Responses

We are very thankful for all reviewers and the recommender for their constructive and detailed feedback. Below we respond to each comment point by point.

Charlotte Pennington – Recommender

You will see below that the majority of comments are minor, but I would like you to pay particular attention to the following: Reviewer 1 makes some important points regarding the methodology, with one comment referring to the different scoring approaches used in the field to categorise problematic gaming, and the potential for these to impact the findings/interpretations. This is where Registered Reports are particularly beneficial, as you justify the scoring approach in advance. I ask that you are particularly attentive in responding to this concern. Please note that Reviewer 1's comments have been provided in a Word document rather than in-text.

We agree this is a highly important point; a detailed response is below.

Both Reviewer 1 and 2 note some confusion with the terminology of "remarkably lower" in the hypotheses, and Reviewer 2 asks whether this could be changed to "significantly lower". Can you please clarify the wording of these hypotheses on Page 4?

We agree with this issue too; a detailed response is below.

Reviewer 1, 2 and 3 note a lack of clarity with regards to the recruitment timeline and strategy. The recruitment timeline could be included within a table and, with your concerns around word count, perhaps this could be included as supplementary material on your OSF page.

In this case, our recruitment process is relatively simple. We have a contract with Bilendi, and after we have provided them the survey and instructions (which they already have from our pilot), they begin data collection via their panel and transfer us the data (in a total of ~month). We have added this information in the Appendix, but due to its brevity, we did not create a table for it. If more details are needed or we have misunderstood something, we are happy to revise and reconstruct this section.

Reviewer 4 makes a very good suggestion about conclusions – could the findings suggest the abandonment of such constructs altogether rather than any attempts at improved conceptualisation?

This is a difficult question; a detailed response below.

In my own assessment of your manuscript, I noted the following which should be addressed in a revision: The Abstract is a little unclear in parts, e.g., the term "related ontologically diverse screening instruments" is difficult to read – could the word 'related' be removed here? And "each of which representing a different ontological basis" - is this meant to be "each of which represent"?

We agree the abstract was very unclear; we have streamlined and clarified it.

Page 2, Introduction: What do you mean by “essence” when outlining the research questions? This term isn’t used elsewhere until a lot later, so this can be confusing for the reader. I also recommend changing the word ‘good hypotheses’ to ‘informed hypotheses’ or something similar.

We have switched to “informed hypotheses” and changed “essence” to “overlap,” which is clearer.

What will be the conclusions, as like in your pilot research, some of the measures overlap (e.g., in prevalence rates) but one or more do not?

This is an important question. To not further complicate and lengthen the manuscript, we hope it can be resolved here (as these documents will also be publicized and stored). As to each H1, H2, and H3, they all consist of sub hypotheses, so it feels most pragmatic to draw conclusions mainly from the sub hypotheses – unless all sub hypotheses of a main hypothesis are fully corroborated or refuted. We have added this information in Appendix 2 part “Theory”.

I am a little confused by the difference in H1 and H2: if you expect the ICD-11 and DSM-5 based “gaming disorder” prevalence rates to be lower than DSM-IV (H1), then how does it follow that the ICD-11 and DSM-5 will overlap with DSM-IV in essence (H2)? Perhaps this is just my misunderstanding, so I am happy for you to just respond to this in your response to reviewers if it is.

Figure 1 can be used as a reference here: even if some prevalence rates (e.g., GD-1) are smaller than others (e.g., GD-2), the smaller group can fully overlap with the larger group, i.e. everyone who meets GD-1 criteria also meet GD-2 criteria (but not vice versa). Figure 2 shows the opposite scenario: only some of GD-1 (with smaller prevalence) overlaps with GD-2 (higher prevalence). Indeed, the overlap is not perfect in a sense that, due to size difference, perfect overlap cannot be reached literally; what we mean is the highest possible overlap. We did not add anything more about this to the manuscript to save words, but we can do that in the next iteration if needed.

For H3, you state that this is compared to the general population – where are you getting the estimates for the general population?

A detailed response is below (R1).

In the Introduction, H1 specifies that ICD-11 and DSM-5 will have lower prevalence rates than DSM-IV, however, in the Method section, a reiteration of this hypothesis also mentions THL1. These should be consistent throughout.

We noticed that our use of acronyms has been confusing (see R1 responses below); H1 speaks of self-assessment based problems, which is the same as THL1 – thus, we have now removed many acronyms from the manuscript to make it more consistent and readable. We now primary use ICD-11, DSM-5, DSM-IV, and Self-assessment labels. On the other hand, we decided to use GRHP for “gaming-related health problems,” which saves us dozens of words throughout the manuscript and streamlines some of the hypotheses.

