanism (or common conceptual footing) to reconcilliate the two, the authors are pushed to conclude that the SESOI is unfeasible, while the "surprisingly large" ES from the past literature also did not pass some of the internal checks present in a skeptical reasoner. So the outcome of the several paragraphs long justification is an arbitrary set of ESs. There's inherently nothing wrong with setting an arbitrary target, or one that is doable given some budget. My point is only that looking at a wider hypothetical range would be more informative. That way, the reader would gain a comprehensive outlook of what power does the given design/test provide for any given ES. Btw, I liked the justification for the target ratio of type I/II error rates.

Reply: We hope that the presented power curve will provide a more comprehensive outlook of statistical power for our design. We agree that the discussion might be long, but 1) explicitly discussing the SESOI and feasibility constraints is important, especially in literature where such a discussion is uncommon, and 2) every study is informative for some effects of interest, and not others, and we believe clearly specifying which effects of interest the study will yield information about is worthwhile.

Changes in the manuscript: See Sample Size Determination section, line number: 392 -393, and Supplementary Materials – Figure S1, Table S1, line numbers: 89-92.

#10. It is fine to compute power for individual SEM parameter estimates (not "for the structural equation model" as put down), but I think it always makes sense to report whether the SEModel has decent power to pick up significant model-data discrepancies if these are present. That can be done for an approximate fit hypothesis using the RMSEA (see https://www.quantpsy.org/rmsea/rmsea.htm) and be reported at least in SMs.

Reply: We calculated the power for the model fit using the calculator suggested by the reviewer. We found in a sensitivity power analysis that our model (alpha = .05; df = 37; sample size = 2000; Null RMSEA = .01; Alt. RMSEA = .05) will have a power of 1.00 to detect RMSEA of .05. We included the code for the power analysis in the supplementary materials.

Changes in the manuscript: See Sample Size Determination section, line number: 400 – 403, and Supplementary Materials, line numbers: 4441-4572.

#11. Re: assuming factor loadings of .50… This is a serious design blunder, IMO. If the employed scales have such an abysmal overrepresentation of error variance (loading of .50 implies 75% of the total variance being error), it has serious consequences for the efficiency, precision, and likely also the accuracy of the design. Yes, in most of the social science research, measurement properties of the measures are hidden away behind convenient sum scores, so I don't want to scold the authors for paying higher-than-usual attention to some of their auxiliaries. But still, if it is the case, this should be discussed.

Reply: We will address this issue as a limitation of our study. As the reviewer noted, it is an acceptable limitation in social science research, and there are no better measures available. We will report the factor loadings in our manuscript and stress the need for measurement development in the section on future research in the discussion.

#12. Just an idea always worth considering, IMO. Maybe it would make sense to try to screen out careless responders (e.g., based on longstring detection or some sort of insufficient variance in responding pattern, or being a multivariate outlier indicating random response pattern). If done, it should only be applied to pre-treatment measures, as carelessness itself may have been affected by the treatment, and exclusions based on that would induce bias.

Reply: 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."

Changes in the manuscript: See Measures section, line numbers: 778-781, and Supplementary Materials - In Lab Baseline – Stage 1 & Before Tournament Stage 3 and 1-month follow-up section, line number: 503-504.

#13. For the description of stages, I found it hard to understand at times, what follows what. E.g., "participants will provide informed consent and fill out baseline

questionnaires"... "Next, the researcher will apply sensors to obtain cardiovascular measurements, and participants will fill in the baseline questionnaires". What comes first? Cardiovascular measurements or questionnaires? The figure with the procedure workflow is clear, but the text description was sometimes difficult to follow. Maybe it's because I'm not acquainted with the subject matter, but it was difficult for me to keep track of what was measured, when, and for what purpose.

Reply: We clarified this issue. We agree that the first part of Stage 1 was not clearly explained.

Changes in the manuscript: See Procedure section, line number: 445-458.

#14. In general, there is little information on how the measures will be ordered. Within the questionnaire block, why not randomize to minimize the order effects? The same with items within scales. Will their order be randomized?

