We sincerely thank all three reviewers and the editor for timely and encouraging reviews. We're glad we were largely able to communicate our plan and goals in the Stage 1, and think that the revision has been able to address the comments raised here. The primary changes include:

- The addition of positive controls that need to be passed before proceeding with the planned analyses (R1)
- A clarification of Study 3 Hypothesis 2, breaking it down into two component parts and better explaining the function and role of the joint test (R3)
- The addition of pre-specified limitations (R2)
- Clarifications to our approach to genre classification (R2)
- An expanded explanation of our sample size justification and the information gained by simulation (R1)

Specific responses follow line-by-line below.

We want to flag one unrequested alteration from the initial submission, resulting from an issue arising with our data-sharing agreement with Microsoft. Specifically, Microsoft is no longer able to provide the full name of 3rd-party titles (i.e., games not published by Xbox Game Studios). Instead, they will provide hashed IDs for each game alongside a genre label.

This is frustrating news for us, and reduces the value of the dataset for secondary use. However, we are fortunate in that it only requires one relatively minor change to the planned analyses: Study 3 will need to use Microsoft's provided genre labels, rather than those sourced through IGDB. We have documented this change in the main text below.

We also discovered while finalizing our contract with the panel provider that our budget is sufficient for a longer US diary study than originally planned, so this has been increased to 30 days from 21.

Reviewer 1

Review by Chris Chambers, 16 Sep 2024 13:51

> I read this Stage 1 submission with great interest. Since I am not a field expert, I leave specialist assessment of the theory and rationale to specialist reviewers and focus my review on general Registered Reports evaluation criteria and methodological rigour.
>
> Overall I felt this was a very clear and impressive submission that tackles a series of important research questions in an innovative way. The combination of digital trace data with longitudinal psychological data, together with a focus on reproducibility and transparency, seems (to this non-specialist) to

> be an ideal vehicle for moving this field forward. I also judge that the three programmatic components of the submission are sufficiently substantive to justify separate Stage 2 outputs.

Thank you for the kind words.

> I have very few comments, but offer the following suggestions to help maximise the quality of the Stage 1 proposal:
>
> 1. I suggest including a summary of the sample planning precision estimates in the section Sample Size Determination. It is fine to include the bulk of this in supplementary information, but there should be enough content in the main manuscript to provide a general overview (whereas at the moment there really isn't enough meat in the main manuscript).

We agree that this was buried too deep to be accessible to readers, in part because we don't want readers to put too much stock in the specific precision estimates given that these will vary based on a wide range of factors we feel ill-positioned to exhaustively explore. Rather, they should be seen as broadly indicative.

We have added additional detail in the sample size determination section to be clearer on this point:

> "Due to a lack of prior data and results for the varied measures in the data, we conducted simulation analyses based on the size and structure of the data and one reasonable parameterization of the distribution of variables and the relations between them. This method does not replace a formal power analysis, as the sensitivity of the tests will be determined by a wide range of interacting characteristics of the data (e.g., random slope SDs, autocorrelation coefficients, the true effect size, and so on) that would need to be simulated across a range of values—a prohibitively difficult task given that it will not affect our sample size, which is fixed by resources.
>
> However, these simulations may be able to give a broad indication of the expected precision of our tests. For example, in Study 1 H1, the 95% CI of the simulated estimate of a 1-point within-person change in gaming need satisfaction predicting general need satisfaction is .12 scale points wide. In Study 2 H1b, the 95% CI of the simulated estimate of a 1-hour change in daily late-night gaming predicting hours of sleep is .15 hours wide. In Study 3, the 95% of the simulated estimate of a 1-hour change in platform gaming on general mental wellbeing is .06 scale points wide. Together, while unlikely to reflect the exact precision of our tests with the true data, these arbitrary but reasonable estimates provide initial evidence that our tests have sufficient precision to detect small to medium relationships."

> 2. "We do not preregister any further exclusion criteria; in case of further quality checks (e.g., using careless; Yentes & Wilhelm, 2021) identifying additional responses to exclude, we will report results with both minimal and maximal exclusions applied." This sounds ok but, of course, any post hoc exclusion criteria should applied strictly to exploratory analyses. I'm assuming this is what the authors intend; if not, then these further quality checks must be precisely prespecified.

