**Managing Board of PCI Registered Reports**

---

**Dear Managing Board of PCI Registered Reports,**

We were pleased to receive the review of our Stage 1 Registered Report entitled *Can playing Dungeons and Dragons be good for you? A registered exploratory pilot program using offline Tabletop Role-Playing Games (TTRPGs) to mitigate social anxiety and reduce problematic involvement in multiplayer online videogames* in your email of February 12, 2023, and your willingness to consider a revised version of our manuscript. We are grateful for the promptness of the review process and thank all peers involved for their insightful comments that were helpful in improving our manuscript.

First of all, we would like to apologize for the delay in completing this resubmission, especially given the promptness of your review process. This delay is primarily due to the resumption of the academic year coupled with a substantial teaching and administrative workload. Importantly, we have decided to delay the beginning of data collection (baselines) to the first or second week of April 2023. Figure 2 (i.e., "Study design and representation of the various steps of the psychological assessment") has been updated accordingly. The recruitment phase (eligibility) started end of February and is planned to finish by the end of March. The start of data collection is planned for mid-April. Delaying the beginning of data collection also allowed us to send back the current revised Stage 1 Registered Report before the actual beginning of data collection.

To facilitate the review of our revised manuscript, the recommender's and reviewers' comments are numbered and reported in table cells, below which are our responses. For ease of reference, all modifications to the revised manuscript have been highlighted in yellow. We thank you for your consideration of this revised manuscript and we look forward to hearing from you.

Please also note that professional proofreading of the paper will be performed prior to submission of our Stage 2 Registered Report, pending our revised Stage 1 Registered Report being convincing enough to be recommended for publication.

Best regards,

Joël Billieux, on behalf of all authors

---

**Joël Billieux** (Corresponding author at: joel.billieux@unil.ch)
Institute of Psychology, University of Lausanne, Switzerland

**Recommender – Veli-Matti Karhulahti**

---

**1.0.**

Dear Authors,

Three reviewers have generously provided detailed rapid feedback, considering your hard deadline. They are all positive, but some critical things need to be carefully considered.

- **Answer:** We are very grateful and would like to warmly thank these reviewers for their time and willingness to review our Stage 1 Registered Report. Their comments were very helpful and definitively improved and/or clarified the study protocol.

**1.1.**

The MS sits between an exploratory pilot and a confirmatory intervention: a key goal is to explore feasibility, but there are also hypotheses to be tested. As reviewers point out, hypothesis testing would require solid corroboration/falsification rules and clarity when success would be left undecided. A complete data analytic plan regarding how efficacy will be measured would be needed for assessing hypothesis testing. It also remains possible to register this as an exploratory pilot, in which case evaluation is more flexible (but you cannot make confirmatory claims at Stage 2). Although I personally see the exploratory option most feasible – especially considering your time limit – below is a list to help you revise if you wish to pursue hypothesis testing (skip this if you choose the exploratory path).

1.1.A. There are discrepancies between the hypotheses on p. 8 and the expected outcomes on p. 22. E.g., PO1 concerns gaming frequency, but this is not among the previously named hypotheses. It's important to consistently justify each hypothesis; you may also set expectations without testing them (= no confirmatory claims at Stage 2), but they need to be clearly distinguished from tested hypotheses.

1.1.B. Justify the smallest effect of interest. Currently only the term "reduction" is used, but we need to be more specific. E.g., reduction of gaming by 1min/day would hardly be meaningful. Each effect/hypothesis used for confirming effectiveness needs a justification, respectively. See, e.g., Anvari et al. (2022; https://doi.org/10.1177/17456916221091565).

1.1.C. All outcomes are currently expected both at the end of the TTRPG-based program (P1A) and at the 3-month follow-up (P1B). We need to agree beforehand which of these, or what combination thereof, corroborate/falsify hypothesis. E.g., what if we see no reduction at P1A but reduction at P1B, would this corroborate hypotheses?

1.1.D. Considering that some effects will not be meaningful, please specify when the result will be considered null, i.e., what are the results that will conclude the intervention had no meaningful effect or a non-meaningful effect.

1.1.E. Carefully consider how dropouts are assessed. E.g., what if you have 50% (10/20) dropouts and find meaningful effects in the remaining participants, would this be considered corroborating hypotheses?

1.1.F. What about missing data, e.g., if a participant fails to deliver P1B data, will this be considered a dropout? What is the overall rule structure, considering all scenarios, for corroboration and falsification of hypotheses?

1.1.G. A complete data analytic plan would be required for each to-be tested hypotheses.

Because constructing a robust hypothesis testing design within the present time limitations may be challenging, you may also choose a simplified confirmatory design where only feasibility is tested. Following the main goal of the study ("to test the feasibility, e.g., number of dropouts – ability of the participants to complete regularly the online assessment"), you could formalize this into feasibility hypotheses.

1.1.H. Define what counts as dropout and justify success/failure by the number of dropouts, e.g., in relation to common dropouts in similar interventions. Consider the degree of flexibility, e.g., with confidence intervals.

1.1.I. Define and quantify online assessments to be completed by participants and justify a sufficient completion rate that will qualify successful and unsuccessful intervention.

The above would allow you to make confirming claims about the practical feasibility of the intervention at Stage 2 with relatively little revision. Note that you can (and should!) also report the current primary/secondary outcomes, but only as non-confirmatory, tentative results that will inform future efficacy testing of the design.

In case you choose either of the two confirmatory designs, please add each hypothesis separately in the design table with justifications. Note that currently some of the explanations are not fully sufficient. E.g., regarding sample justification, you have stated it to be non-relevant, but there should be a justification for having n=20 and not, e.g., n=1 or n=200. I see this is already touched on p. 11. See, e.g., Lakens (2022; https://doi.org/10.1525/collabra.33267). Also, the rationale for confirming and disconfirming hypotheses still appears to be highly relevant for this design (if tested as confirmatory).

Note that if you choose not to test any hypotheses, a fully exploratory approach is totally ok and does not need the design table (or any of the other confirmation concerns either). In this case, make sure to remove the hypotheses and/or clearly state that they will not be tested.

- **Answer:** The recommender made a correct and important point. After consultation with the full research team, and taking into account the nature of our study, we decided to go for an exploratory pilot. Accordingly, no confirmatory claims will be made at Stage 2. In the revised text (title, abstract, main text), we now refer to a "registered exploratory pilot".

- The following changes have also been made according to the list of related concerns expressed by the recommender:

    - **1.1.A.:** As we reframed the study as an exploratory pilot, we removed the formal hypotheses (p.8). Terms such as "efficacy" were replaced by "effect". Sentences are now formulated in an exploratory way.

      Example of modified sentence: *"This pilot also aims to explore whether our program, which is designed to mobilize social skills by exposing the participants to socially engaging situations in real life, affects assertiveness, and self-concept.".*

      In this example, changes made are highlighted in yellow.

      In our revised document, we only kept the operationalized (i.e., variable-centered) expected outcomes (primary and secondary). All formal hypotheses have been removed in the revised version of Stage 1.

      Thanks also for noticing that we did not refer to gaming frequency before the outcomes section. This has been corrected in the revised text.