Reply: To minimize order effects, the order of questionnaires will be randomized in the questionnaire sets (beginning of Stage 1, Stage 3, and one-month follow-up). However, our software (Microsoft Forms) does not allow us to randomize the order of the items within the scales. We will use the items' order within the scales as they were created and published. This is the usual approach in distributing psychological questionnaires. 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). We hope it will not be a disqualifying factor for our study and have no strong prior beliefs this factor will have any influence on the scores. Furthermore, one could probably argue that items in a measurement tool do not need to be randomized, but placed in the optimal order to create the best possible measure. Such an investigation is beyond the scope of our manuscript.

Weinberg, M. K., Seton, C., & Cameron, N. (2018). The measurement of subjective wellbeing: Item-order effects in the personal wellbeing index—adult. *Journal of Happiness Studies*, *19*, 315-332.

Schell, K. L., & Oswald, F. L. (2013). Item grouping and item randomization in personality measurement. *Personality and Individual Differences*, *55*(3), 317-321.

Changes in the manuscript: See Procedure section, line number: 451-452.

#15. In the stage, I, is gaming for 2 minutes enough to provide a reliable picture? Just asking.

Reply: Two minutes should provide a reliable measure of performance. We planned 2-minute matches to keep high ecological validity, as it mimics the length of the rounds of CS: GO tournament matches. In fact, it is common in the psychophysiology literature using well-validated Trier Social Stress Test or performance-based paradigms to focus only on the initial minutes of stressful tasks because these are the most "reactive" periods (e.g., Hangen et al., 2019; Jamieson et al., 2012; Oveis et al., 2020).

#16. It is fine to have confirmatory RQs, with all other things lumped in exploratory analyses. But the reasoning on p.24 does not make sense to me. This one: "We treat them as secondary because we did not include them in the power analysis, and we may not have enough statistical power to infer about the effects of synergistic mindset intervention on them.". Why? Power seems to be an independent issue to me.

Reply: In our reasoning, we followed the Journal Article Reporting Standards of the APA, which suggest researchers distinguish primary hypotheses and secondary hypotheses. The difference is as follows:

First, a study is designed to answer a **primary hypothesis**. The Type 1 and Type 2 error rates for this primary hypothesis are as low as the researcher can afford to make them. **Secondary hypotheses** are questions that a researcher considers interesting when planning the study, but that are not the main goal of the study. Secondary hypotheses might concern additional variables that are collected, or even sub-group analyses that are deemed interesting from the outset. For these hypotheses, the Type 1 error rate is still controlled at a level the researchers consider justifiable. However, the Type 2 error rate is not controlled for secondary analyses.

Of course, an alternative is to treat all secondary hypotheses as primary hypotheses as well, and make sure all have sufficient power. However, this is not practically feasible in the current study. Therefore, we will not draw any strong conclusions if an effect is absent, but as we control our error rate, we would like to draw conclusions if we can reject the null hypothesis. This is the essence of a secondary hypothesis.

#17. The number of items for RESS-EMA scales is not reported. Also, why a different response scale in the baseline and stage III?

Reply: We clarified this issue. We will use six items of the RESS-EMA, one for each affect regulation strategy. We unified the scales format for the RESS-EMA.

Changes in the manuscript: See Measures section, line number: 638.

#18. Several scales are based on only 2 items. I guess also the alpha will be very low.

Reply: We agree that the reliability of short scales might be low - or at least will be difficult to measure accurately, even if the reliability is high. However, as they are not primary measures and we are using them for secondary or descriptive analysis, we consider it as a limitation. Our goal is to include some brief measures that might serve for hypothesis formulation and exploration for future study directions.

#19. Re data reduction contingencies: I'd say that with factor loading below .40, the modeled latent is at higher risk of being "hijacked" by some idiosyncratic factor (or poor construct validity in general) but that is also the case with factor/component scores or sum scores. The latter two only hide that problem. So although far from ideal, I think it would be reasonable to model the latents as long as the model converges and fall back to some observed scores only if that is not the case.