This was indeed the intention: we will report all results with the exclusion criteria as described under the philosophy that all data should be included unless there is a strong rationale to the contrary, but wanted to acknowledge that data quality checks are difficult to pre-specify with data this complex. We may discover further legitimate reasons to exclude data, which we would report alongside the preregistered analyses as exploratory sensitivity checks.

On reflection, this is unnecessary; whether described in the Stage 1 or not, we can always report additional sensitivity checks as exploratory analyses if the need arises. The paragraph now simply reads: "We do not preregister any further exclusion criteria."

> 3. Please add the various positive controls to the design tables, and in the main manuscript (and the interpretration column of the design table for the corresponding row), note the consequences for evaluation of the main hypotheses in the event that one or more of the positive controls fail. I would also strongly encourage a sensitivity power analysis for these positive controls. The success of positive controls can be critical for Stage 2 acceptance (see criterion 2A) so it is very much in the authors' own interests to be sure that the design is sufficiently robust and sensitive to capture these sanity checks.

We have added several positive controls, and conducted basic power analyses to ensure that our study design, if operating successfully, can capture these effects. We've added a section to the main text, and have copied this information into the relevant sections of the design tables. In the main text, this reads:

> "We specify several positive controls, which act as tests to ensure the data we collect is structured and co-related as expected. Passing these tests is therefore a prerequisites for proceeding with our analyses as planned:

Applicable Study	Test	Statistical Power
All studies	A significant positive correlation between self-reported video game play and digital trace playtime during the previous 2 weeks	Assuming n = 9,300 panel surveys (after 10% wave-on-wave attrition), a true population value of r = .2, an alpha of .05, and a one-sided test, power > 99%
All studies	There will be no overlapping sessions for a given individual on Nintendo or Xbox (we allow for possible overlap across different platforms, in	(N/A; fails if a single case occurs)

	case the user has two devices active simultaneously) AND there will be no cases where a player logs more than 60 minutes of playtime on Steam between adjacent hourly measurements	
Study 1	Significant positive correlation between need satisfaction in general and daily life satisfaction	Assuming n = 21,000 diary surveys (after 30% total attrition), a true population value of r = .2, and an alpha of .05, power > 99%
Study 2	Significant positive correlation between social jetlag as calculated by the Munich Chronotype Questionnaire and daytime sleepiness.	Assuming n = 4,440 panel surveys with sleep measures (Waves 2, 4, 6 only + 10% wave-on-wave attrition), a true population value of Spearman's rho = .1, an alpha of .05, and a one-sided test, power > 99%
Study 2	Significant negative correlation between sleep quality (Pittsburgh Sleep Quality Index sleep quality component) and Warwick-Edinburgh Mental Well-being Scale (WEMWBS).	Assuming n = 4,440 panel surveys with sleep measures (Waves 2, 4, 6 only + 10% wave-on-wave attrition), a true population value of Spearman's rho = -.1, an alpha of .05, and a one-sided test, power > 99%
Study 3	Significantly higher playtime in shooter games for men as compared to women (Lange et al., 2021)	Assuming telemetry data for n = 1,000 (as attrition during surveys does not prevent us from collecting gameplay data), a true population difference of d = .3, and an alpha of .05, power > 99%

Table 2. Positive controls used to assess whether data is suitable for hypothesis tests, and estimated statistical power of these tests"

> 4. "Responses where the two duplicate items differ by more than 1 scale point will be flagged for manual inspection of potential careless responding." – please define the precise rule for exclusion. What specific signs will constitute careless responding?

Thank you for catching this. We've decided that the attention check item is itself already a reasonable exclusion criterion, and the sentence now reads "Responses where the two duplicate items differ by more than 1 scale point will be excluded."

> 5. "We anticipate approximately 10% attrition per wave of the panel study, and 30% total attrition for the diary study. " Since participants who are excluded or drop-out are (presumably) not replaced, please specify a minimum sample size that will considered sufficient to answer each research question (and test each hypothesis) and therefore justify a Stage 2 submission. I am assuming this will be some number substantially below 1000 (and probably below 700 - i.e. 1000 minus 30% attrition minus maximum tolerable exclusions).

