Managing Board of PCI Registered Reports

Dear Managing Board of PCI Registered Reports,

We were pleased to receive the review of our Stage 2 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 November 2, 2024, and your willingness to consider a revised version of our Stage 2 Registered Report. We are grateful for the comprehensiveness of the review process and thank all peers involved for their insightful comments that helped improve our manuscript.

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.

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

Dear Joël Billieux and co-authors,

Thank you for your detailed and informative Stage 2 manuscript. We were lucky to have all three reviewers return to assess the work. In general, the reviewers are impressed but also see a few issues that we must address. Below I spell out the key points that I'm expecting to be addressed, overlapping with some (but not all) reviewer feedback.

Answer: We are very grateful and warmly thank the reviewers for their time and willingness to review our Stage 2 Registered Report. Their comments were helpful and contributed to improving our manuscript.

1. At Stage 1, we discussed three options: a) set criteria to confirm effectiveness, b) set criteria to confirm feasibility, c) not set any criteria and make the study fully exploratory. We agreed with the last option. As two reviewers remind us, this means that the study cannot, at any place, communicate the intervention to be effective and/or feasible; such claims always depend on selected criteria and should they be made, the criteria and their justifications need to be mutually agreed at Stage 1. Therefore, I kindly ask you to remove words and phrases that claim feasibility or effectiveness, including terms like "high" that are a matter of subjective evaluation. The study is very strong without such claims; a detailed exploration like this doesn't need to speculate about success. The readers can decide themselves how effective and/or feasible the reported results appear to them. I will carefully re-review all words/phrases on the next round and if debatable sections remain, we can discuss them separately one by one. [To further clarify: it's different to register analyses vs interpretations; this paper did the former but not the latter – and it's perfectly ok!]

Answer: We would like to thank you and the reviewers for pointing out this issue. In accordance with this important point, we revised the manuscript to systematically remove (both in the results and discussion sections) words and phrases that claim feasibility or effectiveness.

Examples of changes made (all highlighted in the main text):

- **The sentence "The participation rate in the TTRPG program was high, as reflected by the number of sessions attended by the eighteen participants who completed the whole intervention." was transformed into "Among the eighteen completers, most (10/18) participated in all 10 sessions, while some (6/18) missed a single session."**
- **"Only two participants dropped out" was replaced by "Two participants dropped out".**
- **"Again, completion rates were very high (see Figure 3 for details)" was replaced by "Completion rates are reported in Figure 3".**
- **"Only one participant completed the 10 weekly psychometric assessments but not the three-month follow-up "was replaced by "One participant completed the 10 weekly psychometric assessments but not the three-month follow-up".**
- **In the first paragraph of the discussion, we replaced "These results are encouraging not only because they support the feasibility of a TTRPG-based intervention program but also because they suggest that this type of intervention can reduce treatment reluctance in this population." by "The results observed are encouraging and suggest that TTRPG-based intervention programs may be used to reduce social anxiety and gaming disorder symptoms.".**

Substantial rewording was also done in the "Abstract" and "Discussion" sections.

For example, this was taken from the previous initial paragraph of the discussion:

"*Evidence of feasibility was supported by a low rate of dropout (18 out of the 20 participants completed the 10-week intervention), high involvement and participation in the study (two of the 18 completers*

***missed more than one session, and the majority participated in all 10 TTRPG sessions), very high completion rate of weekly psychometric assessment (all baseline measurement points and weekly assessments were completed, and only one participant did not complete the three-month follow-up), and a global high attainment of the progressively more challenging objectives within the three consecutive modules of the TTRPG-program.*"**.

After our revisions, this paragraph reads as follows:

"***In terms of feasibility, we observed that most participants completed the program (18 out of the 20 participants completed the 10-week intervention) and were involved in terms of participation (among the 18 completers, 10 missed no session, 6 missed a single session, and 2 missed more than a single session) and weekly psychometric assessment (all baseline measurement points and weekly assessments were completed, and one participant did not complete the three-month follow-up). Moreover, participants were largely able to attain the progressively more difficult objectives implemented in the TTRPG program.*"**.**

We invite readers to refer to the revised manuscript for all relevant reformulations.

2. As reviewers note, there's an unusually large number of changes to Stage 1 content. I tried to double check all of them, and my impression is that they don't meaningfully change planned content and are thus acceptable. However, in the future (especially if you have other recommenders), such changes might not be accepted (e.g., when the original comparison to "medical interventions" has been changed to "psychological treatments", I don't see this problematic but others might consider it reframing the work). Please consider this in your next RR.

Answer: We can only agree on this point, and we would like to thank the recommender for their flexibility on this issue. This is mainly due to the fact we did not perform the professional English proofreading at stage 1, as transparently disclosed during the stage 1 revision process:

"***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." (Cited from Stage 1 Author's response of March 23, 2023, available from PCI RR website: https://rr.peercommunityin.org/download/t_recommendations.reply_pdf.afcd31e5cdd32aca.44445f52525f53746167655f315f46696e616c5f616e737765722e706466.pdf).*"**.**

To avoid complicating the work of the recommender and Reviewers at stage 2, we initially decided not to highlight nor "track changes" the grammatical and formal modifications in the Stage 2 manuscript. However, we were not allowed to do so, and we had to provide a file with all changes (including formal/grammatical corrections) reported in "track changes" mode.

All changes not strictly related to formal/grammatical improvements were disclosed in the cover letter accompanying the Stage 2 submission (see: https://osf.io/y4pez). We proceeded this way to facilitate the review process, and we regret that the "tracked changes" file has complicated the task for some Reviewers.