Reply: We incorporated reviewers' suggestions.

Changes in the manuscript: See Data Preprocessing section, line number: 791-794.

#20. Regarding the use of difference scores, I think it is statistically superior to use residual scores. E.g., regress pre-match baseline score on resting baseline and taking the residual.

Reply: We are aware of the difference score limitations. As we found in our meta-analysis of psychophysiological literature on positive emotions, it is the most common method of calculating the reactivity to affective stimuli (Behnke, Kreibig, et al., 2022). Thus, as we don't want to diverge from the existing psychophysiological literature (also a request from reviewer 1, #2), we will keep using the difference scores for calculating the physiological reactivity. Furthermore, examining raw changes in signals also carries important meaning about the "magnitude" of effects given that the data are presented in the same metrics they were collected in. If we are able to find a statistician who is able to help us with multiverse analysis, we will include the residual scores as one option for physiological measures.

#21. Re missing data: the plan is to exclude participants with missings on a per-analysis basis. If you use SEM, that makes little sense. Why not use full-information maximum likelihood to impute the missing values? Mplus can do that easily. Deletion means discarding information and is only ok when the data are missing completely at random.

Reply: We incorporated reviewers' suggestions, and we introduced the full-information maximum likelihood to impute the missing values.

Changes in the manuscript: See Data Preprocessing section, line number: 818-819.

#22. Any conceptual or statistical reason to exclude outliers with a such number of observations and type of variables?

Reply: As suggested by the reviewer, we will use FIML estimation to handle the missing data in the SEM models. We will identify outliers with the median absolute deviation (MAD), with a cutoff of 3, as recommended by Leys et al. (2013, 2019). We will then delete the data if the data is identified as an error. We do not expect any measurement errors and encoding errors in affective experience data and gaming data. Even if a data point is identified as an outlier, we will not delete it if it represents real data rather than an error. For the cardiovascular data, we will double-check the identified outliers. If we find biologically impossible values, we will delete them. We will report the number of outliers for a given variable.

Changes in the manuscript: See Data Preprocessing section, line number: 818-827.

#23. After controlling for the effect of the intervention, your model implicitly assumes that the covariances between the residuals of mediators are all zero. Is that what you want? If untrue, this represents a model misspecification that will propagate throughout the model and bias the other estimates.

Reply: We added to the model the residual covariances among mediators. We also added them to the power analysis. We used data from our previous study (Behnke, Gross, et al., 2022) to estimate the covariances between the residuals of mediators. Adding the residual covariances among mediators indicated that we need 1800 cases. Thus, as we secured money to recruit 300 participants (and expect 10-20% of the sample to be reduced due to physiological recording problems and voluntary attrition), we decided to increase the number of matches for the 250 gamers to eight.

Changes in the manuscript: See Sampling Plan section, line numbers: 392-393, Procedure section, line number: 591, Table 1. Design Table, and Supplementary Materials - Power Analysis Script and Main Analysis Script, line numbers: 3813-4440, 4579-4619.

#24. RMSEA and CFI are okay approximate fit indices, but why is the plan for model fit evaluation missing the only formal model test, the chi^2 test? I'd definitely want to see that. Maybe also the SRMR.

Reply: We agree that the noncentrality-based indices (RMSEA and CFI) were insufficient. We will also present absolute fit indices χ^2 and SRMR. We will evaluate multiple fit indices as evaluating any single index might be problematic (e.g., a significant $\chi 2$ 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).

Bergh, D. (2015). Sample size and chi-squared test of fit—a comparison between a random sample approach and a chi-square value adjustment method using Swedish adolescent data. In *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings: Rasch and the Future* (pp. 197-211). Springer Berlin Heidelberg.

Geiser, C. (2012). *Data analysis with Mplus*. Guilford press.

Kenny, D. (2023, February 17). Measuring Model Fit. http://www.davidakenny.net/cm/fit.htm

Changes in the manuscript: See Analytic Plan for Primary Hypotheses section, line number: 884-891.