Page 7, “If neither mental nor physical effects are nonsignificant or below d=.22, we consider H3a/H3b not supported”, do you mean “significant” here? Can you also provide a reference to equivalence testing so a reader could read up on this to understand this (relatively new) technique (e.g., will you be using guidance by Lakens, 2017?).

Corrected, and a reference added.

Page 9, you state that it would be “impossible for us to collect a representative sample” – is this based on resources (too large sample size)? Please clarify why this is impossible (e.g., “the number of resources required makes it impossible.”).

We have removed this part entirely to save words; we realized there is no need to justify the lack of significance testing, as significance testing is not by default a gold standard.

The hypotheses are outlined in both the Introduction and Methods section, but this is a little confusing because the Methods further outline sub-hypotheses (e.g., H1a-H1d etc.). There needs to be consistency. One option is to move all hypotheses to the Introduction and remove these from the Method, and the other option is to remove from Introduction and just have in the Method. In the former, the Method could then simply refer back to these hypotheses, e.g., “To test the hypotheses of XX to XX, we will..”. I will leave the authors to decide what is best to do here.

All hypotheses are now moved to Introduction.

In the Method, what do you mean by “nothing will be corroborated”: do you mean that the results (prevalence) would be inconclusive?

Corrected.

I also note that you would like to submit to a particular journal upon Stage 2 acceptance – please note that the word counts for RRs at this journal include the references section too. I was unable to check the word count due to your document being in an online PDF, but I just wanted to notify you of this. I think you could remove the code detail within Page 9 H3 – if you make your analytical code available (which is a mandate to adhere to TOP guidelines), then readers/reviewers can see the calculations/code there instead.

We are thankful for this important practical point; indeed, we overlooked this detail in the word limit. This seems to make fitting the manuscript within 5000 words very difficult if not impossible if we want to maintain methodological rigor. We have tried to cut down the number of words even more, but we can also see the journal has a note “longer contributions negotiable under special circumstances.” We understand and accept that the manuscript might not fit in our target journal, but perhaps there will an opportunity to negotiate an extra 1000–2000 words at Stage 2 (we do not expect this but express our willingness to negotiate if it is possible).

Please can you add the funding information to the title page of the manuscript.

Added.

Reviewer 1 (Linda Kaye**)**

Thank you for allowing me the opportunity to review this RR. I am glad to see work in this area going through the RR process and appreciate the authors' efforts in detailing the rationale, methodology and analysis plan with a good level of detail. I have included my comments in line with the reviewer criteria below, as well as some more general observations/comments at the end. I wish the authors all the best with their research and hope my comments are helpful in their research endeavours.

These are important questions and I am glad to see the conceptual bases of "gaming addiction" being queried as this is an ongoing concern in this field. The RQs therefore seem relevant although I have some observations about some of the specific hypotheses which put these in operation (see below). It would perhaps be helpful in the rationale to make explicit reference to the distinction between core addiction criteria/symptoms (e.g., tolerance, withdrawal, mood modification, etc) and gaming-related problems (e.g., interference and impairments in life events) which are noted to often be conflated within empirical work (Colder Carras & Kardefelt-Winther, 2018). E.g., Colder Carras and Kardefelt-Winther (2018) identify four classes of player; IGD class, normative class, engaged class, and concerned class. However, their latter two classes would be misclassified based on their reporting of gaming-related problems only. Also Myrseth and Notelaers (2018) identify five classes; never symptoms, rarely symptoms, occasionally symptoms, problem gamers, and disordered gamers. Again, this suggests that different classes of players may have distinct patterns of gaming-related behaviour/gaming-related problems. Including some explicit statements about how each of the instruments being used are able to measure gaming addiction symptoms vs gaming-related problems would be helpful here I think.

This is a highly important observation, and the issue will be worth investigating explicitly in future research (and especially in scale development). Perhaps the key issue -- in the field in general -- is that we don't have good basic clinical data to assess how criteria/symptoms differ from problems, if they do. It should also be noted that not all ontological perspectives distinguish between criteria/symptoms and problems, so the relevance of distinction tends to depend on the source. Let us use DSM-5 as an example, which defines IGD as follows:

"Persistent and recurrent use of the Internet to engage in games, often with other players, leading to clinically significant impairment or distress as indicated by five (or more) of the following in a 12-month period."