    - **1.1.B.:** Point not addressed given that we are following the suggestion to go for an exploratory pilot and not a confirmatory intervention.

    - **1.1.C.:** Point not addressed given that we are following the suggestion to go for an exploratory pilot and not a confirmatory intervention.

    - **1.1.D.:** Point not addressed given that we are following the suggestion to go for an exploratory pilot and not a confirmatory intervention.

    - **1.1.E.:** No confirmatory statement will be made based on the dropout rate, as we are no longer formulating hypotheses. Yet, given the nature of the study (exploratory pilot) and the study design (multiple single case analysis), we will perform exploratory analyses to identify potential profiles of participants who dropout versus participants who not dropout.

    - **1.1.F.:** No confirmatory statement will be made based on the dropout rate, as we are no longer formulating hypotheses. Participants with missing data will not be removed from the analyses unless number of measurement points per phase is < 3 inasmuch as three measurement points per phase is considered the minimal

standard to reach in a single-case methodology (Tate et al., 2015). The following sentence has been added to the manuscript:

**"***Participants with missing data will not be omitted from the analyses unless the number of measurement points per phase is < 3, as three measurement points per phase is considered the minimal standard to reach in a single-case methodology (Tate et al., 2015).".*

- **1.1.G.:** Point not addressed given that we are following the suggestion to go for an exploratory pilot and not a confirmatory intervention.

- **1.1.H.:** No confirmatory statement will be made based on the dropout rate, as we are no longer formulating hypotheses. In the revised text, we added a section of what constitutes dropout in our study and how we will handle it. This section reads as follows:

*"The number of participants was determined by taking into account the expected dropout rate and the requirement to provide enough inter-subject replication of the experimental effect. TTRPGs are well suited to groups of 3 to 5 players plus one game master. We opted for the upper limit (5 participants per group) to compensate for potential dropout(s). We decided that the minimal number of participants required to continue playing should be 3 to guarantee sufficient social exposure. If more than 2 participants drop out in the same group, the remaining participants will be allocated (if possible) to another group. In this pilot study, dropout occurs when a participant leaves the program permanently, regardless of the number of session(s) completed. Participants who miss a session for acceptable reasons (e.g., being sick) will have the possibility to reintegrate and continue the program (the number of potentially missed session(s) will be recorded for each participant).".*

**1.1.I.:** No confirmatory statement will be made as we now assume a fully exploratory perspective. Yet, a section has been added regarding the handling of missing data regarding the psychological assessment (see our response to point 1.1.F.).

**1.2.**

| |
|---|
| Title: Because "registered reports" include preregistration, it might be more informative to use the former term in the title. |

- **Answer:** Based on your suggestions and our decision to go fully exploratory, we used the term "**registered exploratory pilot**" in the title, abstract, and main text.

**1.3.**

Figure 2: We're in mid-February, which is the time for filling consent forms. Please update how far the recruitment is when you return the revision. It's totally ok if some data have already been collected (e.g., participant demographics are known), but then we just take this into consideration with bias control (author guidelines section 2.6).

- **Answer:** We have decided to delay the beginning of data collection (baselines) to the first or second week of April 2023. Figure 2 (i.e., "Study design and representation of the various steps of the psychological assessment") has been updated accordingly. The recruitment phase (eligibility) started end of February and is planned to finish by the end of March. The start of data collection is planned for mid-April. Delaying the whole study also allowed us to send back the current Stage 1 Registered Report before the actual beginning of data collection.

**1.4.**

P. 10: Will one of the team members serve as a game master or is this an external expert? Please clarify.

- **Answer:** Jonathan Bloch, a member of the research team, is our game master. We were totally transparent about this in the "Author contributions" section:

  *"Jonathan Bloch elaborated the various modules of the 10-week TTRPG-based program under the supervision of Joël Billieux. Jonathan Bloch will administrate – as the game master – the TTRPG program to the four groups of participants."*.

  We also mentioned the following in the "Procedure" subsection:

  *"The Game Master in charge of managing the TTRPG sessions will not have access to the results of the various psychological assessments conducted during the study."*.

**1.5.**

P. 11: Because participants with as few as 1/9 IGD symptoms are included, it remains a bit unclear how this will affect the analytic strategy and the interpretation of results. E.g., there is some evidence that 2/9 symptoms are connected to lower wellbeing (Ballou & Zendle, 2022: https://doi.org/10.1016/j.chb.2021.107140), but it's not clear how the reduction from 1/9 to 0/9 symptoms should be interpreted. Would it imply the participant's health/wellbeing improved?

- **Answer:** Each item of the IGD-10 is scored based on frequency statements (0 = "never"; 1 = "sometimes"; 2 = "often"). For the eligibility screening, we will follow the suggestion by Király et al. (2009) and consider responses "never" and "sometimes" as an absent

criterion (0 point) and responses “often” as a present criterion (1 point). As two items refer to the last DSM-5 criterion (i.e., items 9 and 10), they will be combined during the scoring procedure (Király et al., 2017). This coding is used to match with the categorical structure of the DSM-5 (in which criteria are either present or absent) and identify potentially problematic gamers during the eligibility screening (endorsement of ≥ 5 criteria according to the DSM-5 guidelines). For all statistical analyses conducted, a total score ranging from 0 to 20 will be used instead to increase the variance of the scores and thus increase the likelihood to evidence change.

Accordingly, the dichotomous IGD criteria will only be used for the eligibility screening phase, and we aim to increase scores’ variance by using the polytomous 3-point Likert scale for the psychological assessment that will take place during the study (baselines, during program, follow-up). The above-mentioned points are explained in the section describing the IGDT-10 instrument.

That being said, we totally understand the concern raised regarding the interpretation of a decrease (or increase) of the self-reported GD symptoms. For this reason, no confirmatory statement will be made regarding the efficacy of our program. Also, no cut-off (e.g., in terms of GD symptoms) will be used to interpret a potential effect of our program.

Importantly, our objective is rather to provide – thanks to the comprehensive single case design adopted – an individual and idiosyncratic fine-grained exploratory analysis of primary and secondary outcomes for each participant included (including GD symptoms, but not only).

References cited in this answer:


- Király, O., Bőthe, B., Ramos-Diaz, J., Rahimi-Movaghar, A., Lukavska, K., Hrabec, O., Miovsky, M., Billieux, J., Deleuze, J., Nuyens, F., Karila, L., Griffiths, M. D., Nagygyörgy, K., Urbán, R., Potenza, M. N., King, D. L., Rumpf, H.-J., Carragher, N., & Demetrovics, Z. (2019). Ten-Item Internet Gaming Disorder Test (IGDT-10): Measurement invariance and cross-cultural validation across seven language-based samples. *Psychology of Addictive Behaviors*, *33*(1), 91–103. https://doi.org/10.1037/adb0000433

- Király, O., Sleczka, P., Pontes, H. M., Urbán, R., Griffiths, M. D., & Demetrovics, Z. (2017). Validation of the Ten-Item Internet Gaming Disorder Test (IGDT-10) and evaluation of the nine DSM-5 Internet Gaming Disorder criteria. *Addictive Behaviors*, *64*, 253–260. https://doi.org/10.1016/j.addbeh.2015.11.005


**1.6.**

| |
|---|
| P. 13: The participants will be randomly distributed into 4 groups, but is that optimal? Considering that the study addresses social anxiety, taking into consideration, e.g., |

gender in group distribution seems relevant. Imagine you have 5 women and 15 men; having mixed groups would likely lead to different outcomes vs if all men and women would be in gender-based groups. Which would be better in the light of current knowledge?