#25. Is it correct that you won't be interpreting effect sizes if there will be a significant model-data misfit? Btw, ironically, low loadings help lower the chi^2 value.

Reply: Yes, if the fit indices suggest model misfit, we will not interpret effect sizes.

Changes in the manuscript: See Analytic Plan for Primary Hypotheses section, line number: 891.

#26. "We do not plan to use the same approach for hypothesis 2 because we did not find a way to operationalize the smallest effect of interest for cardiovascular responses". Alternatively, it is fairly easy to compute Bayes factors for individual model parameters by a model selection approach, using just the BIC (Bayesian information criterion) approximation – comparing the BIC of models with and without the given parameter (see Wagenmakers, 2007). No SESOI or prior needs to be specified (a weakly informative unit information prior is implicitly assumed) and BIC can easily be extracted for any model. Presenting the continuous BFs alongside equivalence tests may be informative for the readers.

Reply: We clarified this issue. We selected the SESOI to show the effect sizes that should be practically interesting. By practically interesting, we mean changes that are likely to be observed by people and to make an impact in real circumstances. In our case, this means changes in the affective experience that people are able to notice and changes in performance scores that would (in our opinion) be substantial for competitors. But, we could not figure out what physiological change would make a practical difference. Thus, we did not calculate the

SESOI for cardiovascular changes, and we do not aim to draw conclusions about the practical value of any observed physiological changes. In the reviewers' words, here we aim to see whether the intervention creates "the effect" and not only a "significant effect". On the other issue, as all data and scripts will be fully accessible, all readers will be able to use different analytical strategies to answer new research questions or test the robustness of our findings.

Changes in the manuscript: See Analytic Plan for Primary Hypotheses section, line number: 898, 903-906.

#27. "We will test the robustness of our findings by adding to the primary model the moderation of the negative prior mindsets, negative appraisals, and gaming experience." Robustness? How? The target causal effect is identified regardless. Btw, I think it would make things clearer if you framed your research goals as either testing causal effects (intervention –> mediators, outcome) or examining mechanisms through which the causal effects operate (mediation effects).

Reply: We clarified this issue. We no longer state that the moderator analysis will test the robustness of our findings. Furthermore, we state that our primary research aim is to test the causal effect of the synergistic mindset intervention, but we also designed the study in a way that will allow us to examine the mechanism through which the causal effects operate. We do not see why we should choose only one option.

Changes in the manuscript: See Present Study section, line number: 257-259.

#28. Exploratory Analyses section: "We treat the moderations as exploratory analysis because the initial studies were inconclusive on whether the prior mindsets moderate the effects of synergistic mindset intervention on cardiovascular and performance outcomes (Yeager et al., 2022)." I see your point but the fact that "initial studies were inconclusive" is irrelevant with respect to the inclusion of a moderator to the model. The thing is, that the inclusion of a moderator and modeling (incoming and outgoing) paths to the treatment and outcome nodes is an act of expressing ignorance about the presence of the given paths. Meaning, there may or may not be an effect. It's fine to choose your confirmatory research aims but justifying it based on the inconclusiveness of prior research appears conceptually weak to me.

Reply: We clarified this issue. We did not find strong enough evidence for moderators to include them in the primary model.

Changes in the manuscript: See Exploratory Analyses section, line number: 921-922.

#29. Also Exploratory Analyses section, "We will also test the robustness of our findings by testing alternative operationalizations ,of the variables used in the model. For positive/negative affect, we will use the sum of the positive/negative items instead of the latent factor." The sum score is only a special case of a latent variable model, where you assume equivalence of factor loadings and reliabilities of all measured indicators equal to 1. Therefore, from a measurement perspective, I personally wouldn't qualify the interpretation of the robustness of the conclusions based on employing psychometrically inferior measurement models. Instead, I would only plan using observed sum scores as a fallback plan if they were locally under-identified (say in case of collinearity issues) or producing estimation issues due to the violation of local independence assumption (large

residual covariances). That is far from ideal as sum scores in such case hide serious measurement issues but at least provides you with an opportunity to empirically address your target research questions, albeit more tentatively. Btw, if you really were to resort to observed scores as a fallback, I'd use a PCA component score, where you don't assume equal component loadings at least.

