Thank you for the opportunity to review this interesting proposal. Below I’ve left some comments relevant to the key criteria, and a few thoughts that I hope might be useful.

Good luck with the research!

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

I have no concerns with the validity of the proposed research questions.

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

H2 is a little problematic. It begins with “All video games containing loot boxes…” but I’m not sure the researchers do not intend to identify all games containing loot boxes on the Google Play Store. Instead, they intend to sample 100 (or potentially fewer) games. This will not allow the researchers to test H2: All video games containing loot boxes on the Google Play Store will accurately display the IARC ‘In-Game Purchases (Includes Random Items)’ label. Even if all 100 (or fewer) of the sampled games accurately display the appropriate warning, this cannot support the conclusion that all games containing loot boxes are accurately labelled. The simplest solution here seems to be to re-word H2. And, in fact, it seems your interest is only in games previously known to contain loot boxes (i.e., not all video games containing loot boxes). This may also require a re-wording of RQ2 to indicate that the intended scope of the question relates only to games previously known to include loot boxes (rather than all video games on the Google Play Store that contain loot boxes). This will also require some consideration of the representativeness of the sampled games (see below).

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

For Study 1, the proposed methodology appears appropriate. However, I’d request clarification on two points.

First, and apologies if I’ve misunderstood something, but I was unsure why the consistency rate is calculated as “1 – (games rated consistently across systems/total games rated)”. It seems the consistency rate would simply be (games rated consistently across systems/total games).

Second, I’d like to see a justification for the proposed cut-off of ~95%, as opposed to say 100% or 90%, for accepting H1?

For Study 2, as identified above, the proposed methodology cannot test Hypothesis 2. Thus, some re-wording of the hypothesis / research question, or some change to the methodology will be required. However, it also seems like explicit justification is required for focussing on previously-identified games containing loot boxes rather than looking for games that *currently* contain loot boxes. Is labelling compliance for games that have previously and publicly been identified as containing loot boxes likely to be representative of labelling compliance for new games? It might be, but I think some explicit consideration of the representativeness of the sample (games previously containing loot boxes) for the population of interest (all games containing loot boxes) is needed.

I’m also not convinced games such as Minecraft (i.e., those with substantial third-party content which may or may not include loot boxes) should be assumed to contain paid loot boxes, or treated as containing loot boxes by default. Many players can engage with these games without the “loot box” component: Action needs to be taken by the player (i.e., purchasing or downloading additional content) in order for these game mechanics to be present).

Similarly, I’d be cautious about treating loot boxes and social casino game content as equivalent. The author is correct that both fall under the umbrella definition of “transactions with randomized elements”, but I’m not sure I agree that these should therefore be lumped in together (i.e., as an example of a loot box in a video game). They are conceptually distinct. At the least, I would recommend some clarification in the reporting of these results to discriminate between games containing what would commonly be understood as loot boxes and games containing “social casino” activities.

**1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.**

I believe this criterion has been met for both of the proposed studies (with some relatively minor clarifications as suggested elsewhere in this document).

**1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).**

As mentioned previously, I believe that the methodology for Study 2 (as currently proposed) cannot address Hypothesis 2 (as currently worded).

One further, possibly minor, thing to note. The author notes that for Study 1, they’re at Level 3 of bias control – meaning they’ve not yet observed any part of the data/evidence. In a sense this is true: they have not analysed any data yet. However, given the data they have provided relating to the number of games identified based on ESRB and PEGI lists, a little mental arithmetic makes it plain that RQ1 can be answered based on what is known (i.e., it appears to be impossible that H1 will be supported). This seems to be equivalent to Level 1 of bias control: the answer to one research question is, to some extent, already known.

Specifically, there were 17 titles identified from the ESRB list and 64 titles from the PEGI list. Thus, it seems the numerator (number of consistent cases) could range from 0 (if none of the 17 ESRB items were also in the PEGI list) to 17 (if all 17 ESRB titles were also in the PEGI list), and the denominator (number of titles included in both lists) could range from 47 (if there was complete overlap) to 81 (if there was no overlap). This would mean the outcome could range from 0/81 to 17/47. That being the case, it seems impossible that H1 will be accepted (which requires ~95% consistency). This seems more akin to Level 1 of bias control.