- **Answer:** We believe this remains an empirical question whether variables such as gender, sexual orientation, educational level, or age could have an influence in such a context or facilitate/complicate interactions between participants. In the absence of supporting evidence, we initially decided to rely on a random distribution.

  Yet, we thank the recommender for this important comment because it made the research team think about the randomization process in more detail. We concluded that it will not be possible (nor ideal) to randomly assign participants in the different groups, for the following reasons:

  - Groups will be constituted to maximize heterogeneity in terms of gender, age, and education level (this information will be collected during the eligibility screening).
  - The availabilities of the participants will not necessarily be the same (e.g., different groups will potentially play at different times and days).
  - It cannot be excluded that some participants know each other. In such case, they will be allocated to different groups.

Accordingly, we have removed the term “randomized” from the manuscript and have provided more details regarding the strategy used to constitute our groups of participants in the revised text. This new paragraph reads as follows:

*“Distribution of participants in the various groups will be done according to feasibility constraints, including (1) their availabilities (the various groups play at different times of the day and/or on different days of the week), (2) ensuring that no participants knowing each other are included in the same group, and (3) maximizing heterogeneity in terms of gender, age, and education level.”*.

**1.7.**

P. 20: Qualitative feedback is collected. Please also explain how and what kind of, and how it will be analyzed in this study (if it is).

- **Answer:** Qualitative feedback will be collected after specific sessions (see Table 1) by the Game Master and another member of the research team (see authors contributions). The feedback collected will be summarized in a table and uploaded to the Open Science Framework. If relevant points are formulated, they will potentially be considered during Stage 2, but at this point on time we cannot say more about how we will use these qualitative feedbacks.

**1.8.**

| |
|---|
| P. 22: PO1 mentions frequency and hours, both. In my understanding, frequency refers the number of times of engagement (“three times per day”), not the total time of engagement (“three hours per day”). Please clarify. |

- **Answer:** Thanks for spotting this issue. We corrected this in the revised text.

**1.9.**

| |
|---|
| P. 23: It is noted that deviations will be justified at Stage 2, but I must note that PCI RR guidelines (section 2.10) advise authors to consult the recommender for deviations immediately and prior to the completion of data collection whenever possible. If you choose to have this as a fully exploratory RR, deviations are more flexible. Especially if any confirmatory elements remain, it remains important to notify of them as soon as possible. |

- **Answer:** Although we decided to go for an exploratory RR (and not a confirmatory RR), the team of authors took good note of this. The related sentence has been amended accordingly: *“Any deviation from this pre-registered data analytic plan will be discussed with the recommender and described and justified in the final version of the registered exploratory pilot.”*.

**1.10.**

| |
|---|
| Scales: because at least some of the scales (like DSM-based IGDT-10) include both core and peripheral construct criteria, it feels reporting omega would be better than alpha. |

- **Answer:** The values currently reported are those of the psychometric validation papers. As our multiple single case design study will only comprise 20 participants, and all assessment instruments used have already undergone previous psychometric validation, we did not plan to report internal reliability coefficients. If this is important, we could report the omega on Open Science Framework at Stage 2.

  That being said, we can only agree on the relevance of considering the distinction between peripheral and core criteria of GD (see, e.g., Castro Calvo et al., 2022). Although we will not pre-register different sets of analyses for peripheral and core criteria, we will keep this suggestion in mind when it comes to analyzing our results. We cannot exclude that, for some participants, this distinction will be relevant.

  Reference cited in this answer:

- Castro-Calvo, J., King D.L., Stein D.J., Brand M., Carmi L., Chamberlain S.R., Demetrovics Z., Fineberg N.A., Rumpf H.-J., Yücel M., Achab S., Ambekar A., Bahar N., Blaszczynski A., Bowden-Jones H., Carbonell X., Chan E., Ko C.-H., de Timary P., Dufour M., Grall-Bronnec M., Lee H.K., Higuchi S., Jimenez-Murcia S., Király O., Kuss D.J., Long J., Müller A., Pallanti S., Potenza M.N., Rahimi-Movaghar A., Saunders J.B., Schimmenti A., Lee S.-Y., Siste K., Spritzer D.T., Starcevic V., Weinstein A.M., Wölfling K., & Billieux J. (2021). Expert appraisal of criteria for assessing gaming disorder: An international Delphi study. *Addiction*, 116, 2463-2475. https://doi.org/10.1111/add.15411

**1.11.**

| |
|---|
| Please also consider the reviewers' separate comments. I hope you find the reviewers' feedback and my additions helpful. You may contact me directly for any clarifications if needed. This is a highly interesting and promising study, and I'm happy do my best to support it.<br><br>Best wishes,<br><br>Veli-Matti Karhulahti |

- **Answer:** Thanks again for the very valuable comments that helped us to improve our study and/or to elaborate and think more on some key aspects of the study (e.g., distribution of participants among groups).

**Reviewer – Matti Vuorre**

---

**2.0.**

| |
|---|
| The authors aim to test the feasibility and initial efficacy of a tabletop role-playing game (TTRPG) intervention on social anxiety and dysregulated gaming. It seems that the TTRPG intervention is designed with great care and informed by expertise & experience in role playing games. Well done! The manuscript addresses a real need in addressing an important problem, but also tries to understand the potential psychological effects of ludic activities. I therefore think that the intervention has promise. My review focuses on the evaluation of the intervention. |

- **Answer:** We would like to warmly thank this reviewer for the positive appreciation of our study.

**2.1.**

| |
|---|
| Authors propose to conduct both a feasibility study and a test of initial efficacy. These two aims seem at odds because the former would only track practical issues in the procedure including things like dropout and whether the participants understand what they are doing etc. The latter would require a detailed statistical investigation of a sufficiently large dataset. In my view the project looks like a success regarding the former aim but falls somewhat short regarding the latter.<br><br>My recommendation is that the authors either consider reframing this manuscript to focus on the first – also valuable – aim, or greatly increase the sample size to allow studying the latter.<br><br>The design involves running 20 individuals through the experimental procedure after baselines of varying duration. Effectiveness is then evaluated by comparing participants' outcome scores during and after the treatment to their baseline scores (at last measure and the 3 month follow up). Authors could clarify what the exact comparison will be: is it average baseline vs. last measure/follow up? I understand the data analytic plan will be pre-registered later, but I didn't find sufficient information here to determine whether they will plausibly have enough precision to evaluate effectiveness. Considering the sample size of 20 I don't think the precision will be sufficient. |

- **Answer:** The opinion of this reviewer echoes some of the concerns raised by the recommender. After consultation with the full research team, and taking into account the nature of our study, we decided to go for an exploratory pilot. Accordingly, no confirmatory claims will be made at Stage 2. In the revised text (title, abstract, main text) we now refer to a “registered exploratory pilot”.