Reply: We incorporated reviewer suggestions and no longer plan to use sum scores in our primary analysis. If the model fails to converge, we will explore other reasonable models, e.g., using Principal Component Analysis component scores.

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. We will address the interpretation of the robustness of the conclusions based on potentially psychometrically inferior measurement models as a limitation of our study

Changes in the manuscript: See Data Preprocessing section, line number: 791-794, and Exploratory Analysis section, line number: 927-931.

#30. Third, "For positive/negative affect, we will also try the single difference score (sum of negative emotions subtracted from the sum of positive emotions)". Do you mean the difference in *mean* scores if they are using the same scale? If there will be some missing data, subtracting sums will not work.

Reply: As the reviewer pointed out some weaknesses of mean and sum scores, we no longer intend to use a single score.

#31. I am a fan of testing the robustness of findings by employing alternative operationalizations. But with so many, how are you going to do that specifically? There will be quite a few combinations. Maybe you should consider doing a multiverse analysis for these robustness checks.

Reply: Conceptually, this is an interesting idea, but multiverse analyses are themselves tricky to interpret, as they do not allow for clear statistical inferences, and mainly describe patterns across studies. We have recently worked on methods to allow better statistical inferences (Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., & Finos, L. (2022). *Post-selection Inference in Multiverse Analysis (PIMA): An inferential framework based on the sign flipping score test* (arXiv:2210.02794). arXiv. https://doi.org/10.48550/arXiv.2210.02794 but feel it is too early to implement them. As the description of the exploratory analysis is not mandatory in Stage 1 of the registered report, we will plan to identify if our collaborators on the multiverse paper are interested in implementing a multiverse analysis approach for SEM models - but we can make no guarantees at this time.

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.

Changes in the manuscript: See Exploratory Analysis section, line number: 936-937.

#32. It is practically not feasible for me to review the measures as only links for items are provided.

Reply: Due to the privacy policies, we provide links to the sources of the items used in the study rather than the content of the original items. The items' content should be available while minimizing the risks of privacy policies or any legal violations.

#33. The last one and important. When I read the control condition instruction, it seems obvious to me that this condition likely doesn't elicit the same degree of expectancy as the reappraisal manipulation. The control condition needs to seem smart and face valid to the participants, but be inert w.r.t. the outcomes. I encourage the authors to think about how to make the control condition,far more believable and applicable because at the moment, any post-treatment difference in the outcome between the experiment groups may be due to a likely substantial difference in the strength of demand characteristics perceived by the participants. Apart from that, it does not help if only "the synergistic mindsets group will report the adherence and progress in scheduled affect regulation training" (got that right?). Just a few examples. Maybe it's just me but that sounds trivial to me and I definitely wouldn't expect any effect on my esport performance or affect by reading about Phineas Gage and suchlike stuff:

"What I didn't expect, however, was that the brain is so involved when I'm playing! In fact, everything I need to play - seeing the map, hearing my teammates on headphones and speaking to them, moving around the map with the mouse and keyboard, thinking about my next move, anticipating my teammates' and opponent's moves - are all made possible by different areas of my brain!"

"I now know that my eyes are not responsible for my vision and that I know how to go to the local store is due to my temporal lobe. I learned that when I am stressed, my behavior depends on the cooperation of two parts of the nervous system - one responsible for normal functioning and the other responsible for immediate reactions."

"On the other hand, it is sad how much brain damage can impair further functioning. Fortunately, thanks to such injuries, scientists are learning more and more about how the brain works and how to treat various diseases."