In this definition, the problems (clinically significant impairment) are included in the symptoms by the phrase "as indicated by." On page 20, the DSM-5 also stresses that "Each disorder identified in Section II of the manual (excluding …) must meet the definition of a mental disorder," which is then followed by the problems caused by disorders. However, IGD is not in Section II but Section III, so this does not seem to apply to IGD, which was assumedly created so that the symptoms are/include the problems. We hasten to add that the above interpretation is by no means the only possible one, and we do not claim that there is one "correct" interpretation. But we hope this example illustrates that the distinction between "disorder symptoms" and "problems" is not clear, and they could also be one and the same thing. The latter interpretation has been consistently chosen by almost all validated screening tools, and while we fully agree that many of them would have benefitted from additional health problem control, it would also be methodologically problematic to modify these existing

validated scales or their official interpretation guidelines. It might even be considered undermining our findings if we diverge from the scale guidelines.

To address this issue, we have added a paragraph in the introduction, stating the need for new categories and citing the indicated literature. We have also added in the Analyses section that our interpretations follow the diagnostic manual recommendations. We do not know if these scales/cutoffs are a useful approach for identifying people with actual disorders or problems, but by following the diagnostic guidelines, respectively, we can measure the constructs as they were officially structured (whether they make clinical sense or not).

This would also be important in relation to some of the hypotheses. For example, H3 makes suggestions about those with "gaming disorders" from DSM 4 (GAS7) & self assessment (THL1) but up to this point, we don't have much insight to know whether these two measures make clear the addiction symptoms from the health-related issues and so isn't clear why this specific hypothesis is directed in this way.

Following the previous point, I also think it is potentially problematic to be forming hypotheses which expect poorer health in those who self-assess compared to general population (H3b). That is, people with gaming-related problems (which may be picked up from the THL1 scale) could equally be representative of the general population and not specifically distinct in the way those who are "addicted" may be from a general population. Further H3c expects health related outcomes between those with DSM-4 and THL1 scale to be similar. I am not sure I follow this rationale, based on one perhaps being incorporating more core addiction-related symptoms (DSM-4) and the other not (THL1). The hypotheses here may therefore require additional rationale or be removed if this reasoning is difficult to articulate

This is an excellent point. Indeed, GAS7 is based on DSM-IV ontology and THL1 is based on self-assessment ontology, and we have no strong reason to expect them to be similar (or different). Originally, we had a theoretical justification (they both measure gaming problems, so they assumedly operate similarly), but that's not a strong justification. To solve this, we changed our THL1 cutoff to 2/4, which is the same ("something") as used in GAS7. This lower cutoff allowed us to identify n=138 in our pilot for THL1, and this sample size was enough to calculate prior Cohen's d also for the health of THL1 2/4. Again, the evidence is mixed (see the manuscript for details). We also managed to access previous National Health Institute data with THL1, which provided further mixed evidence. Therefore, we added for H3b and H3c competing hypotheses (H0), i.e. we will assess equivalence and effects for both. For simplicity, we use "H3/H0" in the manuscript (instead of creating confusing H3b-1, H3b-2, H3c-1, H3c-2).

We have also hierarchized all our hypotheses into primary, secondary, and tertiary hypotheses to communicate our (lack of) confidence in them (following APA guidance by Cooper, H. (2020). *Reporting quantitative research in psychology: How to meet APA style journal article reporting standards*. APA). We did not cite Cooper, as our use of "tertiary" is not exactly in line with the source, and we already struggle to stay within the 5000-word limit (i.e. there is no room to explain this in a footnote).

Figure 1 is helpful to articulate the various scenarios. However, I feel some further rationale behind this is needed about why four groups of GD are expected to be derived from the

measures. Why “severe”, “medium”, “minor” & “v minor” ; is this based on existing classification levels of these measures?

The Figure has now been clarified. As these scenarios are just examples, the terminology is not tied to any specific classification. However, we have changed “v/minor” to “mild” in order to use consistent wording with DSM-5 (not to explicitly follow DSM-5, but to reduce variety in terminology in the field).

In the hypotheses (e.g., H1 including H1a-H1d), the term “remarkably” is used to explain expected differences in prevalence rates. Arguably “significantly” would be a more typical term here.

In fact, we originally had “significantly” but we were advised to change that by the Managing Board (Zoltan Dienes) because it can be confused with statistical significance (and hypotheses concern population properties, which are not statistical). We agree “remarkably” is confusing too; as a solution, we now use “meaningfully” in the hypotheses to communicate that the difference is not only directional but also meaningfully so, in terms of the population, as measured by, e.g. the smallest effect size of interest.

There could be more context given about the “health-related problems” before it gets to the methodology where it becomes clear it is referring both to physical and mental health

A note regarding this has been added.