Yet, and importantly, we decided to keep the focus on the primary and secondary outcomes, but adjusted the text according to the concerns of this reviewer. In particular, we removed the formal hypotheses (p.8). Terms such as “efficacy” were replaced by “effect”. Sentences are now formulated in an exploratory way, as you can see from this example:

> Example of modified sentence: *“This pilot also aims to explore whether our program, which is designed to mobilize social skills by exposing the participants to socially engaging situations in real life, affects assertiveness, and self-concept.”*.
>
> In this example, changes made are highlighted in yellow.

Keeping the focus on primary and secondary outcomes is totally feasible with 20 participants, thanks to the fact we rely on a multiple single case design which enables a multiple replication of the effect of the intervention across participants. To determine whether our program affects target variables for specific individuals, we relied on a combination of multiple baselines across participants and multiple assessments per participant (see Figure 2). In this perspective, each participant is their own control, and this is made possible by multiplying the assessment points for each single participant (Tate et al., 2015).

Even though we decided to adjust our study to a fully exploratory design and not to make confirmatory statements, it is important to bear in mind that single-case methodology has unique strengths for assessing the efficacy of a treatment and is considered a clinically relevant and scientifically well-established alternative to traditional group comparison designs (Dattilio, 2006). Accordingly, and this is an important point, our sample size is totally adequate for the anticipated and pre-registered analysis.

Reference cited in this answer:


- Dattilio, F. M. (2006). Does the case study have a future in the psychiatric literature? *International Journal of Psychiatry in Clinical Practice*, *10*(3), 195–203.


**2.0.**

| |
|---|
| Thank you and good luck with the project.<br><br>Respectfully signed,<br><br>Matti Vuorre |

- **Answer:** We respectfully thank this reviewer for their positive comments and their wish of good luck.

**Reviewer – Matúš Adamkovič**

---

**3.0.**

| |
|---|
| Dear Authors,<br><br>Thank you for this interesting submission. The main topic of this RR – piloting the usage of TTRPGs as an intervention to alleviate problems with gaming, self-concept, social anxiety, and loneliness – is highly relevant and innovative. I'm very sympathetic to the fact that the authors decided to submit this pilot as a RR. I'd also like to highlight the rigor of the proposed design. Below, I'll try to provide several suggestions and will also depict some points that, in my opinion, would require further clarification. |

- **Answer:** We would like to warmly thank this reviewer for their positive comments about our work.

**3.1.**

| |
|---|
| Introduction<br><br>The theoretical framework is well-written. I've only two minor suggestions. Please consider adding an estimate of the number of video game players worldwide. Please consider adding subheadings. |

- **Answer:** We added recent statistics regarding video game players worldwide. The section added reads as follows:

  *"Video games are one of the most popular leisure activities worldwide. It is expected that the number of gamers will reach 3.07 billion players in 2023 (Newzoo, 2021)."*.

Reference added in the manuscript:

Newzoo (2021). Global Games Market Report. The VR & Metaverse Edition. https://newzoo.com/insights/trend-reports/newzoo-global-games-market-report-2021-free-version/

- We also added subheadings as suggested.

**3.2.**

| |
|---|
| Goals of the study |

The authors acknowledge that the present RR is a pilot to test the efficacy of their intervention program. They hypothesize that the intervention will reduce GD symptoms, social anxiety, and loneliness. It'll also lead to the improvement of self-concept and assertiveness. The expected outcomes are further summarized in Data analytic strategy and Study Design Table, however, no evidence thresholds (i.e., the evidence needed to dis/confirm a hypothesis) are mentioned. Given it's a pilot study, I've been missing a crucial aspect – a (qualitative) examination of the participants' experiences with the intervention program and the analysis of their feedback. Although the authors briefly mention this in Study Design Table, I think this point requires much more attention throughout the paper. A minor note – the introduction contains many distinct (although related) constructs. I've noticed that, for example, assertiveness, which is one of the focal variables in the study, is first mentioned when describing the potential effects of the intervention program. Please consider introducing the construct earlier in the text.

**Answer:** This comment covers various concerns. Our answer is thus itemized.

- **Evidence Threshold:** The opinion of this reviewer echoes some of the concerns raised by the recommender and some of the reviewers (see, e.g., Recommender point 1.1.). After consultation with the full research team, and taking into account the nature of our study, we decided to go for an exploratory pilot. Accordingly, no confirmatory claims will be made at Stage 2, and no threshold will be used to determine potential efficacy of the program. Yet, and as explained in detail to another reviewer (see point 2.1.), the single case design adopted in the current exploratory pilot allows for considering primary and secondary outcomes for each participant included.

- **Qualitative Analysis:** We thank this reviewer for their relevant comment about the importance of qualitative examination of participants' experiences. Qualitative feedback will be collected after specific sessions (See Table 1) by the Game Master and another member of the research team (see Author contributions). The feedback collected will be summarized in a supplementary Table and uploaded to the Open Science Framework. If relevant points are formulated, they will potentially be considered during Stage 2, but at this point in time we cannot say more about how we will use these qualitative feedbacks.

- **Constructs included in the introduction:** We took care to homogenize and reduce the number of constructs used in the introduction. The term "assertiveness" was removed from the introduction, and we now refer to broader constructs of "social skills" and "self-concepts". The way these constructs are operationalized and assessed is comprehensively detailed in the psychological assessment section.

**3.3.**

Procedure and participants

These two sections are, again, well-written and provide details that will allow independent researchers to carry out a replication study. Figure 2 increases the understanding of the procedure. I, however, got a bit puzzled by the frequency of the psychological assessment. Could the authors clarify it in the text or create a table (maybe not necessarily a table and a graphical extension to Figure 2 will suffice) that will summarize which measure will be administered at what time point? The inclusion/exclusion criteria are clearly summarized and reasonable given the nature of the study. However, why do the authors think that prior experience with TTRPGs should be an exclusion criterion? Furthermore, the authors justify the sample size based on the expected dropout rate and inter-subject replication of the experimental effect. Could the authors elaborate on that? For example, what dropout rate do the authors expect? Will they try to contact the participants who drop out of the study to learn about their reasons? I'm asking this because participants who drop out from the study may be those who felt that the intervention had no (or even adverse) effect on them. Consequently, this could overestimate the success of the intervention. Is there a possibility to control for that?

- **Answer:** This comment covers various concerns. Our answer is thus itemized.

  • **Frequency of Psychological Assessment:** As our study employs a single case design (and not a classical group comparison design), it was necessary to multiply the number of assessment points per participant. Thus, to determine whether our program affects target variables for specific individuals, we relied on a combination of multiple baselines across participants and multiple assessments per participant (see Figure 2). In this perspective, each participant is their own control, and this is made possible by multiplying the assessment points for each single participant (Tate et al., 2015).

  • **Readability of Figure 2:** In Figure 2, the intersections between the "Week number" columns and the "Psychometric instruments" rows indicate what psychometric instrument is administered at what time across the baseline, intervention, and follow-up phases. For example, in Group 1, during the baseline phase, all five psychometric instruments will be administered at Week 1, whereas three psychometric instruments will be administered at Week 2 and Week 3 (i.e., IGDT-10, LSAS, UCLA-LS). The rationale underlying the notion of how many times (and when) each psychometric instrument is administered is detailed in the "Psychological assessment" subsection of our manuscript. Additionally, we have now detailed the latter in the legend of Figure 2 in the revised version of our manuscript.