Reply: As clarified in response to #2, we are using a control intervention that, in one or another version, was contrasted with the mindset interventions in many studies, with a total *N* > 15k, and published in *Nature* twice (Yeager et al., 2016, 2019, 2022). Our work aims to answer the initial study authors' call for new large-scale trials in diverse populations and contexts (Yeager et al., 2022), so we don't want to deviate far from the original studies. This is why we adapted both interventions to the esports performance context.

Before the adaptation, we considered two other options, but none of them was better than the original control intervention:

1) Creating a new control intervention – if we aimed to create the intervention, following the same procedures to maintain the quality of the original one, it would require another whole project just to create it – as it is clearly described in Yeager et al., 2016. Given the grant timeline we are obliged to follow, it is not possible to create the new intervention from scratch and run a study presented in the manuscript.
2) No intervention at all/waiting list – we believe it would make an even bigger difference in the study's demand characteristics.

Creating a new intervention or using no intervention/waiting list would also create another problem of results comparability between our and original SMI studies.

Thus, we decided to create a cover story for this project in which we say we are creating a *psychological training program for the future generation of gamers.* As with all training programs, it is created from different elements – in our case, modules. We will tell participants that they will see one of the modules that we plan to use in the training program, but before including them, we have to test which elements of the program are beneficial to gamers – some gamers will receive the module about the brain, whereas some about stress and emotions. In this way, the participants should be less confused about why they are learning about the brain.

Furthermore, teaching about physiology and the brain is usually a part of mental training programs for athletes. Usually, it takes the form of a psycho-educational talk where the athletes learn about the foundations of how their body and mind work when the term 'physiological arousal' is introduced (Behnke et al., 2019; Röthlin et al., 2016). From this perspective, a module about the nervous system is the typical beginning of mental training for athletes.

Finally, the examples mentioned by the reviewer are authentic statements provided by the CS: GO gamers whom we asked for help adapting the control condition.

That being said, we appreciate the skepticism of the reviewer, but when comparing the original control intervention (with all its limitations) to other options (with unknown limitations), we decided to keep the original control intervention. This is a strategic decision, and we are ready to take the risk of testing the synergistic mindset intervention against the current control condition and to consider the demands differences as a possible limitation.

We also added some modifications to our study. First, we added the question for the control group to the daily questionnaire in which they will be asked to: "List some of the gaming situations that happened to you today". To decrease the demands' difference between the conditions.

Second, when presenting the project flow in Stage 1, the experimenter will emphasize to the participant that:

> *We are testing different modules during the research, but due to logistics and time constraints, we will only show you one of them. We want to test how each module works separately before adding them to the training program.*

Third, in the exploratory analysis, we will test the effects of synergistic mindset intervention when controlling for the differences in intervention evaluations.

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., ... & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of educational psychology*, *108*(3), 374.

Changes in the manuscript: See Procedure section, line number: 469-472, 534-540, 579-580.

To sum up, I think the present proposal has merits and can provide rich data and relatively robust findings. What worries me most is (1) the outlook of poor measurement and (2) pretty obvious differences in demand characteristics of treatment and control conditions. I also feel the study underutilizes the data, where the authors are depriving themselves of the opportunity to examine interesting questions (specifically using the longitudinal measurements, modeling more complex models, and

looking at the follow-up). But I am a fan of the principle that authors should be free to study what they want. Anyway, this got way too long (sorry for that) so please, feel free to integrate what you see fit and react only to what requires a reaction.

Good luck with the revision!
Best wishes,

Reply: Thank you for the thought-provoking comments. We hope that the existing limitations are not a disqualifying factor for our study but issues that can be addressed in the limitation section.

Ivana Piterová (Reviewer 3)

Thank you for the opportunity to review the Mplus scripts for this RR.

#1. Power analysis script: In this script, the semicolon at the end of the line (NAMES =...) is missing, so the code firstly reports an error, but after this correction, the code works well and produces the reported results, so I can confirm the reproducibility of the calculations with this script in Mplus, listed in the supplements.

Reply: Corrected.

#2. Primary analysis script: The mediation code is correct and on simulated data produces the expected results without errors or warnings.

Reply: That's good to hear.