3. A major related point concerns effect sizes. At Stage 2, the paper has introduced new benchmarks into a Stage 1 section; these new labels are linguistically very favourable (already 0.61 = large). Even though the footnote reports these benchmarks as unplanned, this has huge impact by framing results to a positive light post hoc, especially because readers of the final article cannot see the originals (without separately seeking Stage 1). Because PCI RR doesn't generally support any standardized benchmarks but rather encourages authors to think and justify how and why certain effects are meaningful (or meaningless), I propose a compromise to remove all benchmark labels and simply report the effects via some clear and neutral way. As the cited Vannest and Ninci say: "For those engaging in research, a benchmark is only useful when a body of literature has been meta-analyzed for the typical results… Benchmarks are not as useful as direct

interpretations of the change in relationship to factors related to client needs, context, and prior intervention work."

Answer: We agree with the limitations associated with benchmarks. Consequently, we removed all references to benchmarks in the manuscript in the following sections:

- **Data analytic strategy (in the "Methods" section)**
- **Efficacy of the TTRPG program (in the "Results" section)**
- **Discussion**

Please note that we have also modified Table 2 as well as the notes below Table 4 accordingly. In addition, Vannest and Ninci's (2015) reference has been removed.

This is a representative example of benchmarks being removed from the revised results section:

"*The results obtained using non-overlap Tau-U indices between the baseline (A) and the intervention (B) phases indicated a large inter-individual heterogeneity, with symptoms decreasing to various degrees among participants (see Table 2). Avoidance of social situations symptoms decreased to various degrees in sixteen participants (88.89% of the sample) of whom thirteen showed a TAU-U between 1.00 and .60. Fear of social situations symptoms decreased in seventeen (94.44%) participants of whom ten showed a TAU-U between 1.00 and .60. Gaming disorders symptoms decreased in fifteen (83.33%) participants of whom five showed a TAU-U between 1.00 and .60. Time spent gaming decreased in ten (participants (55.56%) of whom only two showed a TAU-U between 1.00 and .60. Perceived loneliness decreased in eleven (61.1%) participants of whom six showed a TAU-U between 1.00 and .60.*".

4. Following the above quote, I'd personally like to see in more detail *what* the reported changes were, beyond averages. For example, Nuyens et al. (2023) have recently noted that IGDT-10 conflates core and peripheral criteria and is not well positioned to distinguish between problematic and non-problematic gaming. This isn't a problem for the pilot, but for a scientifically curious reader (like myself -- I'm genuinely excited about the intervention mechanisms), what matters are the very symptoms that changed. As is well known in the field, e.g. #1 ("When you were not playing, how often have you fantasized about gaming, thought of previous gaming sessions, and/or anticipated the next game?") does not refer to any problems, and #8 ("Have you played to relieve a negative mood [for instance helplessness, guilt, or anxiety]?") can be fully met by just gaming to relax when tired. I find it problematic to speak of "recovery" in cases where such experiences are absent. On the other hand, reporting e.g. #9 (Have you risked or lost a significant relationship because of gaming?) would be highly valuable effectiveness information. A matrix with specific symptom distribution changes would be informative and help us understand where the effects come from and further specify the possible mechanisms (this is why we do pilots, after all).

Answer:

4A: Symptom-based analyses

We agree that a symptom-based analysis would have been relevant, especially given that some symptoms included in the DSM-5 framework are considered "core features" (indicative of a problem) while others are considered "peripheral features" (not necessarily indicative of a problem). However, these analyses were not pre-registered as our design was not optimal for testing those effects. Indeed, providing single-case analyses on these symptoms considered individually is compromised because (1) we have only 1 item per symptom, and (2) the range of the response scale used in the IGDT-10 is limited (response can be 0, 1, or 2). Consequently, variance is very limited. Moreover, we further reduced this variance by excluding persons with too high levels of problematic gaming (see "Participants" section). However, we added to the OSF repository a table reporting the proportion of each IGDT-10 item endorsed among participant before the intervention, post-intervention, and at the follow-up: https://osf.io/2g658. Future studies conducted in clinical people should indeed adopt a design allowing for symptom-based analysis (e.g., by using several items for each gaming disorder symptom).

In response to this comment, the following footnote was added:

"*Following a request made by the PCI RR recommender during Stage 2, we added to the OSF repository a table reporting the proportion of IGDT-10 items endorsed by each participant before the intervention, after the intervention, and at follow-up: https://osf.io/2q658*.".

4B: We agree that the terms "improvement" or "recovery" might not necessarily be relevant in studies conducted in sub-clinical populations. In the current study, the terms "recovery" and "improvement" are used to adhere with the labels proposed by Jacobson & Truax (1991) to define clinically relevant changes in psychotherapy research, even if our study is conducted with non-clinical participants. Although we mentioned it clearly and transparently in the "Data analytic strategy" section of the Stage 2 paper, we now specified it both in the "Results" and "Discussion" sections. These are examples taken from the revised manuscript:

"Secondary outcomes considered included assertiveness, self-concept, and perceived loneliness. As a reminder, the terms "recovered", "deteriorated", and "improved" are used to adhere to the labels proposed by Jacobson and Truax (1991) to define clinically relevant changes in psychotherapy research, even if our study was conducted with non-clinical participants." (from the "Results" section)

"Of note, the terms "recovery" and "improvement" are used to adhere with the labels proposed by Jacobson and Truax (1991) to define clinically relevant changes in psychotherapy research, even if our study was conducted with non-clinical participants." (from the "Discussion" section)

5. The above could apply to other measures as well. For example: if the intervention positively contributes to reduced social gaming, it would make perfect sense that some participants increase their loneliness scores (losing contact to close online friends). Such hypothesis too might benefit from item-specific analysis of loneliness scores. As another example, I checked LSAS and found items like these to be rated for avoidance: "Participating in a small group activity", "Talking face to face with someone you don't know very well", "Meeting strangers", "Entering a room when others are already seated", "Speaking up at a meeting", etc. Factually, anyone who participated in the intervention didn't avoid such situations; the scores should drop in these items with any (control) activity that involves such engagements. For scientists interested in the mechanism, imo the most valuable information would be the scoring distribution, allowing us to learn about the possible transfer network beyond the obvious. I won't address other measures but would encourage taking a look into their operationalisations too.