Analytic decisions etc are well detailed and provides insight into how these correspond to RQ/hypotheses. The analysis appears feasible but I feel more information about the sampling strategy would help know whether the sample size is feasible and whether it is representative (see next section)

A mixture of monothetic and polythetic approaches to cut-offs is proposed. It is good that these different scoring approaches are noted. However given recent findings that different scoring approaches are better or worse than others at capturing differences between “problematic gaming” groups (Connolly et al., 2021), it may be worth considering whether a consistent approach should be used (or do analyses which test both monothetic and polythetic for all analyses)

Connolly, T., Atherton, G., Cross, L., & Kaye, L. K. (2021). The Wild West of measurement: Exploring problematic technology use cut off scores and their relation to psychosocial and behavioural outcomes in adolescence. *Computers in Human Behavior, 125*, e106965. https://doi.org/10.1016/j.chb.2021.106965

Here we must highlight that the polythetic/monothetic difference is one of the key differences between the ontological systems; to test the differences between the systems, we must follow their official rulesets. We have now added a complementing section about this in Introduction. On the other hand, we completely agree that what counts as “endorsement” of criteria (in polythetic, monothetic, and other cutoffs) seems to have no strong evidence for or against any chosen approach; one of the few reliable ways to do that would be to clinically validate endorsement rates with samples that have been reliably found be significantly impaired due

to gaming. As we will use the officially stated cutoff in each scale, we have now added alternative endorsement rates to be exploratively tested, too.

"we include two control questions (Oppenheimer et al. 2009) in the survey and remove those responses that fail both. Participants who report not having played videogames within the past six months will not fill out the gaming-related screening instruments, and they will be considered not meeting the problem criteria that concern the present study"- I find this a little problematic in the way this is written. The last part of this sentence by implication suggests that those people who have played games in the last 6 months do meet the "problem criteria" which I feel is not a justified assumption.

Rephrased.

There is currently not sufficient information about the recruitment strategy in relation to where people will be sampled from. If a representative player sample is sought, then some details about the relevant contexts they can be sought from is required. This information would be important especially in ensuring that any replication efforts are matched in this regard

New information added.

Description of the outcome measures provides information about descriptors of the scoring ranges as reference points. There is not a control group but that is not unusual for research on this topic which is conducted in this way. The authors however may consider whether it is worth recruiting a control group to take the Physical and mental health measures as a control however especially as some of their hypotheses are making comparisons to the general population. Without a control group, this comparison may not be achieved.

We highlight that our sample N=8000 is not a gaming sample, but nationally representative general population. When we do comparisons between groups (as identified by each scale), those who are left unidentified will form a nationally representative group for comparison (e.g., if problem group = 200, then general population = 7800). That said, we did not previously make any differences between gaming and non-gaming in the general population, and we have now added one more explorative test where we divide the general population into gaming and non-gaming to assess if there are differences between them (thus having two separate group comparisons, controlling if gaming alone is relevant here).

In the Introduction, the term "abnormal technology use" is used. I would advise the authors to avoid this term and instead perhaps use "problematic technology use" or similar. The cognitive load in reading this is quite heavy as acronyms are used quite a lot to describe the different instruments (and sometimes interchangeably between the initials of the scale and the diagnostic criteria they refer to). It would be most helpful to reduce acronyms down to aid readability.

We have reduced acronyms as much as possible.

R2 Daniel Dunleavy

1. I found each of the four sub-hypotheses for H1 to be clear (p. 4). I was initially hesitant, finding them to be poorly defined (i.e., what "remarkably lower" meant). However, I found the description of the interval-based method on p. 5 to be sufficiently clear to satisfy my concerns. I feel that the method used here is described in enough detail to be replicated by another set of researchers and further that it is an adequate method for assessing the hypotheses. The other two hypotheses (H2 and H3) are clear and testable - particularly H3, which clearly specifies the smallest effect size of interest (and justifies its selection).

As noted earlier, we have changed ”remarkably” into “meaningfully” for consistency.

2. Recruitment - p. 5 - The authors are recruiting using a company called Bilendi. I'd just like to see a little more detail about the recruitment timeline (expected start-finish dates) and how participants will 1) be incentivized, and 2) complete the survey (i.e., what software or platform will they use to complete the survey?).

We have added a clarifying note about recruitment, and the appendix now provides more details about the process. Although we cannot know the dates at this point (they depend on review time), we start the process immediately when possible and it is expected to take ~1 month.

3. Sampling plan - pp. 7-8 - The authors provide a reasonable sample size justification and the required sample size to estimate prevalence rates.

