Dear Dr Karhulahti and the Managing Board,

Thank you for the opportunity to submit a revised version of this manuscript to *PCI Registered Reports*., which has now been uploaded to the Open Science Framework and is available at this link: https://osf.io/f3yab/?view_only=33e5875516d144ed98509c7871242b31. I include the complete text of Dr Karhulahti's comments and the three reviews by Drs Moshirnia, Macey and Chin in *black italics*; my point-by-point responses in purple; and amended or newly added text in blue below.

Leon Y. Xiao

--

Dr Karhulahti's Recommender Comments

Dear Leon Xiao,

Thank you for submitting your Stage 1 manuscript to PCI RR. To my knowledge, this is the first RR in the domain of law, and as such a highly interesting manuscript to handle. I have now received all three reviews, collectively representing expertise of gaming and law as well as the related methods. The reviews are very positive, but also highlight issues that need revision. In general, the reviewers are consistent and do not express conflicting views, for which I will merely follow-up on some of their points and add a few comments of my own. I start by moving chronologically through the MS with minor issues, and in the end, I discuss some bigger methodological issues.

Response 1: I would like to begin by thanking Dr Karhulahti and the three reviewers. I really appreciate how promptly I have received such fair, detailed, and reasoned feedback. I attempt to address every concern that has been raised below.

1. On page 2, I would remove the part "rather than, e.g., wealthy players" because wealthy players can also be at-risk players. Later the same page, a parenthesis is not closed ("and therefore be…").

Response 2: I have removed the reference to wealthy players and closed the parenthesis.

2. On page 5, you introduce Netherlands for the first time, while previously addressing only Manx and UK law. It would improve readability to briefly note earlier that Netherlands is also a candidate (yet its role differs from the other two).

Response 3: I note that references to the Netherlands had to be revamped due to a court decision that was published on 9 March 2022, after the reviews were sent to me and during the revision process. Basically, loot boxes that were previously deemed illegal are now confirmed to be legal. This means that the Netherlands would no longer even be a potential candidate to study now. However, I still provide updated information on the Netherlands for context and why I decided against using it previously (p.3 and p.6).

> The first type constitutes gambling under existing law in many countries, as recognised by various European national gambling regulators, including in the UK, Denmark, and Belgium[41–44], although only the Belgian regulator has actively enforced the law[11]. In contrast, the Dutch gambling regulator also previously opined that the first type constitutes gambling [45] and has enforced the law by imposing a financial penalty on Electronic Arts for allegedly illegal loot box implementation in its *FIFA* games [51,52]; however, that interpretation has since been successfully appealed and was overruled by the highest Dutch administrative court. Therefore, the Netherlands is the first country where the first type of loot boxes has been confirmed not to constitute gambling.
>
> …
>
> Secondly, the legal position in relation to loot boxes in the Netherlands changed in March 2022 [71]. Previously, the Dutch gambling regulator *incorrectly* interpreted the law and has actively enforced existing gambling law to regulate the first type of loot boxes by sanctioning allegedly non-compliant companies (specifically, imposing a financial penalty on Electronic Arts for allegedly illegal loot box implementations in its *FIFA* games[51,52]). This is unlike other countries (*e.g.*, the UK) whose regulators came to the same interpretation of their gambling laws but have not sought to take enforcement actions against potential contraventions. The present Dutch position is that the first type of loot boxes are confirmed to be generally lawful [71]. The Dutch Apple App Store would therefore likely be experiencing change to reflect that new regulatory position, which would render it inappropriate to study for answering the present research question. Even assuming that the regulatory change did not take place, it would have been appropriate to study the Netherlands because the previously enforced Dutch regulation focused on the presence of the ability for players to transfer loot box rewards to other players in exchange for real-world money [10,11,38]. A previous loot box prevalence study attempted to assess the presence and prevalence of this so-called 'cashing out' process: however, Zendle *et al.* (2020) importantly failed to reliably do so [4], possibly due to video game companies actively preventing this from happening such that the availability of third-party cashing out platforms is extremely transient. Even if the presence of cashing out features could have been reliably assessed, the previous Dutch regulatory position meant that only a reduction in the prevalence of 'cashing out' features would have been observable and that a reduction in paid loot box prevalence was not necessarily observable and, indeed, highly unlikely to have been true because the removal of paid loot boxes was not legally required. This

> contrasted with Belgium, where a reduction in paid loot box prevalence should be observable as an outright removal of the feature is required to comply with the law, as compared to only amendments to a certain aspect of some loot boxes' implementation that Dutch law previously required. This is demonstrated by how the same video game company removed paid loot boxes entirely from a game in Belgium [56], but did not remove paid loot boxes from the same game in the Netherlands and only changed them such that cashing out is no longer possible [72].