Answer: We can only agree with the relevance of this comment. Although we decided not to add additional non-preregistered analyses at Stage 2 (beyond those reported above, see Recommender, point 4), we now considered in the discussion the fact that committing to the program *per se* can influence the assessment of social situation avoidance. In our future work, we will consider designing our studies to allow symptom-based analyses.

The related section, in the revised discussion, reads as follows:

"*Regarding social anxiety symptoms, fear of social situations was reduced for almost all participants (17/18 cases), which was also the case for the avoidance of social situations (16/18 cases). The magnitude of decrease was the most pronounced for the avoidance of social situations during the intervention compared to the baseline phase, according to the BC-SMD analyses. It cannot be ruled out that the result regarding avoidance of social situations has been at least partly influenced by the fact the participants might have considered the TTRPG sessions* per se *as an exposure to social situations.*".

6. I really like the supplement case report. It's great to see, on an individual level, what changed and didn't change, in what context. It's unfortunate that it wasn't possible to triangulate it with any qualitative data that could've helped better understand reporting. We've found it highly informative to add open questions after symptom items to understand participants' reasoning, this can be easy to implement for interventions too. For outliers, I'd encourage some conventional negative case analysis to better understand why they occurred. I see some of this has already been done informally (in discussion); of course it's also possible that there's nothing more to find, in which case you can report that explicitly.

Answer: **We agree with the recommender that it would have been highly relevant and interesting to triangulate our quantitative results with qualitative data (about the intervention) obtained from participants. Unfortunately, it was finally not possible to do it in the context of this pilot. This was transparently acknowledged in a footnote in our Stage 2 paper:**

"*It was initially planned to collect high-quality qualitative data during this final session. Unfortunately, this was not possible for practical reasons. Indeed, there was not enough time in session 10 to implement this data collection. Instead, the game master conducted informal feedback during the debriefing of the last session; however, data were not collected in a structured way nor recorded, hindering the possibility of conducting qualitative analyses. To remedy this issue, we implemented a qualitative question in the last online assessment; however, most participants either did not answer or provided very short, non-comprehensive answers. These data were thus not analyzed nor stored in the Open Science Framework.*".

Our plan is to collect qualitative data in the clinical pilot (i.e., the second step of the current research program) that is planned for 2025 in a group of people with clinically relevant gaming disorder.

7. Minor things:
a. One reviewer comments about the changes not being visible. You can ignore that comment; for me the tracked changes were clear and the managing board (Chris) specifically asked not to add extra marks to the main document.

Answer: Well-noted. In the revised version of Stage 2, we highlighted the changes to facilitate the reviewer process (this was agreed beforehand in an email exchange with the Recommender on the 1st of December, 2024).

b. At the end of Stage 1, we discussed separately about final technical edits after agreeing that no hypotheses or feasiblity are tested. I notice that in a few places "hypothesis" remains, especially p. 9 and it has also been added to conclusions. Please remove these to protect readers from misinterpreting the work as hypothesis testing.

Answer: The revised Stage 2 register report no longer refers to "hypotheses". We also modified that in the study design table (available from the OSF: https://osf.io/b965v).

I hope you find my above notes and the authors' feedback sufficiently clear and useful; the study operates with numerous critical small details so I wouldn't be surprised if there are some misunderstandings/errors involved in our readings. As earlier, you can contact me directly to consult with specific questions, to streamline the next review round. However, I will invite all reviewers back for a second review and am naturally going to consider their views carefully as well.

Answer: It was a real challenge to conduct this highly complex study as a registered pilot while simultaneously using the PCI RR platform for the first time. It was also our first multiple single-case study with so many participants and the first time we created and tested a TTRPG-based program. So, the study was extremely demanding at many levels, and we humbly admit that we are satisfied by the way we managed to handle it. We really did our best to be totally transparent and to follow all the canons of open science the best we could. For those reasons, we were especially grateful to the two Reviewers who recognized our efforts and merits.

Again: this is a highly interesting pilot that, over some years of further development, could become useful in practice. The present work is just the very first step -- let's ensure that we don't hype the results but focus on what was learned, what can be improved, and how to next proceed optimally toward formal effectiveness testing.

Answer: We got this and carefully revised the manuscript to avoid hyping or over-stating the results (see, e.g., our answer to point 1).

Reviewer – Matti Vuorre

I have read the Stage 2 RR "Can playing Dungeons and Dragons be good for you?" and provide my observations and evaluation below. Overall it seems the authors have conducted the study as described in the Stage 1, but I am less satisfied by some aspects described below.

Answer: **We indeed did the maximum to follow every single step described at Stage 1. However, the study design and procedures of the present study were complex, and we couldn't avoid some deviations. We have transparently disclosed every single deviation made in footnotes and explained them in the cover letter submitted with the stage 2 report (this letter is now available from the OSF). We sincerely hope that this Reviewer will be more satisfied by the additional changes made in the Stage 2 report**.

