Dear Prof Chambers and the Managing Board,

Thank you for the opportunity to submit a revised version of this stage 1 manuscript to *PCI Registered Reports*.

I include the complete text of Prof Chambers' recommender comments and the two reviews by Prof Etchells and Dr Sauer in *black italics*; my point-by-point responses in purple; and amended or newly added text in blue or red for emphasis below.

Leon Y. Xiao

--

Recommender Comments by Prof Chris Chambers, 18 Nov 2022 09:40

Major Revision

I have now obtained two very helpful reviews of your submission. As you will see, both evaluations are cautiously positive while also noting various aspects of the design and rationale that would benefit from clarification or modification. Key issues to address include the testability of the hypotheses, the timeframe of the data extraction (with the useful suggestion by Etchells to use a broader and more principled range), and the justification of specific assumptions and elements of the rationale.

Response 1: Thanks to Prof Chambers for arranging scheduled review for this submission. I am very grateful for the very helpful and constructive feedback and how promptly they have been provided.

I address all elements of the two reviews point-by-point below. Notably, I have amended RQ2 and H2 (and the Study Design Table) per Dr Sauer's suggestion; changed the timeframe for data extraction to up to approximately when the labels were introduced per Prof Etchells' advice; and proposed to separate out *Roblox* and *Minecraft* for dedicated scrutiny and discussion.

One particular issue that requires careful attention is whether the answer to RQ2 is a foregone conclusion given known information (noted by Sauer). In order to be eligible for consideration as a Registered Report at PCI RR, the conclusions of the research must not be known (or inferrable with certainty) before the study is conducted. Please consider this issue carefully and, in turn, whether the bias control level for your submission is set appropriately.

Response 2: I have addressed this issue in detail in Response 14. In short, I believe particularly given the amendments to the methodology through this revision, Study 1 should be able to claim Level 3 bias control and Study 2 should be able to claim Level 6.

I have added a brief positionality statement following Dr Veli-Matti Karhulahti's suggestion for my previous registered report concerning the loot box 'ban' in Belgium: https://rr.peercommunityin.org/articles/rec?id=264.

Finally, If I may draw Prof Chambers to a question in Response 10 regarding whether the limitations inherent to Study 2's proposed data sampling should be stated in the introduction or may alternatively potentially be allowed to be moved down to the discussion section in due course, in case of in-principle acceptance?

Based on these reviews, I am happy to invite a revision, which I will return to the reviewers for another look

--

Review 1 by Prof Pete Etchells, 15 Nov 2022 11:25

This is a well-considered and straightforward study, and I have no major concerns in general.

Response 3: Thanks to Prof Etchells for taking the time to review this proposal.

Study 1 carries with it a certain level of risk, given that the relevant games lists from both PEGI and and ESRB have already been collated, but I appreciate the necessity of this, and the author is honest in their reporting. However, I do wonder whether the time frame limitation is somewhat arbitrary. While I appreciate that using the entire database (c. 31000 games) is beyond the scope of the study, limiting to the year leading up to September 2022 could be better justified. As such then, I would suggest expanding the time frame scope to capture a broader range of games and be grounded in a sounder rationale. Given that the announcement to include the relevant warning tags came in April 2020, I wonder whether this would serve as a useful time point to start with. Expanding the time frame would also allay concerns that readers may have about the potential for 'data peeking'.

Response 4: I agree with Prof Etchells suggestion. I had previously considered using that cut-off date, but the problem is that one cannot determine when each game was rated (the ESRB does not share this) and thus which was the first game that was rated after the labels were introduced (*i.e.*, the cut-off). This problem does not have a perfect solution, but I propose a potential rough solution. I have amended the methodology for generating the ESRB List to more or less allow for all games that have been rated and granted the label between 13 April 2020 and 21 September 2022 to be identified.

I will concede that the new methodology is also imperfect because the ESRB does not note the exact date on which a rating was granted so I could only estimate how many games were rated between 13 April 2020 and 21 September **2021** (in contrast, I was able to determine exactly how many games were rated between 22 September 2021 and 21 September 2022), but I believe it is the best that could be done and should be acceptable in the circumstances.

The relevant paragraph now careful explains what data I have already collected and what data I will hopefully now go on to collect (the list of games that have been given the label by the ESRB between *approximately* 13 April 2020 and 21 September 2021).

Prof Etchells is right that this may allay some concerns that readers might have regarding data peeking. I do also believe that this change also addresses the relevant concerns raised by Dr Sauer, which I address under Response 14 below.