Reviewer 3 Anonymous

The methods section largely contains enough detail that replication would be possible. Sources of ambiguity may include how and from where participants will be recruited via the 3rd party (online advertisments), and how the survey data will actually be collected; i.e., whether in-person or online. These considerations may impact overall representativeness.

In line with our earlier responses, we have now added this information (and more details in the appendix).

In most cases, explanations are adequate. There may be quesitons about what constitutes *exploratory* investigation in some tests, but the authors have justified the need for exploratory analyses where necessary.

We have further labeled our hypotheses primary, secondary, and tertiary, to express our confidence in them. We have also modified H3.

One future suggestion, that may be worth bearing in mind, is the digital convergence between gaming and gambling, with microtransactions and other predatory features that resemble gambling, including lootboxes. It strikes me that, given the overlap between gambling and gaming disorders in the DSM-IV and DSM-5 criteria, there might be the potential for confounding among participants. Controlling for this possibility might be something to think about.

This is a good point; we have added a gambling measure (BBDS) in the survey.

Reviewer 4 David Ellis

1. It would be useful if the Abstract and Introduction made it clear what 'risk groups' might look like. I assumed these were groups whereby their scores suggest some form of diagnosis/classification? However, as I read on I realised this is also about other aspects of physical and mental health. Is a 'risk group' both of these? A table summarising all measures would be helpful within the method section to provide further clarity in this regard.

We have added a table in the appendix. We have removed the word "risk" from the entire manuscript and, instead, use "criteria meeting" and "[manual] based" to communicate that they meet the specific criteria.

We were tempted to just use simply "gaming disorder" for all measures, but we try to fight against this general trend. We know it is a bit tricky, as we try to pursue a balance between accuracy and clarity. Each construct (despite being mixed in meta analyses and meta reviews) is arguably independent. Only ICD-11 criteria currently allow for official disorder diagnosis; DSM-5 criteria are tentative and are meant to indicate problems (but not official disorder); DSM-IV criteria are modified from gambling and thus rely on scholars' own theories (no official disorder); self-assessment, by definition, refers to the participant's subjective perception of having problems with gaming. By using "criteria meeting" (and referring to gaming-related health problems in general) and "[manual] based" we address all these possibilities – which may or may not be one construct. We hope to strike a balance between accuracy and clarity by this approach (as much as possible), and refer to these all constructs now systematically as gaming-related health problems (GRHPs).

2. The research questions could also do with some clarification or possible re-ordering. 'Who' really refers to overlap (or lack of ontological overlap) in the sample as I understand it. It feels like it should be the first research question. The second question could then consider prevalence as this is making more population-based inferences (I think). I am not familiar with the interval-based method outlined so can't comment further. Wondered a bit as to how this goes beyond a series of chi-squared tests that could also consider ontological similarities. Research question C is clearly about Health and that is clear.

We have clarified this by changing "essence" to "overlap" and have moved all sub hypotheses to Introduction to improve readability. We did not change the order of hypotheses, as other reviewers found the current order suitable.

3. I suspect the results, regardless of direction, will have implications for many notions of technological 'addiction' or related 'disorders'. If ontologies are different, then this suggests a problem for comparisons across different studies. If they are all the same, then this doesn't only indicate that scholars should direct efforts toward assessing the clinical relevance of multiple constructs as suggested. On the contrary, it begs the question as to why researchers have continued to re-invent the wheel and made little progress regarding a consensus. Creating a new disorder or ontology doesn't create new resources for pracititoners or clinics. What are the societal costs of stigmatising the most popular form of play in children and adolescents? I guess what I am getting is that results from this work

might suggest the abandonment of such constructs altogether rather than any attempts at improved conceptualisation.

We very much agree with this remark and, in terms of positionality, believe that many, if not all, current screening tools are incapable of properly measuring actual problems. Many of us also believe that “gaming disorder” as a construct needs restructuring and some of the other constructs (such as “internet gaming disorder” in DSM-5) do not deserve a diagnostic category of their own (as the APA concluded in 2013, too).

That said, we do not believe that the results of our work can suggest the abandonment of such constructs *altogether*. This is because our methods are limited to self-report surveys, which cannot be used as direct clinical evidence. The overall validity and specificity of these constructs should be assessed and triangulated via clinical and other in-depth investigations across cultures; only in this way we can produce evidence that has justified power to suggest the abandonment of the construct(s).

That said, we have added a nod toward this dialogue by changing the term “clinical relevance” into “clinical (ir)relevance”, thus highlighting that relevance also involves the possibility of lacking relevance.