Some of the comments made by this Reviewer made us realize that they were perhaps unaware of the letter that accompanied our submission, which disclosed very comprehensively and transparently all the changes made (see also the other Reviewers' feedback on this matter). We think this letter was crucial to facilitate the review process (and we are thus very sorry if this Reviewer was unaware of this important document). This letter is for us an essential piece of this exploratory registered pilot, and we decided to upload it on the OSF repository: https://osf.io/y4pez.

One of my main comments regarding the Stage 1 manuscript was that the study was framed to provide both a feasibility test and a test of an "initial" effect. This framing was not improved upon in later Stage 1 submissions, and I find this Stage 2 report to suffer from the same problem. Calling this a "registered exploratory pilot" that tests an "initial effect" is in my view not justified because the concrete differences between a "pilot" and an "initial effect" from a non-pilot and non-initial effect are not sufficiently clear. The motivation to provide comprehensive hypothesis tests of effects, instead, is clear throughout, e.g. in the conclusion that this is "the first quantitative study of the therapeutical use of TTRPG that endorses all the canons of open science, from pre-registration of design and hypotheses to open data and material."

Answer: In accordance with this important point, we revised the manuscript to systematically remove words and phrases that claim feasibility or effectiveness. We also carefully revised the main text (and the study design table) to avoid referring to hypotheses.

Moreover, none of the tests of feasibility were specified in advance: In a version of the Stage 1 manuscript ("Current study"), I read that "Against this background, the current study proposes an exploratory pilot experiment that aims to test the feasibility (e.g., number of dropouts, ability of the participants to understand and engage in a tabletop role-playing game, ability of the participants to complete regularly the online assessment)". This sentence does not appear in the Stage 2 report, which instead states that "In terms of feasibility, we were interested in how many sessions participants would miss and how many participants would drop out from the program entirely, the ability of the participants to complete the weekly online psychometric assessment, and the ability of the participants to understand and engage in a tabletop role-playing game as well as to succeed in the various objectives of increasing difficulty implemented in the TTRPG program." It is not clear to me if those are exactly the same aims? And, it is not specified what would e.g. count as a low dropout rate; yet 10% of participants dropped out with 10 out of 18 remaining participants completing all sessions, and this was decided to count as "low dropout". I don't necessarily disagree but what would have counted as high dropout? These issues seem to me to go against the idea of (pre-)registering a feasibility study.
Consequently, I am not able to fully evaluate to what extent the feasibility results support the feasibility of the study protocol. It seems fine, but since no cutoffs for "fine" were reported in advance I don't think it is fair to describe those analyses as (pre-)registered.
Similar issues apply to results regarding "initial" effects: For example, what proportions of very large, large, etc. effects would support each of the hypotheses? Authors state in the design table that "Our study is not testing a specific theory or model" and "We will not provide "general" interpretation (unless in the unlikely case where all participants present with the same pattern of results)." This is fine, but none of the results should

then be described as tests of hypotheses regarding some (general) effect. It seems unfair that the interpretation of the presented results will inevitably be colored by the advertised registration of this study, even though it doesn't extend to these "hypothesis tests".

Answer: As said, the study was very complex. Regarding feasibility, we simply realized during the study that we were not specific enough at Stage 1 about how feasibility will be assessed. This is why we were totally transparent regarding these matters when we submitted the Stage 2 paper.

This was stated transparently in the letter joint to the submission of the Stage 2 paper, as follows:

"*We initially planned – mainly based on the reviewer's comments at Stage 1 – to conduct exploratory non-pre-registered analyses using the game master's assessments of objective achievement. Yet, we finally abandoned this idea and only used these assessments to support the feasibility of our TTRPG program. The proportion of achieved, partially achieved, and non-achieved objectives are reported in Table 1, and the raw data are available from the following link: https://osf.io/65w3j.*

In this study, we preferred considering objective achievement assessed by the game master as an indicator of feasibility. Indeed, it was deemed important that a large proportion of participants successfully attained the progressively more difficult objectives implemented in the three modules composing the TTRPG program, both from a motivational (i.e., self-efficacy enhancement and positive reinforcement) and engagement (i.e., participant's commitment is a good indicator of treatment adherence / therapeutic alliance) perspective.

***In the Stage 2 Registered Report, these aspects are now reported in detail, and we added a footnote to make this clear and transparent.*"**.

Based on the concerns raised by this Reviewer and the Recommender, we revised the manuscript to systematically remove words and phrases that claim feasibility or effectiveness (see the examples provided in response to comment 1 of the Recommender). We also carefully revised the main text to avoid referring to hypotheses.

- Stage 1 read "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)." but Stage 2 does not have this sentence and instead refers to a three-week baseline measurement. Those do not seem the same but it is possible that I do not understand this correctly; a clarification for the deviation would be appreciated.

Answer: This issue was transparently discussed in the letter joint to the submission of the stage 2 paper, as follows:

"*We initially planned to have baselines of three to six measurement points (i.e., different numbers of measurement points for each group), with the onset of the experimental phase randomly determined for each group (Group 1 starts after a three-week baseline, Group 2 starts after a four-week baseline, etc.). This methodological approach corresponds to international standards for defining an experimental single-case methodology with enough data points per phase. However, we had to adjust our design to guarantee the feasibility of the study. The reason is that the recruitment process took more time than anticipated, and it was crucial for us to finish the 10-week intervention phase before the beginning of the holiday season in Switzerland (early July) to avoid massive dropouts at the end of the study. Accordingly, we decided to apply a similar 3-week baseline for each group, implying that our design became quasi-experimental instead of experimental (Tate et al., 2015).*".