Study 2 seems well thought through, with a clear processing pipeline for gathering game data. I appreciated the level of detail regarding sampling method, but felt that this became a bit cumbersome with regards to Roblox and Minecraft. The author does an admirable job of providing a clear justification for scenarios where these two games (or a combination thereof) will or will not be included in the final sample, but I wonder whether it just makes sense to include them both anyway. Given the unique nature of these two games in terms of compliance issues, they serve well as test cases for future sandbox games that may run up against third-party content, and therefore would be well-placed to be included in such considerations.

Response 5: I appreciate that Dr Sauer has also pointed out concerns with *Roblox* and *Minecraft*, although I suspect that Dr Sauer might hold the opposite view that these two should not be included in the sample at all and be treated differently (see Response 11).

What I propose to do is to have the (hopefully) 100-game sample exclude *Roblox* and *Minecraft*. But I will assess and report their compliance status separately (specifically, whether they are displaying the label on the Google Play Store). I think this both highlights that they are quite uniquely important (as Prof Etchells suggested) and also that they are a bit different from other games with randomised monetisation mechanics (as Dr Sauer suggested).

--

Review 2 by Dr Jim Sauer, 17 Nov 2022 22:14

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!

Response 6: I am grateful to Dr Sauer for providing detailed feedback on this stage 1 submission. I note at the onset that, since my original submission, Dr Sauer and his collaborators have published Garrett *et al.* (2022) testing these labels under experimental conditions. I have added references to that study at appropriate places. I am hoping to address different questions, so I believe these two studies would complement each other in due course.

> Garrett, E., Drummond, A., Lowe-Calverley, E., & Sauer, J. D. (2022). Current loot box warnings are ineffective for informing consumers. *Computers in Human Behavior*, 107534. https://doi.org/10.1016/j.chb.2022.107534

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).

Response 7: Thanks to Dr Sauer for pointing this out. I have reworded as follows. I have also amended the Study Design Table accordingly.

> Research Question 2: Are video games **previously known to be high-grossing and contain loot boxes** and presently containing loot boxes on the

> Google Play Store accurately displaying the IARC 'In-Game Purchases (Includes Random Items)' label?
>
> Hypothesis 2: Video games **previously known to be high-grossing and contain loot boxes** and presently containing loot boxes on the Google Play Store will accurately display the IARC 'In-Game Purchases (Includes Random Items)' label.

I address the representativeness point below under Response 10.

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).

Response 8: Dr Sauer is correct. This was an oversight on my part and has been fixed.

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?

Response 9: This was an issue that also came up with my previous registered report in Belgium regarding the loot box 'ban': there is no objective basis for determining what is a relevant cut-off for regulatory / compliance situations. I note that this also applies to H2, where I have suggested similar cut-offs.

In all honesty, I do not think I can come up with a better justification than merely to say that I think that this is fair and that I think a policymaker would think similarly. I detailed how I will differently interpret the results if they fall within '⩾ 80% but < 95%' and '< 80%' to preregister how I will later speak about the data. Others naturally could interpret the results differently and apply other cut-offs. These are what I did with the Belgian study, although, of course, I am open to improving on this approach if possible.

I have added the below following both instances where these cut-offs were mentioned:

> These cut-offs and corresponding interpretations were based on the author's own opinion on what is a 'satisfactory' self-regulatory measure and what he deemed most policymakers would agree with.

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.

Response 10: I have reworded RQ2 and H2 as detailed in Response 7. I have now added the justification (mainly efficiency) for selecting the sample thusly and the relevant limitations in red.

> Rather than to assess the 100 presently highest-grossing Google Play Store games as to whether they contain loot boxes (as previous studies have done[18,20,34,35]) and then to check whether they are displaying the label, it is more economical and efficient to instead examine games previously known to contain loot boxes. If a game that was known to contain loot boxes is displaying the label, then it is no longer necessary to assess whether said game still contains loot boxes through gameplay, as this can be reasonably assumed. Only those games previously known to contain loot boxes but are not displaying the label need to be re-assessed through gameplay. This expediency is desirable because it is hoped that the present study's results could be published promptly and thereby contribute to the efforts of the UK Government's Department for Digital, Culture, Media & Sport's technical working group that is developing industry self-regulation for loot boxes with the aim of reducing harm[47]. The sample selection (as detailed below) will be based on previously highest-grossing games (many of which will likely still remain high-grossing and popular games presently)[18,20,34,35]. This therefore represents a sample of particular interest for players, parents, policymakers, and the age rating organisations. However, some limitations should be noted. Firstly, the compliance rate amongst this sample of historically (and potentially presently) high-grossing games is not necessarily representative of that of financially worse performing games (which might be less scrutinised by players and other companies and therefore less likely to comply or, contrastingly, might be performing worse financially because they have accurately displayed the label) or the overall situation on the Google Play Store. Secondly, these games were previously highlighted in published academic work as having contained loot boxes[18,20,34,35], and, therefore, their operating companies might have since become more likely to comply (when

> compared to a newly published game that has not yet gained any notoriety), as companies have reportedly complied with the Belgian 'ban' on loot boxes only following the publication of Xiao and media reporting thereof[35,47] and four years after they were originally supposed to have done so.