  • **Prior experience in TTRPG:** We reasoned that prior involvement in TTRPG might have influenced some key psychological factors assessed in the study (e.g., social skills, self-concepts, loneliness, or social anxiety symptoms) and would have thus constituted a confounding factor. Also, we found important that all participants included in the study follow a comparable and progressive exposition to playing TTRPG. Eventually, mixing participants with and without TTRPG prior experiences (or with different levels of TTRPG

prior experiences) would have resulted in unbalanced situations among participants (potentially easier for some participants and harder for others, in function of their prior experiences in playing TTRPG).

A footnote was added in the revised manuscript to explain the rationale behind this exclusion criteria:

*"We reasoned that prior involvement in TTRPGs might have influenced some key psychological factors assessed in the study (e.g., social skills, self-concepts, loneliness, or social anxiety symptoms) and thus would have constituted a confounding factor. Also, we considered it important that all participants included in the study undergo comparable and progressive exposure to TTRPGs. Eventually, mixing participants with and without prior experience with TTRPGs (or with different levels of prior experience with TTRPGs) would have resulted in unbalanced situations between participants (potentially easier for some participants and harder for others, in function of their prior experience with TTRPGs)."*

- **Dropouts:** No confirmatory statement will be made based on dropout rate, as we are no longer formulating hypotheses. Yet, given the nature of the study (exploratory pilot) and the study design (multiple single case analysis), we will perform exploratory analyses to identify potential profiles of participants who drop out versus participants who do not drop out.

In the revised text, we added a section of what constitutes dropouts in our study and how we will handle them. This section reads as follow:

> *"The number of participants was determined by taking into account the expected dropout rate and the requirement to provide enough inter-subject replication of the experimental effect. TTRPGs are well suited to groups of 3 to 5 players plus one game master. We opted for the upper limit (5 participants per group) to compensate for potential dropout(s). We decided that the minimal number of participants required to continue playing should be 3 to guarantee sufficient social exposure. If more than 2 participants drop out in the same group, the remaining participants will be allocated (if possible) to another group. In this pilot study, dropout occurs when a participant leaves the program permanently, regardless of the number of session(s) completed. Participants who miss a session for acceptable reasons (e.g., being sick) will have the possibility to reintegrate and continue the program (the number of potentially missed session(s) will be recorded for each participant)."*.

**3.4.**

| Psychological assessment<br><br>Just a minor suggestion – since all the measures have been well-established in the psychological literature, the descriptions of the measures could be shortened/moved to |
|---|

supplementary material. The description of the TTRPG program is detailed and all the supplementary files help understand the procedure.

- **Answer:** Thanks for the positive comment regarding the description of the TTRPG program. Regarding the psychometric instruments, it is important for us to keep the comprehensive descriptions which allow to understand how the central variables of the study are operationalized from a psychometric perspective. We hope this reviewer will agree with our decision.

**3.5.**

Data analytic strategy

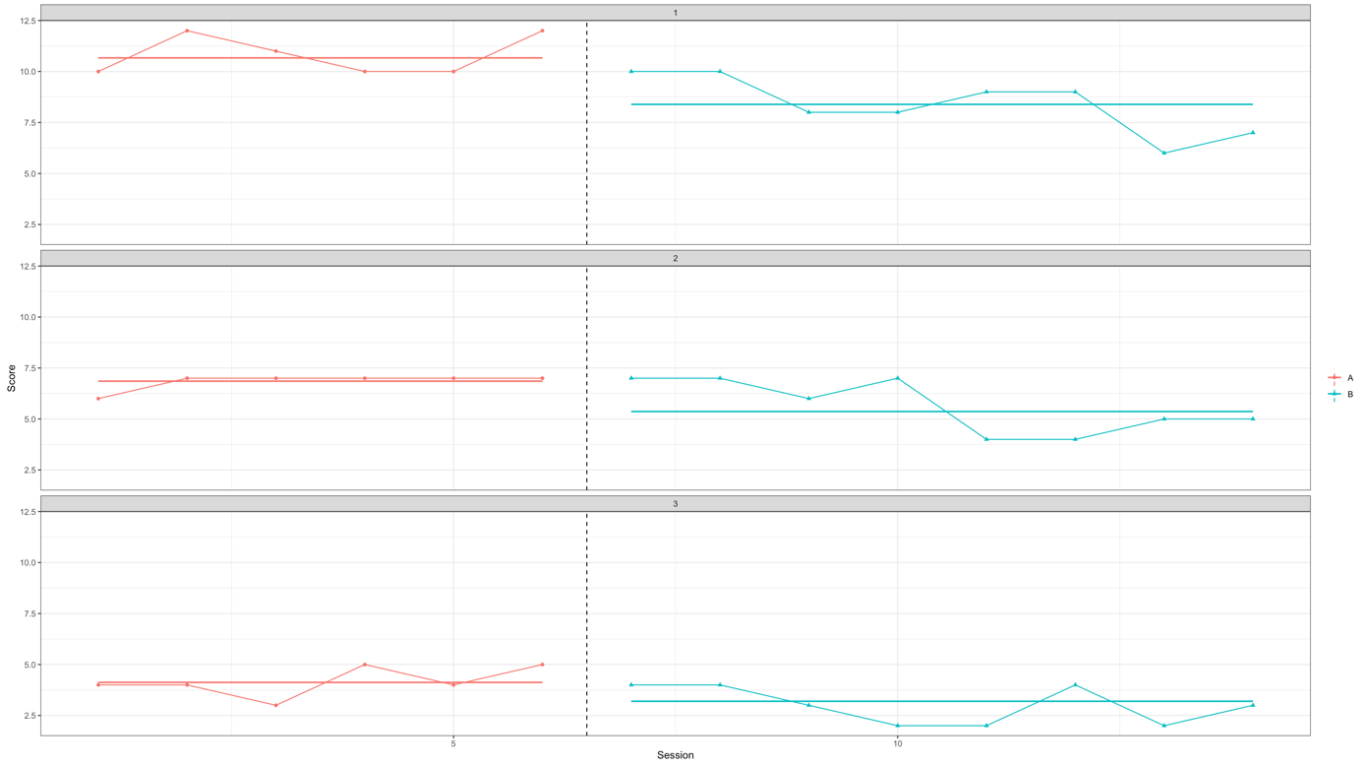
Although I'm not familiar with single-case data analysis, the proposed analytical workflow appears to be well-thought. I especially appreciate the authors' decision to use multiple analytic approaches and test the robustness of their findings. As I mentioned above, please consider specifying the evidence thresholds (not necessarily based on p-values given the sample size). Please consider providing the analytic code (with a simulated dataset) at Stage 1. I also wonder – will the authors control for potential confounders (besides the inclusion/exclusion criteria)? A minor comment – the link for the SCDA package / Rcmdr plugin doesn't work.

**Answer:** We would like to thank the author for the positive comment on the data-analytic strategy. This comment covers various concerns. Our answer is thus itemized.