We also transparently mentioned this issue in a footnote in the Stage 2 registered report.

- The statement "In this sense, our study confirms the soundness of single-case analysis to test treatment efficacy, as a more traditional group approach would only have emphasized the global (i.e., group-based)

small or small to medium effect on symptoms, masking the important heterogeneity in the response to intervention." is incorrect and should be removed. There is no statistical quantification of heterogeneity, but simply a casual comparison of significant to not significant. A "more traditional group approach" using e.g. multilevel models would allow such quantification.

Answer: The sentence pointed out by this Reviewer was corrected. The revised sentence reads as follows:

"*In this sense, our study confirms the soundness of single-case analysis to test treatment efficacy, which in the current case allowed us to emphasize the important heterogeneity of the participants' responses to the intervention.*".

- In the design table authors say "We will not provide "general" interpretation (unless in the unlikely case where all participants present with the same pattern of results)." Yet they draw conclusions such as "The current study demonstrated that a 10-week structured TTRPG-based intervention is feasible and effective in reducing symptoms in a sample of sub-clinical socially anxious gamers." Any general conclusions such as that should be removed as per authors pre-specified plan.

Answer: We revised the manuscript to systematically remove words and phrases that claim feasibility or effectiveness (see the examples provided in response to comment 1 of the Recommender). We also carefully revised the main text to avoid referring to hypotheses.

- I found the Stage 2 report evaluation cumbersome because authors did not submit the required (or I did not find) manuscript with highlighted changes from Stage 1. I can tell that some language has changed substantially. I have closely read the submission and it seems that these changes are not to e.g. theoretical rationale, but it is harder to tell lacking the document with highlighted changes.

Answer: To avoid complicating the work of the recommender and Reviewers at Stage 2, we initially decided not to highlight nor to "track changes" of the grammatical and formal modifications in the Stage 2 manuscript. However, we were not allowed to do so, and we had to provide a file with all changes (including formal/grammatical corrections) reported in "track changes" mode.

All changes not strictly related to formal/grammatical improvements were disclosed in the cover letter accompanying the Stage 2 submission. We proceeded this way to facilitate the review process, and we sincerely regret that the "track changes" file has complicated the evaluation of this Reviewer. In fact, we think, based on the feedback provided by this Reviewer, that they unfortunately did not have access to the cover letter mentioning and justifying all the changes. For this reason, the letter is now available from the Open Science Framework.

- For example, the write-up of the data analytic strategy section has changed a lot from Stage 1 to Stage 2 rendering its evaluation unnecessarily difficult.

Answer: The modifications made to the "Data analytic strategy" section have been disclosed and justified in the covering letter accompanying the Stage 2 submission. We regret that the evaluation of Stage 2 was complicated for this Reviewer. We really tried to facilitate the review process to the maximum by adopting a fully transparent approach and detailing all the changes made.

- Please include figures and tables where they belong in the manuscript. I will not review future submissions with floats at the end of the document.

Answer: Figures and tables are now included in the manuscript. The study design has been removed from the main text and is available from the OSF.

Respectfully signed

Reviewer – Matúš Adamkovič

Dear Authors,

I was very pleased to read this Stage 2 report. Honestly, this is perhaps the most detailed and in-depth pilot/feasibility study I've ever read. The level of transparency and reporting practices is outstanding. I compared the Stage 1 protocol with the Stage 2 report, and all the deviations I could detect (and even a few more) were disclosed. The reasons for these deviations were understandable, and none of them altered the study design in a way that would make me question adherence to the preregistered protocol or the validity of the piloting procedure.

Answer: We have no words to qualify how this comment was pleasant to us. We warmly thank this Reviewer for recognizing our merits.

I have some minor suggestions to improve the paper (especially the discussion section), but frankly, I wouldn't mind if the paper was published in its current version. My suggestions are as follows:

- Please consider moving the footnotes regarding the deviations into one paragraph in the main text (perhaps merging them with the current paragraph on study limitations, as it already covers some of the deviations), so readers can easily see what was changed and why.

Answer: We have considered this and tried various options. Yet, we finally decided to keep the footnotes approach. First, by doing so, we comply with the PCI RR guidelines. Second, using a single paragraph to mention all deviations complicates the global understanding and readability of the paper (because some footnotes concern the participants, some the methods, some the instruments, etc.). We also assume that footnotes report elements that are not mandatory for understanding the study, but that are available for interested readers. In contrast, the limitations paragraph reports essential deviations that warranted being reported in the main text rather than in footnotes.

- The program had a very low attrition rate, which is very promising. The authors attribute the high engagement to (the appeal of) TTRPGs. However, I would appreciate a short discussion on what other factors (e.g., rapport with the game master, group dynamics, conscientiousness, etc.) might have contributed to the high engagement, and how these could be investigated and potentially accounted for in future studies.

Answer: This is a very relevant and justified remark. We added these elements in the "Discussion" section of the Stage 2 registered report, as suggested:

"*A potential explanation for these very encouraging results is that TTRPGs share many features with video games (especially MMORPGs and online RPGs) and similarly allow for fulfilling basic individual needs such as relatedness, autonomy, and competence. Of course, we cannot exclude the possibility that other aspects (e.g., financial compensation, relationship of trust with the game master, specific group dynamics, conscientiousness of participants) may have influenced the motivation and involvement of participants. The observed results are encouraging and suggest that TTRPG-based intervention programs may reduce social anxiety and gaming disorder symptoms.*".