Correct, we will not replace participants who drop out. While it's challenging to specify a minimum sample based on strong scientific grounds for each individual hypothesis test(given that we deemed it unfeasible to conduct a trustworthy simulated power analysis for data this complex), we now specify a common-sense minimum that reflects what we believe would be the lower bound for success in terms of data collection:

> "The minimum sample size required to proceed with our planned hypothesis tests is 50% response rate throughout the diary (total N after 30 days ≥ 15,000) and panel (total N after 6 waves ≥ 3,000) surveys. This ensures that we do not impute more data than we collect."

Reviewer 2

> **Review by valtteri kauraoja, 13 Sep 2024 14:13**
>
> I thank the managing board of PCI Registered Reports and the authors for the opportunity to peer review this interesting stage 1 report using digital trace data alongside longitudinal wellbeing data to explore the quality and context of play on a large scale.
>
> **Reviewer's disclosure:**
>
> I am a first-year PhD-student, and lack expertise to comment on things like measures and statistical R-analysis. I hope my comments are still useful for the authors.

We sincerely appreciate your feedback here.

> **Technical:**
>
> -The ORCID-link supposedly for Przybylski A. is actually the link to Ballou's ORCID.

Thank you for catching this, now fixed.

> -The authors may consider registering Limitations already at Stage 1, e.g. for reflexive reporting

We agree that this is a useful exercise. While we do not exhaustively describe all limitations here (some of which will vary by study and be better addressed in context in their respective Stage 2 Discussion sections), we have added a paragraph to the end of the manuscript highlighting some overarching limitations. This reads:

> Across all studies, the absence of third-party Nintendo data means that we are missing ~30% of playtime on that platform. Importantly, the distribution of genres among 3rd party games on Nintendo differs from the genres of 1st party games, and thus the relationships might differ for this missing 3rd party data. Across all platforms, idle time—periods when games are left running but not actively played—and account sharing could inflate playtime metrics, introducing bias. The playtime

figures we report should be interpreted as an upper bound for the time spent actively playing during the study period. In all studies, our approach is observational and thus not positioned identify causal relationships between gaming and wellbeing.

In Study 1, reliance on self-reports of activities displaced by gaming introduces the risk of social desirability bias; participants might overstate intentions to engage in socially esteemed activities like exercising, which may not accurately represent their actual behavior in a counterfactual universe where they did not play games.

For Study 2, collecting sleep quality reports in the evening rather than in the morning may compromise data accuracy, as retrospective assessments can be less reliable than immediate reports upon waking.

Lastly, the lack of title information for 3rd party Xbox games means that we are reliant on Xbox's provided genre labels for categorizing games in Study 3. While Xbox's taxonomy largely corresponds to the open source IGDB database, discrepancies may nonetheless influence our estimates of genre-specific playtime and wellbeing.

> **Basic Psychological Needs in Games and Wellbeing (Study 1)**
>
> The simultaneous validity testing of BANG hypotheses and expanding it is impressive and commendable. I hope its use will give qualitative context to the idea of problematic displacements through games. Consider the subjectivity and limitations of self-reporting complicated displacements and related information.

The subjectivity of reporting counterfactuals is without a doubt a limitation worth highlighting. We now raise this in the added limitation section (see above), and will foreground this in the eventual Stage 2 output.

> **Game Genres and Wellbeing (Study 3)**
>
> The choice to use structured metadata repositories for genre categorization is sufficiently justified, and the acknowledgment of the limitations of self-report and researcher-ascribed taxonomies is valuable.
>
> One of these justifications is the accomodation of genre fluidity and evolution (p. 13). This is true within the context of contemporary, user-generated tags and genres. However, I recommend consideration of the fact that once analysis is being done, the genre classifications will have to become a fixed set and can desync with the genres presented by the database.

We're glad that this approach was well-justified, and we hope that the unfortunate development with regard to our data-sharing agreement with Microsoft does not undermine our position too badly.

> Page 13 talks of the community repositories in generalities, and the service used in the study is only specified in the method-section. It then remains unclear if the genre classifications on Internet Games Database are controlled by developers, service admins, the public, or some combination of these. I read it as implied that the study will not consider "themes", only "genres" within the service, but I wish it was explicitly stated and the definitions considered. The categorization used by IGDB can be questioned and the choice to include or exclude these different layers of classification should receive careful justification despite the already explored limitations of all genre categorization.