- **Evidence Threshold:** The opinion of this reviewer echoes some of the concerns raised by the recommender and some of the reviewers (see, e.g., Recommender point 1.1.). After consultation with the full research team, and taking into account the nature of our study, we decided to go for an exploratory pilot. Accordingly, no confirmatory claims will be made at Stage 2, and no threshold will be used to determine potential efficacy of the program. Yet, and as explained in detail to another reviewer (see point 2.1.), the single case design adopted in the current exploratory pilot allows for considering primary and secondary outcomes for each participant.

- **Analytic Code and Simulated Dataset:** We have also provided in the open science framework a simulated dataset for three participants on the IGDT-10 (one of the primary outcomes), the related analytic code (R script for the NAP test and for the between-case standardized mean difference) as well as the obtained results. The simulated dataset, related results, and analytic codes are available from the OSF: https://osf.io/3pgt7/. Yet, to ease the review process, please find below the main outputs as well as a short description of the obtained results (for individual and between cases differences):

**Supplementary Figure 1. Results from the between-case standardized mean difference test (BC-SMD) on the IGDT-10 total score.**




**Supplementary Table 1. Results from the non-overlap of all pairs test (NAP) on the IGDT-10 total score.**

| Participant | Estimate | SE | 95% CIs |
|---|---|---|---|
| 1 | 0.94 | 0.05 | 0.64 – 0.99 |
| 2 | 0.77 | 0.12 | 0.46 – 0.93 |
| 3 | 0.82 | 0.11 | 0.51 – 0.95 |

Results indicate a large effect size for participant 1 (meaning that any randomly drawn observation from the intervention phase has a 94% probability of being lower than any randomly drawn observation from the baseline phase), whereas effect sizes are moderate for participants 2 and 3. Thus, according to the NAP test, the intervention is moderately to largely effective in decreasing gaming disorder symptoms.

**Supplementary Table 2. Results from the between-case standardized mean difference test (BC-SMD) on the IGDT-10 total score.**

| BC-SMD estimate | SE | 95% CIs | Degrees of freedom |
|---|---|---|---|
| -0.28 | 0.61 | -2.70 – 2.19 | 2.19 |

Results indicate that – for the whole sample – gaming disorder symptoms decreased by 0.28 standard deviation from the baseline phase over the course of the intervention, which corresponds to a low effect size according to Cohen's criteria (1988).

- **SCDA package / Rcmdr:** we have slightly modified and simplified the description of the analytic plan. The SCDA package/Rcmdr is unnecessary inasmuch as the two other packages used also provide a graphical presentation of the data for each participant. We have also provided the specific references regarding the R packages used for running the analyses. The data analytic procedure will be pre-registered when and if our Stage 1 exploratory pilot is accepted. All data analytic codes will be available for reproducibility purpose at Stage 2.

*"Any deviation from this pre-registered data analytic plan will be discussed with the recommender and described and justified in the final version of the registered exploratory pilot. NAP tests and between-case standardized men difference will be computed on R 4.2.2 (R Core Team, 2023) with the following packages:*

- *SingleCaseES package (Pustejovsky, Chen, Grekov, & Swan, 2023; https://jepusto.github.io/SingleCaseES/)*
- *scdhlm package (Pustejovsky, Chen, & Hamilton, 2023; https://CRAN.R-project.org/package=scdhlm)"*

The following references have been added:

Pustejovsky, J. E., Chen, M., Grekov, P., & Swan, D. M. (2023). SingleCaseES: A calculator for single-case effect size indices (Version 0.7.1) [R package]. https://jepusto.github.io/SingleCaseES/

Pustejovsky, J. E., Chen, M., & Hamilton, B. J. (2023). scdhlm: Estimating hierarchical linear models for single-case designs (Version 0.7.2) [R package]. https://CRAN.R-project.org/package=scdhlm

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Autria. https://www.r-project.org/

**3.6.**

| Transparency<br><br>The authors (will) share all the data and materials at the study's OSF project (https://osf.io/3pgt7/). I'd like to appreciate this level of transparency and authors' adherence to open science practices. |
|---|

- **Answer:** Thank you so much for the positive comment.

**3.7.**

| I hope the authors will find the suggestions useful. Looking forward to reading the revised version of the RR. |
|---|

Best,

Matúš Adamkovič

- **Answer:** Very useful comments indeed. We particularly appreciate the suggestion of simulating an analysis and provided the data analytic code already at Stage 1. Thanks for that.

**Reviewer – Charlotte Pennington**

---

**4.0.**

This study aims to assess the initial feasibility and efficacy of a tabletop role-playing game in reducing symptom severity of possible gaming disorder, social anxiety and loneliness and increasing self-concept in 20 participants. It is a proof of principle study which will inform a larger confirmatory study of this intervention. An experimental, multi-case study design is used with outcomes assessed over a 10-week period and analyzed using a variety of effect sizes, across a series of complementary analysis steps. The Stage 1 report is well detailed and clear, and I have assessed it in accordance with the PCI RR guidelines as follows below. I want to make it clear here that I am no expert in multi-case designs, so the Recommender may want to include another reviewer for this specific design element. It does not seem that confirmatory analyses relating to NHST/p-values will be used, and it also seems that no power analysis is required because of the pilot nature of the study. The figures are fantastic, aiding understanding and clarity of the design and various elements.

- **Answer:** It is a real pleasure for us to read such comments on our work. We want to warmly thank this reviewer for their overall assessment of the study.

**4.1.**

1A. The scientific validity of the research question(s).

The research questions are informed by an existing evidence base and there is a clear need and rationale for this research. Ethical approval has been granted. One minor point is that without looking at the image in Figure 1, it was not clear to me that this was an offline game (despite you using the term 'offline/real world setting'). Perhaps I am being old school here, but could you refer to this as an offline boardgame? In my mind I was envisaging an online videogame with offline elements, which confused me.

- **Answer:** Thanks for the positive comments (e.g., rationale, ethical approval). The minor point is very useful as we want that a large audience understands well our work. We have thus amended the title of the project to make clear that TTRPG are **offline** games. We also specified this in the revised abstract. With these two changes, we believe that it will be clear to all readers from the onset that we are here using an entirely **offline** game.

**4.2.**

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

This is a pilot study to assess the initial feasibility and efficacy of an intervention based on MMORPGS.

Hypotheses are proposed that are plausible given the background literature. It would be good to know exactly how these hypotheses would be supported, or not, by the data. The analytic plan uses a range of effect size estimates and complementary steps – what will be the precise criteria for stating that the intervention is indeed feasible and efficacious? In what instances may the data be inconclusive?

In addition to this, can your study get at the 'underlying mechanism' for why this treatment may work? Specifically, is it able to assess which component(s) of the game may impact the outcome variables? (The components being: character creation, advancement system, teamwork, heroic fantasy-based world). It appears the data is analyzed at different stages, as the players progress through the different 'modules' of the game – are these game-specific elements going to be specifically assessed, or are you looking at the overall impact of the intervention (i.e., the full game)?

- **Answer:** This comment covers various points. Our answer is thus itemized.