- I would also suggest addressing an implementation issue–in my reading, the success of the program might have largely depended on having a highly skilled game master. Could the authors reflect on this and suggest potential solutions that could be applicable to a broad range of clinicians and other health care professionals?

Answer: That is a very relevant point. Yet, given the pilot nature of the study, we respectfully think this is a bit premature to discuss these aspects in the present manuscript. Indeed, although a skilled and committed game master is required for administrating the full program, it is possible that clinicians with less TTRPG expertise – but that display an interest in it – could implement some specific

ingredients of our program in individual or group therapies. The present study is just the initial step. The next steps anticipated (in the best-case scenario) are as follows: (1) testing the feasibility and initial efficacy of the program in a group of clinical participants (i.e., clinical pilot); (2) conducting a larger scale RCT; and only after – in case of convincing evidence – developing training sessions for interested clinical psychologists. But all of this is probably a matter of years, as correctly emphasized by the Recommender Veli-Matti Karhulahti.

- Finally, I would like to see more discussion on the generalizability of the program to clinical populations or a mention of what aspects/features of the intervention might need to be adjusted for such populations. While the authors note that the pilot was conducted on a subclinical sample, a more detailed discussion would be helpful.

Answer: **It is an empirical question whether the program will work in a clinical sample of participants. We plan to conduct a clinical pilot with the same program and a single group of five clinical participants in 2025 (also using a multiple single-case design). The participants will be individuals seeking treatment at the Center for Excessive Gaming in Lausanne, Switzerland. This clinical pilot is necessary to determine whether adjustments to the program should be made for this specific population. In response to this comment, we modified a paragraph in the discussion section:**

"*All in all, the results of the current pilot are promising and pave the way for future studies in clinical samples of socially anxious gamers. However, the question remains as to whether the program should be adapted (e.g., adjusting the difficulty of in-game objectives or the number of participants per group) for a clinical population.*".

To conclude, the authors did a great job piloting the program, especially considering how well they managed the challenges of conducting longitudinal interventions in real-life settings. I'm looking forward to seeing other teams implement the procedure as a way to alleviate problematic behaviors and add evidence on its efficiency and usability.

Answer: **We warmly thank this Reviewer for recognizing our merits and the complexity of this study.**

Matúš Adamkovič

Reviewer – Charlotte Pennington

Overall, this is an excellent Stage 2 report that clearly and transparently reports the findings of this study and interprets the data in a balanced manner. I am particularly impressed by its clarity and the Discussion section which considers the impact of all necessary deviations to the protocol and interprets the data in line with these. I am also impressed by the Conclusion that outlines how this is the first pilot evaluation that follows open science principles throughout and responds to a call made in the field to conduct more robust and well-designed empirical studies. The clarity of this manuscript has made it one of the easiest Stage 2 reports that I have reviewed for PCI Registered Reports, and the authors should be very proud of this.

Answer: We have no words to qualify how this comment was pleasant to us. We warmly thank this Reviewer for recognizing our merits and the transparency/clarity of our approach.

My overall decision is for minor revisions to be made and some clarifications to be outlined in the response to reviewer's document for this manuscript to fully meet the Stage 2 requirements of PCI RR, as follows:

2A. Whether the data are able to test the authors' proposed hypotheses (or answer the proposed research question) by passing the approved outcome-neutral criteria, such as absence of floor and ceiling effects or success of positive controls or other quality checks.
I agree that this criterion is met. This is a pilot evaluation to test the feasibility and efficacy of a 10-week TTRPG programme and to assess whether it can reduce gaming involvement, problematic gaming, social anxiety, and perceived loneliness. The data collected are able to answer the primary and secondary research questions and all deviations are transparently outlined throughout – these deviations appear necessary to make the project feasible and the study contributes to the literature to suggest that the TTRPG intervention is feasible and may be used to reduce social anxiety and gaming disorder symptoms. I am also impressed that the authors have outlined each of these deviations in the limitations section (e.g., how they were unable to test a reduction in time spent gaming at the three-month follow up) and have interpreted their findings in line with these. I have one minor comments in this regard:
The Results state that "Only two participants dropped out of the TTRPG program…", but these participants were not replaced with other participants to adhere to the participant sample size outlined in the methods (n = 20, four groups of five participants). As such, it would be good to clarify the final sample size (n = 18) within the Method section itself (and elsewhere where relevant) to ensure this is clear. I also thought that this would be a useful addition to the Abstract.

Answer: Thanks for this point. We made the requested changes.

2B. Whether the introduction, rationale and stated hypotheses (where applicable) are the same as the approved Stage 1 submission. This can be readily assessed by referring to the tracked-changes manuscript supplied by the authors.
I partially agree that this criterion is met. Looking at the tracked changes document, you can see that several changes are made to the Stage 1 Introduction and Methods sections that are not usually permissible at Stage 2. However, a closely look at these changes does show that many of these are simply rephrasings or expansions of previous sentences, so nothing 'untoward' or different is added. The authors should note that this isn't usually appropriate for RRs, however. The Editor may also wish to specifically look closely at the "Current study" subsection to ensure they are happy with the changes made there.

Answer: Thanks for your flexibility on this issue. See also our response to comment 2 of the Recommender.