We have substantially expanded our description of IGDB, and how/why we use it. The relevant methods section now reads:

> Game genre will be obtained by cross-referencing game titles in the digital trace data with the Internet Games Database (igdb.com), which catalogues and categorizes games according to 19 distinct genres (e.g., Platformer, Role-playing (RPG), Simulation). IGDB is, to our knowledge, the only database with complete metadata coverage of games across platforms that offers an API for programmatic data retrieval. The platform is crowd-sourced; community members can submit contributions (e.g., a new game or alternative categorization), which are vetted by admins and moderators before appearing in the database. The database is thus dynamic as some entries may change over time (although for popular games with many contributions this is rare); we will use the genres as they appear at the time of study completion. We will use the first and primary value of the "genres" field on IGDB as this is the most parsimonious categorization of games*, and do not consider other variables such as "themes". A complete list of genres on IGDB can be found in Appendix A.
>
> *We considered allowing games to have more than one genre, but decided not to allow this as it would change the interpretation of our coefficients—rather than a 1-hour change in genre playtime representing the unique change of playing an additional hour of a game with that genre only, it would represent some unknown combination of (1) playing an additional hour of a game with that genre only and (2) playing the same amount of time as before, but "adding" a secondary genre to one's existing gameplay.

> **Method:**
>
> It would be interesting to see both details and analysis of the Nintendo data

> (Table 1). What is the exact definition (or list) of "close partners"? The note on the sales-dominance of 1st party games is valuable, but increased analysis could help the paper show how the lack of 3rd party data could affect, for example, genre-related data. This consideration would increase data validity and transparency.

We have clarified this. It now reads: "games published in whole or in part by Nintendo", and includes a footnote pointing towards a list of all known Nintendo-published games (https://nintendo.fandom.com/wiki/List_of_Nintendo_games). We acknowledge that 1st party titles likely reflect a different distribution of genres than 3rd party titles, but cannot easily address this. We have added a point in the registered limitations to this effect (see previous comment).

> Page 14 has the only mention of "game modes" in the paper. Such data is not mentioned in Table 1, so I deduce that this data will be extracted through the surveys. Details on the extraction and relevance of this data would be useful.

Thank you for catching this—the term "game mode" was misused here. We unfortunately do not have access to game mode information—we'll know what games people are playing, but not what mode they're playing in (beyond a self-report item that asks what kinds of multiplayer gaming they've engaged with in the previous 24 hours). For example, we won't be able to say whether someone playing EAFC24 is playing ultimate team, or manager mode.

Secondly, on reflection this paragraph was out of place—while we do fully intend to report descriptive analyses of who plays games on what platform (using the nationally representative screening data), and who plays what games/genres (using the digital trace data), this is a rather distinct goal from the current planned outputs. Given this, we've elected not to commit to a particular output format or descriptive analysis by including this in the stage one, and have removed the entire paragraph where "game mode" appeared.

> **Exclusion Criteria and Missingness**
>
> Consider if additional work into understanding and excluding false hours from digital tracking could help increase validity and reflexivity. While technical problems and system clock manipulation are concerns, I would also consider the risk of "idle hours", leaving games on while not actively playing. Ensuring that self-reported and digital trace data correlate is a good way to mitigate this risk, but I didn't see specifications on how closely the hours should correlate to be considered valid.

This is an exceptionally important point. Unfortunately, the tools at our disposal for excluding idle and menu time are very limited, and in our view an inherent limitation of the

method. We have added a point to the new limitation section to highlight this point, and will be transparent about this throughout all studies:

> "Across all platforms, idle time—periods when games are left running but not actively played—and account sharing could inflate playtime metrics, introducing bias. The playtime figures we report should be interpreted as an upper bound for the time an individual spent actively playing on linked platforms during the study period."

Given that we cannot be sure that self-report and digital trace data have the same underlying basis (e.g., in the case of a person self-reporting play of 3rd party Nintendo games or play on an uncaptured platform like Playstation), we are hesitant to specify a particular lower bound here—previous studies have found correlations as low as r = 0.35 (https://doi.org/10.1111/jcc4.12056), and ours could conceivably be even lower.