  - **Evidence Threshold:** The opinion of this reviewer echoes some of the concerns raised by the recommender and some of the reviewers (see, e.g., Recommender point 1.1.). After consultation with the full research team, and taking into account the nature of our study, we decided to go for an exploratory pilot. Accordingly, no confirmatory claims will be made at Stage 2, and no threshold will be used to determine potential efficacy of the program. Yet, and as explained in detail to another reviewer (see point 2.1.), the single case design adopted in the current exploratory pilot allows for considering primary and secondary outcomes for each participant.

  - **Underlying Mechanisms and treatment efficacy:** This exploratory pilot also aims to explore whether our program, which is designed to expose the participants to increasingly socially engaging situations in real life, has the potential to affect social skills (e.g., assertiveness) and self-concept (i.e., perceived discrepancy between the ideal and actual selves, see Higgins, 1987). Although no confirmatory statement will be made (see previous point and answers to other reviewers), our rationale is that the program affects some key psychological "processes" or "dimensions" (e.g., assertiveness, discrepancy between ideal versus actual selves), which will ultimately result in reducing symptoms (e.g., social anxiety, GD symptoms). From a process-based and trans-diagnostic perspective, targeting specific psychological processes can contribute to mitigating psychopathological symptoms (Kinderman, 2005; Billieux et al., 2015; 2023). For example, if someone improve their self-concept and social skills, they will less likely "escape" or "avoid" real-life contacts through online gaming. These aspects (and references) are not included in our Stage 1 report, but we plan to develop those points in the discussion at Stage 2. We thank this Reviewer for having raised this point.

References mentioned in this answer (and potentially implementable at Stage 2):


- Hamonniere, T., & Billieux, J. (2023). Individually delivered mindfulness-based cognitive therapy in concomitant problematic substance use and emotional symptoms: A process-based case study. *Clinical Psychology & Psychotherapy*, in press. https://doi.org/10.1002/cpp.2827

- Billieux, J., Philippot, P., Schmid., C., Maurage, P., de Mol, J., & Van der Linden, M. (2015). Is dysfunctional use of the mobile phone a behavioural addiction? Confronting symptom-based versus process-based approaches. Clinical *Psychology and Psychotherapy, 22*, 460-468. https://doi.org/10.1002/cpp.1910

- Kinderman, P. (2005). A psychological model of mental disorder. *Harvard Review of Psychiatry*, 13(4), 206–217. https://doi.org/10.1080/10673220500243349


**4.3.**

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

I am not an expert in “Experimental multiple single-case design” so the Recommender may want to recruit another reviewer who can comment on this specific part. However, I commend the authors for Figure 2 which clarified this for me. This seems like a rigorous step-wedged design.

Page 10 – for lay readers, can you briefly describe what multiple single-case design’ is before outlining its advantages?

What mitigation is in place if you cannot recruit participants meeting inclusion criterion 6: “endorsing at least one criterion on the Internet Gaming Disorder Test (IGDT-10; Király et al., 2017) assessing gaming disorder symptoms; and (7) having a score ≥ 56 (threshold for subclinical social anxiety) but ≤ 96 (threshold for clinical social anxiety) on the Liebowitz Social Anxiety Scale (LSAS; Liebowitz, 1987) assessing social anxiety symptoms”.

What do the four different groups refer to exactly? I think these might refer to the different timepoints at which the participants enroll into each module/stage of the intervention? But it would be good to explicitly state this to avoid confusion.

I do not think a power analysis is necessary given that confirmatory analyses are not being conducted but would appreciate a response to this to make sure (see my above point regarding my expertise on multiple single-case designs).

What happens if participants drop out of the study at different stages? Will you recruit additional participants to make the final target sample size of 20 across the four groups? Can the same analyses be conducted if the groups are unequal?

It would be good to know the explicit rationale for 20 participants across 4 groups.

- **Answer:** This comment covers various points. Our answer is thus itemized.

  - **Brief description of multiple single-case methodology.** As requested, a basic definition of the methodology used has been added in the revised introduction. It reads as follows:

    *"A single-case design is an evaluation method that can be used to rigorously test the success of an intervention on a particular case (i.e., a specific participant). An extension of this evaluation method is the multiple single-case approach used in the current study, in which several (instead of one) cases are considered to highlight potential differences and similarities between them (e.g., factors influencing dropout, effect of the program on primary/secondary outcomes). Evidence arising from multiple-case studies is generally considered as stronger and more reliable than from single-case designs (Baxter & Jack, 2008).".*

    *Baxter, P., & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. The Qualitative Report, 13(4), 544-559. https://doi.org/10.46743/2160-3715/2008.1573*

  - **Potential recruitment problem.** The recruitment phase (eligibility) started end of February and is planned to finish by the end of March. The start of data collection is planned for mid-April. The eligibility screening has started and will finish late March or early April. In the event that we cannot reach the required number of participants (i.e., 20 participants), several options could be considered including:
    - Conducting the study on 3 groups instead 4 groups of participants (which would be a sufficient sample for conducting a robust multiple single-case design).
    - In case of difficulties securing simultaneously 4 groups of participants and to avoid any delay or inconvenience for those who volunteered and are eligible for the study, a non-concurrent multiple baseline procedure can be adopted rather than a concurrent multiple baselines.
    - Slightly diminishing the cut-off used for social anxiety.

  - **What the four groups of participants refer to.** As explained in the introduction and in Figure 1, the offline TTRPG program implies playing in small groups of 5 participants plus a game master. So, basically, our design implies that four different groups of five participants undergo the entire intervention (the 10

sessions and 3 modules). The same game master will manage the four groups, as clearly explained in the manuscript.

- **Power Analysis.** No power analysis is required to determine the number of participants in a multiple single-case design.

- **Dropout.** In the revised text, we added a section of what constitutes a dropout in our study and how we will handle them. Participants who drop out will not be replaced. Dropouts do not impact the statistical analyses, as the statistical analyses are performed per participant (not per group of participants). This section reads as follow:

  *"The number of participants was determined by taking into account the expected dropout rate and the requirement to provide enough inter-subject replication of the experimental effect. TTRPGs are well suited to groups of 3 to 5 players plus one game master. We opted for the upper limit (5 participants per group) to compensate for potential dropout(s). We decided that the minimal number of participants required to continue playing should be 3 to guarantee sufficient social exposure. If more than 2 participants drop out in the same group, the remaining participants will be allocated (if possible) to another group. In this pilot study, dropout occurs when a participant leaves the program permanently, regardless of the number of session(s) completed. Participants who miss a session for acceptable reasons (e.g., being sick) will have the possibility to reintegrate and continue the program (the number of potentially missed session(s) will be recorded for each participant)."*.

- **Justification regarding the number of participants (and number of participants per group).** We wanted to include as many cases as possible in our study. We came to the conclusion that four groups is the maximum number of groups that our gamer master will be able to handle in relation to the duration of our research grant (see Author contributions for more information).

  TTRPGs are well adapted for groups of 3 to 5 players plus one game master. We decided to opt for the upper limit (5 participants per group) to compensate for potential dropout(s). We decided that the minimal number of participants required to continue playing should be 3 to guarantee sufficient social exposition. If more than 2 participants drop out of a same group, the remaining participants will be allocated (if possible) to another group.