I have the following comments regarding some specific changes:

The Stage 1 Abstract used to state "Outcomes assessed include social skills, self-esteem, loneliness, assertive, and gaming disorder symptoms." However, these are now split into primary and secondary outcomes: "Primary outcomes assessed include gaming disorder symptoms, time spent gaming, and social

anxiety symptoms. Secondary outcomes assessed include assertiveness/social skills, self-concepts, and perceived loneliness". I have looked at this closely and this change reflects the primary and secondary outcomes stated in the Stage 1 report, so the addition in the Abstract is to simply clarify the variables. One minor point here is that it would be good if the Abstract followed the same order as the Data Analytic Strategy with regards to these variables: specifically, the Abstract outlines gaming disorder prior to time spent gaming, but the Data Analytic Strategy presents these the other way around.

Answer: Thanks for notifying the inconsistency. This is now fixed.

Most of the changes to the Methods text are outlined in footnotes which is excellent to see. However, were any of these changes discussed and approved with the Editor, e.g., the change not to have a clinical psychologist verify a IGDT-10 diagnosis? Personally, I can see why this change was required and deemed not necessary but just a note (as an Editor of PCI RR myself) that any changes should be discussed with the Editorial Board before implementation to ensure the Stage 1 IPA stands.

Answer: Thanks for remembering this. Almost all changes conducted were anticipated in the risk analysis of Stage 1, and thus available from the PCI RR website (e.g., reduction of the conservative cut-off used for social anxiety symptoms). Yet, we had to decide "on the spur of the moment" for some changes in the study design (e.g., adopting a 3-week baseline for each group). The fact is that without these changes, it would simply no more have been possible to conduct the study, so we reasoned that the Recommender would judge whether these changes are or not acceptable. Of course, all the changes were transparently disclosed and discussed in the letter accompanying the Stage 2 manuscript. Next time, we will take care to notify the Recommender sufficiently in advance.

The following sentence has been omitted from the Methods section: "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". However, it is now clarified within the Data Analytic Strategy section, so such change is fine (and I just wanted to note I had looked at this closely as usually changes to Stage 1 text are not permitted). In the "Data Analytic Strategy", the authors have removed the sentence "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". I would clarify here that all deviations are now reported in the footnotes and were approved by the recommender(?).

Answer: This issue was transparently discussed in the letter joint to the submission of the Stage 2 paper, as follows:

"*We initially planned to have baselines of three to six measurement points (i.e., different numbers of measurement points for each group), with the onset of the experimental phase randomly determined for each group (Group 1 starts after a three-week baseline, Group 2 starts after a four-week baseline, etc.). This methodological approach corresponds to international standards for defining an experimental single-case methodology with enough data points per phase. However, we had to adjust our design to guarantee the feasibility of the study. The reason is that the recruitment process took more time than anticipated, and it was crucial for us to finish the 10-week intervention phase before the beginning of the holiday season in Switzerland (early July) to avoid massive dropouts at the end of the study. Accordingly, we decided to apply a similar 3-week baseline for each group, implying that our design became quasi-experimental instead of experimental (Tate et al., 2015).*"

We also transparently mentioned this issue in a footnote in the Stage 2 registered report.

2C. Whether the authors adhered precisely to the registered study procedures.
I agree that this criterion is met. In the main, the registered protocol is followed and, where necessary, any deviations appear to have been necessary and are outlined transparently in footnotes and in the Discussion section. The OSF page is extremely well organised and clear, with all materials, code, and data publicly available and presented in a FAIR manner.

Answer: We are happy that this Reviewer considers the criterion is met, we really did our best to follow carefully the registered plan and to transparently disclose and describe all deviations. Loïs Fournier, who organized the OSF page, is particularly happy with this comment.

2D. Where applicable, whether any unregistered exploratory analyses are justified, methodologically sound, and informative.

I agree that this criterion is met. The authors outline the findings for each of their primary and secondary outcomes and transparently delineate between these throughout.

Answer: We are happy that this Reviewer considers the criterion is met.

2E. Whether the authors' conclusions are justified given the evidence.
I partially agree that this criterion is met. I recommend the following revisions and/or clarifications:

Abstract: I recommend revising the final sentence of the Abstract to be a little more tentative given that this is (a) a pilot study and (b) there were n = 18 participants; specifically I recommend changing the term 'can' to 'may' in the following: "and can be used to reduce social anxiety and gaming disorder symptoms".

Answer: This correction has been done. We also applied the same modification in the first paragraph of the discussion:

"*The results observed are encouraging and suggest that TTRPG-based intervention programs may be used to reduce social anxiety and gaming disorder symptoms.*".

Results: In the Results section, the authors report that 10/18 completers participated in all 10 sessions, and then go on to interpret this as "high". Do you really think that a 55.56% completion rate of all 10 sessions is "high"? You may want to be a bit more tentative here to suggest that is it relatively high? I also note that this is 'brushed over' in the Discussion section by stating "and the majority participated in all 10 TTRPG sessions"). I do understand that 55% is still a 'majority' but it's not as impressive, I'm afraid, as the authors currently make out. I think further discussion of this, perhaps in the Limitations section, is warranted – what steps could be taken by future research to ensure that this is either met or exceeded? It's also noteworthy that you paid participants and this may have influenced their engagement and the success of the project; something I feel needs to be, at least, noted within the Discussion.

Answer: Thanks for pointing this out. First, we now take care to use neutral terms to describe the feasibility results, to avoid subjectivity and let the readers make their own assessment of our pilot results (see also the examples provided in response to the first point of the Recommender).

In terms of participation, 55% completed all the sessions, but only two of the 18 completers missed more than one session. This implies that 88.88% of the completers missed either zero session or one session. For us, this was highly encouraging. But as said above – we revised and avoided qualifying our feasibility and efficacy results, given the pilot nature of the study.

We now also mention in the first paragraph of the discussion that the participants were paid, which might have influenced their commitment.