Ultimately, discrepancies between digital trace data and self-report data warrant their own in-depth analysis that we feel is out of scope in the current proposal. However, a positive relationship of some kind feels like a crucial quality check, and we have therefore specified a positive control on this basis (see response to R1 above).

> I considered the report novel, well-written, and impressive in its depth and scale.

Thank you.

Reviewer 3 (anonymous)

> The submitted programmatic registered report provides a plan of a large-scale study with three particular studies focusing on three broad effects of video gaming. As typical for this type of RR, the theoretical introduction is quite general presenting rationalization for quite heterogeneous research aims and hypotheses, so it can be used in future manuscripts only in case of elaboration.

We agree, the general introduction sets up the piece as a whole, and in any individual Stage 2 manuscript would be substantially condensed or cut. The study-specific introductions will be the foundation of each Stage 2 output.

> I see three main strengths of this RR. This first one is using real-time first-hand data, when the agreements with the VG providers allow for this. The second one is a sample, that seems as suitable for the planned analyses, although power analysis has not been conducted (operating by various constraints). Next, I appreciate the longitudinal data collection with focus on mental health, that allows investigating real-time changes.

These were the important elements in our minds as well. While power analysis was deemed more challenging than it would be scientifically valuable (given a sample size fixed by resource availability), we hope that the contextual information we provided in the data simulations partially addresses this.

> The hypotheses are clear and justified, I only have a comment to H2a and H2b in Study 3, which are not directional and it is not completely clear how they will be evaluated.

This is an important point, as we recognize the hypotheses in Study 3 are structurally different than those of Study 1 and 2, by virtue of testing an idea that the research literature often implicitly adopts (certain genres affect wellbeing while others don't), but which lacks a coherent theoretical basis, and thus a clear direction.
To address this, we made several clarifications to our approach in study 3. The first was to reword and separate out our hypotheses, which we realized are twofold: that (some) genres are different from 0, and that some genres differ from each other. Our hypotheses now read:

> H2. Biweekly playtime in one or more of the 23 game genres is associated with individual changes in general mental wellbeing (H2a, "within-person") and with wellbeing differences between participants (H2b, "between-person").
>
> H3. Genres differ in how biweekly playtime relates to individual changes in general mental wellbeing (H3a) and to wellbeing differences among participants (H3b).

Next, we clarified our modelling and interpretation approach. We apply the same model as before, and still use joint tests as planned, but we now specify separate joint tests to be conducted on that model and give much more detail on how these joint tests are implemented. For example, the Study 3 Design Table now reads:

> **Approach**: For H2 and H3, we will conduct a joint Wald test on the coefficients in the above model. A joint test simultaneously assesses multiple related hypotheses, allowing us to determine whether the playtime effects for any of the 23 genres differ significantly from zero (H2) and from each other (H3). The joint test the estimated coefficients and their covariance matrix to determine if a set of parameters jointly equals specified value; this test follows the chi-squared distribution. The error rate is controlled in a similar manner as would be achieved by correcting the alpha level for all 23 surrogate hypotheses.
>
> **Tests**: For H2, we will test whether any genre's playtime is associated with changes in mental wellbeing by evaluating if at least one genre-specific coefficient is different from zero. we assess the probability of the data given the null that the genre coefficients as a group are not statistically different than 0 vs the alternative that at

least one is non-zero. In R, this is specified as:

hypotheses(model, joint = "_within", hypothesis = 0)

For H3, we will test the joint hypothesis that all genre-specific coefficients are identical:

linearHypothesis(model, genre1 = genre2, genre2 = genre3, genre3 = genre4 .... genreN-1 = genreN)

In our view, this makes the Study 3 modelling clearer and more internally consistent. This allows us to make a broad interpretation about whether genre-level playtime is related to mental health, and whether genres differ from each other. This sets the stage for potential exploratory post hoc analyses about how particular genres differ.

> Based on the above, I have almost no recommendations for the authors. The planned instruments, process of data collection and statistical analysis seem as appropriate for the stated aims and hypotheses.

We're glad that our approach to data collection and analysis is in line with the study objectives and meets your expectations.