  The following section was added in the revised text:

  *"TTRPGs are well suited to groups of 3 to 5 players plus one game master. We opted for the upper limit (5 participants per group) to compensate for potential dropout(s). We decided that the minimal number of participants required to continue playing should be 3 to guarantee sufficient social exposure. If more than*

*2 participants drop out in the same group, the remaining participants will be allocated (if possible) to another group. In this pilot study, dropout occurs when a participant leaves the program permanently, regardless of the number of session(s) completed. Participants who miss a session for acceptable reasons (e.g., being sick) will have the possibility to reintegrate and continue the program (the number of potentially missed session(s) will be recorded for each participant).".*

**4.4.**

| |
|---|
| 1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.<br><br>Yes, the methodology is detailed and sufficient. There is an explicit link to the code and data on the OSF. Diagrams are included to aid the reader's understanding. |

- **Answer:** Thanks for the positive comment.

**4.5.**

| |
|---|
| 1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g., absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).<br><br>This is the greatest thing, in my opinion – although participants undergo this intervention at different stages in the four groups, is a control group required who play a different/neutral game where the game elements expected to drive effects (e.g., teamwork etc.) are not present. Perhaps the design mitigates against this need?<br><br>It may be worth including a question on engagement or enjoyment as an attention check.<br><br>It would be good to add an attention check within one of the questionnaires, e.g., within the loneliness questionnaire, an additional question could be added which simply states: "for this question, select the option X" (with X being one of the response options used in the questionnaire). |

- **Answer:** This comment covers various points. Our answer is thus itemized.

  **Control group.** No control condition is required because in our multiple single case design, each participant is their own control, and this is made possible by multiplying the assessment points for each single participant (Tate et al., 2015). To determine whether our program affects target variables for specific individuals, we relied on a combination

of multiple baselines across participants and multiple assessments per participant (see Figure 2).

**Engagement/Enjoyment.** Two qualitative feedbacks will be collected (session 3 and final session 10). Aspects such as engagement and enjoyment will be collected at that occasion.

**Attention check.** Thanks for the advice. We will add an attention item to each psychological assessment conducted. Attention items will also be added to the material made available on Open Science Framework.

**4.6.**

| |
|---|
| I would omit the specific name of the hospital from the Introduction; it doesn't add anything but may provide too much detail. |

- **Answer:** We have removed the name of the specific hospital in the introduction.

**4.7.**

| |
|---|
| Is 'race' a common terminology used for the defined categories ("race (e.g., human, elf, orc"))? |

- **Answer:** This is indeed common terminology in "Medfan" universe (such as the one of Dungeons and Dragons and World of Warcraft). The Game Master will present the world of Dungeon and Dragons during the first session of the program, to clarify what means such terminologies in this context and avoid potential misunderstandings or bad vibes.

**4.0.**

| |
|---|
| Overall, this is a fantastic Stage 1 submission with rigor and detail. I recommend minor revisions to aid further clarity on design and analysis elements. |

- **Answer:** We would like to warmly thank this reviewer for this very reinforcing comment. Furthermore, the other comments made were also very useful and appreciated.

**Statistical expert – Zoltan Dienes**

---

**Instruction received by the Recommender (email dated February 13, 2023):**

Please find below additional notes from Zoltan Dienes, our statistics expert. I asked if he could double check because we didn't have time to seek one more reviewer to provide specific feedback on the multiple single case design. Again, note that Zoltan's comments concern primarily hypothesis testing; the exploratory approach would be more flexible. Because Zoltan's feedback was unofficial, you don't have to formally respond to it. But I encourage taking it into consideration carefully.

**5.1.**

> I think the analyses need considerably more specification. As you say they are hypothesis testing – so the inferential basis of confirming or refuting hypotheses needs to be clear. They use significance testing with the "reliable change" method, which when I checked was just a t-test. They are concerned about effect sizes being clinically relevant. As you say they must then specify what is clinically relevant – and intuitively illustrate that relevance by indicating what that means with the raw dependent variable. Some sort of "inference by intervals" may best suit their inferential concerns. As they are significance testing, they are in the business of error control. How will they deal with familywise error rate over several DVs – and 20 subjects? Under what conditions will they assert the hypotheses given in the first column of the design table – which are worded as generalities as if they applied to a population of subjects – as refuted or confirmed? If they mean their hypothesis to apply to a population of subjects, how will they generalize to the population? If they do not mean it to apply to a population, what is their claim precisely? What claim are they testing? They need to justify why the claim they have in mind is best tested by a multiple case design. Why forgo the claims that can be justified by conventional by-subject analyses? If they want to say that the treatment effect varies by subject, they need to explicitly test the variability over subjects. As it stands, the claim they are testing and the inferential basis for confirming or refuting it has not been locked down nearly tightly enough.

- **Answer:** The opinion of this statistical expert echoes the main concerns raised by the recommender and the reviewers (see, e.g., Recommender point 1.1.). After consultation with the full research team, and taking into account the nature of our study, we decided to go for an exploratory pilot. Accordingly, no confirmatory claims will be made at Stage 2, and no threshold will be used to determine potential efficacy of the program. As we reframed the study as an exploratory pilot, we removed the formal hypotheses (p.8). Terms such as "efficacy" were replaced by "effect". Sentences are now formulated in an exploratory way.

**Additional changes (not requested by the Recommender or Reviewers)**

---

**6.0. The team of authors also decided to apply the following changes during related to the recruitment process.**

- One inclusion criterion has been slightly modified. In the initial protocol, only MMORPG players were considered eligible. We decided to:

  1. Include people playing online MMORPGs and/or RPGs. The criterion has been modified as follows: “being a current MMORPG or online RPG player”.

  2. Include participants with an extensive experience of playing MMORPGs or RPGs even if they are currently involved in playing other types of videogames (e.g., multiplayer online battle arena).

  Those changes have been conducted as we started to advertise the study and realized that some people wanted to be included in the study but (a) played online RPGs such as Diablo, or (2) were familiar with online RPGs/MMORPGs but are currently involved in other types of videogames. The research team decided that such potential participants should not be excluded from the study. The modification of this inclusion criteria implied that some corrections have been made in the revised manuscript (title, abstract, and main text). All changes are highlighted in yellow.

  In addition, the following section has been added in the methods:

  *“Participants playing online RPGs (e.g., Borderlands, Diablo, Final Fantasy) – which do not technically qualify as "massive" multiplayer because they involve fewer players – were also considered eligible as those games share most features of MMORPGs (e.g., advancement mechanics, interactions between players). Furthermore, it was decided that participants with an extensive experience of playing MMORPGs or RPGs are also eligible for the study even if they are currently involved in playing other types of videogames (e.g., multiplayer online battle arena).”*

- We added a specification regarding availably of the participants in the following sentences: “The first 20 participants who complete the online survey, agree to participate in the experiment, are available to intend and play at the time proposed by the research team, and meet the inclusion criteria will be invited to take part in the study and will be distributed into 4 groups (see Figure 2).”.

- We decided to add **Iliyana Georgieva** as a new co-author. Iliyana is a master student contributing to the project, who was not yet enrolled in the study when we submitted the Stage 1 Registered Report. The contribution section has been updated to account for the important role of Iliyana in the study.

- The Author contributions section has been updated. We realized that several aspects were not anticipated at the time the protocol was submitted (in relation to the recruitment process, especially). We anticipate a new update at Stage 2.