The following points relate to the Discussion:
The Discussion starts by outlining the positive findings regarding feasibility and participant's engagement with this programme. It then explains that one explanation is that TTRPGs share many features with videos games and similarly allow for fulfilling basic individual needs. I would clarify here that the inclusion criteria was being being a gamer with past or current experience of playing MMORPGs or online RPGs and, as such, this intervention is likely to work for those with experience with, or interest in, playing video games. That is to say, the generalizability of this intervention for those who are not gamers is unknown.

Answer: We indeed cannot know if this intervention would work for non-gamers. It is thus an empirical question whether our program works with socially anxious people not involved in playing videogames.

The second paragraph of the Discussion states that "Regarding reliable clinical change (post-intervention and follow-up), the effect was less pronounced", and then goes onto discuss this generally. It would be good to be more specific here – what specific findings demonstrate that this effect was less pronounced? (remind the reader). Similarly, in the Results section you clearly outline that many of these effects were short-lived, but this doesn't seem to come across within this Discussion section, which is key to evaluating the pilot.
What do you mean by the term 'diminished' in the following sentence: "Self-reported time spent gaming diminished post-intervention for a majority of the sample"? This could be clearer.

Answer: According to this comment, we rephrased the concerned sentence:

"*Crucially, single-case analyses showed that our TTRPG program was very efficient for a subgroup of participants, while its effect was more limited or null/negative for a few participants (see Table 2 for details). In this sense, our study confirms the soundness of single-case analysis to test treatment efficacy, which in the current case allowed us to emphasize the important heterogeneity of participants' responses to the intervention. In terms of reliable clinical change at post-intervention, the effect of the intervention was less pronounced (i.e., less participants benefitted from our program) than what was observed with TAU-U analyses.".*

We also rephrased the sentence this Reviewer found unclear:

"*Self-reported time spent gaming decreased during the intervention compared to the baseline phase for a majority of the sample (10/18 cases) whereas for a significant portion of the sample (8/18 cases), no change or even an increase of self-reported time spent gaming was observed (see Table 2).*".

The following points relate to the figures:
Figure 2 states 5 participants in each of the 4 groups, but it would be good to update this to the final sample size with 5 participants in 3 groups and 4 participants in 1 group. Alternatively, you could put 'target' and then 'achieved' to clarify this.

Answer: Figure 2 is about the study design and not about the results. This is why it indicates five participants per group. At the beginning of the pilot, we secured 5 participants per group. But one stopped after one session, and another after 4 sessions. Moreover, the study flowchart (Figure 3) is very clear with the fact 2 participants dropped-out from the study. For this reason, we decided not to modify Figure 2. We hope this Reviewer will be convinced by our arguments.

Figure 3 presents a study flowchart but I am a little confused by the 2nd box which outlines those excluded from the study. Specifically, you state that n = 123 did not meet the inclusion criteria but then at the bottom you state that n = 40 did not meet 'other inclusion criteria'. What is this other inclusion criteria and why can't you combine this in the first 'did not meet inclusion criteria' statement? Otherwise this figure is fantastic and very clear.
I hope this comments prove useful to the authors in revising their manuscript and I want to reiterate that this is a really strong manuscript that I have enjoyed reading and learning from.

Answer:

The N = 40 correspond to a lot of different cases (e.g., not speaking French, not being a gamer with past or current experience of playing MMORPGs or online RPGs, not reporting motivation to commit to playing a TTRPG for 10 consecutive weeks and/or to undergo weekly psychological assessment during the three-week baseline, intervention, and follow-up phases). To avoid complicating too much the Figure, we decided to group them and only put the details for the most informative cases (i.e. those reporting too much or not enough symptoms of social anxiety and/or gaming disorder).

Among the *n* = 123 participants who fully completed the online eligibility survey, there were *n* = 83 participants not meeting the inclusion criteria due to their scores on the IGDT-10 and/or LSAS-24. For these *n* = 83 participants, we listed the specific criteria for their non-eligibility in Figure 3. In addition, there were *n* = 40 remaining participants not meeting the inclusion criteria, not due to their scores on the IGDT-10 and/or LSAS-24, but rather due to other combinations of criteria that we listed in the "Participants" section of our manuscript. The rationale for employing the term "other inclusion criteria" is that – considering our comprehensive list of eligibility criteria – there were numerous combinations for non-eligibility (endorsing one or more non-eligibility criteria from, e.g., not being 18 or older, not being fluent in French, never having played MMORPGs or online RPGs, having prior experience playing TTRPGs) that we aggregated to prevent Figure 3 from being unnecessarily complex and confusing.

Signed,
Dr Charlotte R. Pennington,
School of Psychology, Aston University, Birmingham

Additional changes (not requested by the Recommender or Reviewers)

---

1. We decided to add to the OSF repository the letter that accompanied the submission of our stage 2 exploratory registered report. We did that to maximize transparency, as this letter is not available from the PCI RR website. This letter can be downloaded from here: https://osf.io/y4pez.

2. The following sentence was added in the acknowledgment:

"*We also thank* Peer Community In Registered Reports *(*PCI RR*), the recommender of the present registered report (Veli-Matti Karhulahti), and the three reviewers (Matúš Adamkovič, Charlotte Pennington, and Matti Vuorre) for their insightful, constructive, and comprehensive comments, which enabled us to improve the study design prior to data collection and to strengthen the presentation and discussion of the results obtained.*".

3. Charlotte Eben's second affiliation has been updated.

4. A few additional typos were corrected (not highlighted to facilitate the review process).