This is more a note to the Recommender, Prof Chambers: may I potentially reserve the right, if the submission is given in-principle acceptance, to move the limitations with the sample selection (marked in red) down to the discussion section in due course where it might be more appropriate? I think that would be where this section would have been written (and would have been expected to be seen) had this not been a registered report.

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).

Response 11: As I detailed in Response 5, I propose now to **not** include *Minecraft* and *Roblox* in the general sample and to instead consider them separately. I hope this goes in some ways to alleviating Dr Sauer's concerns.

As to assuming that both games contain loot boxes, this was a decision made because of practicality. *Roblox* publicly admits that loot boxes are implemented by third parties: https://devforum.roblox.com/t/guidelines-around-users-paying-for-random-virtual-items/307189, although *Minecraft* does not do so. I do not think it would be worthwhile to 'prove' a point that is well-known. I do appreciate that there are differences between *Minecraft* and *Roblox*. The latter appears to encourage more loot box engagement than the former, as loot boxes are naturally in many *Roblox* sub-games, but I do believe engagement with a private server or purchasing from the Minecraft Marketplace might be the 'easiest' way to engage with loot boxes in *Minecraft* (https://www.minecraft.net/en-us/marketplace/pdp?id=9443a18e-935b-4062-b732-4b0fcdf7df38). I propose to bring such points up in the discussion in a separate paragraph dealing with games allowing user-generated content-type loot boxes.

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.

Response 12: I appreciate Dr Sauer's perspective. I think it would be difficult to use an alternative shorthand term than 'loot boxes' to refer to all 'transactions with randomized elements' in this paper (and in loot box studies in general). The latter term is too unwieldy. I have debated publicly on this distinction with Dr Zendle and his team in *Addiction*. Our letter: https://doi.org/10.1111/add.15829; their response: https://doi.org/10.1111/add.15976.

My own view is that this is perhaps not a binary matter. Traditional social casino games appear relatively easy to separate out from 'traditional' loot boxes; however, there have been reports of social casino type slot machines in city-building-type mobile games that provide virtual currency, and I have seen one social casino game (simulated slots) contain traditional loot boxes for a card collecting sub-system. I think the boundary can begin to blur. We have in the literature used the term 'loot boxes' to describe mechanics that are not a 'box' (*e.g.*, virtual card packs). I will note here that the ESRB and PEGI have adopted their own way of addressing this issue with their label 'contains random items,' which perhaps is more confusing than just using 'loot boxes,' but that is an empirical question for another day.

Regardless, I have shored up the definition for a social casino game and added the following to ensure that I will report the compliance rate for both 'social casino games' and traditional 'loot boxes,' as I did for the study on the Belgian 'ban':

> However, the relevant compliance rate (see below) amongst 'social casino games' (which will be identified using the definition above) and non-'social casino games' will be additionally separately reported to provide nuance.

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).

Response 13: I believe this should now have been addressed given the changes and additions detailed in Responses 7 and 10.

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.

Response 14: Thanks to Dr Sauer for pointing this out. I found engaging with this point very interesting.

The reason that the ESRB List (as it originally was) contained fewer games than the PEGI List was that the PEGI List covered a significantly longer period of time. The ESRB List ran from September 2021 to September 2022 (12 months), whilst the PEGI List was from the 'beginning of time'/likely April 2020 to September 2022 (29 months). I would suggest that the ESRB List (as it will now be; running from approximately April 2020 to September 2022) should contain a similar number of games as the PEGI List, so it cannot be said that H1 could not possibly be supported.

Under the old methodology, I would also suggest that the numerator would have potentially ranged from 0 to potentially 64 because a game that did not appear on the ESRB List (as it then was; *i.e.*, not one of the 17 ESRB games) could have still had the label attached by the ESRB but it did not appear on the ESRB List (as it then was) because it was rated prior to 21 September 2021. In other words, although the ESRB List contained only 17 games, it did not mean that the ESRB only gave the label to those 17 games. This would also apply to the new ESRB List, as it might be incomplete: it is possible that a game that should have been on the ESRB List and but was not would later be identified through the PEGI List when said game is searched for using the ESRB search tool.

In any case, I believe under the new methodology of expanding the ESRB List to start from approximately April 2020 (and I have not created said list yet and do not intend to do so until this protocol is hopefully approved, and data collection properly commences) will mean that this should no longer be an issue.

Given that I would no longer have the complete ESRB List, I therefore believe that Level 3 bias control will be achieved here. Please do correct me, if I am wrong.