3. On page 6, you introduce (i) and (ii), but as one reviewer points out, they are not stated as RQs. Please reformulate them into explicit RQs. When doing that, carefully ensure that your hypotheses/methods match with and able to answer the RQs. If you have a reason to do otherwise, please explain in the response letter.

Response 4: Thanks for highlighting this. I have now included the following to clearly set out each of five research questions (pp.6–7):

> "The following research questions will be addressed.
>
> > Research Question 1: Has the Belgian ban on paid loot boxes been effective, such that virtually no Belgian video games contain paid loot boxes?
> >
> > Research Question 2: Has the Belgian ban on paid loot boxes been effective, such that virtually no Belgian video games deemed suitable for children contain paid loot boxes?
> >
> > Research Question 3: Has the Belgian ban on paid loot boxes been effective, such that the prevalence rate of paid loot boxes in Belgium is different from that of another Western country which has not restricted loot box sales?
> >
> > Research Question 4: How effective has the Belgian ban on paid loot boxes been at reducing the prevalence rate of paid loot boxes in Belgium?
> >
> > Research Question 5: Is it possible for a player to circumvent the Belgian ban on paid loot boxes and purchase them from within the country?"

4. On page 7, you note how the Belgian Gaming Commission will be contacted and their response discussed. I expect the response will take time and it is possible that you will not have it by the time of Stage 2 review; thus, I suggest obtaining a permission to share their response publicly and storing it in the OSF when it comes (if after Stage 2). Alternatively, if *no permission for sharing is gained, you could add a summary to the OSF so that future readers will find the information via DOI.*

Response 5: This makes sense. I have amended the relevant passage at p.8:

> "… Permission to publish the Commission's response, if any is received, will be sought, and a summary will be made available at the data deposit link (<[OSF deposit link]>)."

5. On the same page, a small typo (just before "because")

Response 6: This has been fixed.

6. Methods: as the reviewers point out, the time of data collection is very critical. If possible, I would suggest finding out the Top 100 list in Belgium for the compared time (June 2021). While post hoc analysis is not possible, it would at least allow the reader to assess the fluctuation of the titles on the list (and perhaps you to address that briefly in the discussion, if relevant).

Response 7: This is sensible. I did not capture the highest-grossing list for Belgium in June 2021 at the time. I have tried to do this just now, but unfortunately my free access to data.ai/App Annie and other analytics companies' data does not allow me to have access to the historical highest-grossing list in Belgium on 21 June 2021 (the exact Xiao et al. UK data collection date). I have requested for free academic access to the data from App Annie and Sensor Tower and am waiting to hear back.

A less ideal, but workable, solution is also available in the alternative: thanks to Dr David Zendle who captured this information, I actually have access to the Belgian ***Android*** top ***50*** highest-grossing lists on 21 June 2021 (captured at around 14:00 on that date). This is not the iPhone list nor is it the top 100, but it should be very close, given that Zendle et al. (2020) previously found near identical loot box prevalence rates on the Android and iPhone markets.

I will in the Discussion section mention how different the Belgian lists on each date are by reporting how many games are identical (*i.e.*, the overlapping rate). I will upload the relevant data to OSF.

7. On page 10, you say that "game will be assumed by the coder to contain paid loot boxes without the need for the coder to identify and screenshot such a mechanic." I might have missed something, but I don't see how third-party involvement would automatically ensure that loot boxes are present. E.g., if there is a known avenue for generating paid loot boxes in sandbox games that cannot be interfered by companies, please cite that.

Response 8: Specifically, this relates to Roblox and Minecraft (which likely will be in the highest-grossing list in due course because of their popularity).

In Roblox, the company recognises that loot boxes can be implemented by third parties, and the company will moderate such user-generated content. There are rules requiring loot box probability disclosures by Roblox, for example: https://devforum.roblox.com/t/guidelines-around-users-paying-for-random-virtual-items/307189.

However, it is not known whether I would be able to find any official rules specific to Belgium published by either Roblox or Minecraft stating that loot boxes may not be implemented by third parties in Belgium. My proposal is that if I find such a requirement for Belgium, then the game will be deemed 'compliant' in so far as they have taken some compliance action (even though paid loot boxes will probably still be available somewhere in the game if I looked hard enough). However, in contrast, if I cannot find such an official requirement, then the game will be deemed 'non-compliant' because the developer/publisher has not sought to at least attempt to enforce the law.

That is the approach we adopted when deciding whether Minecraft and Roblox are compliant with loot box probability disclosure requirements in the UK.

I have added the following clarification (p.12): 'However, the game will be deemed compliant with the law and coded as not containing paid loot boxes if…'

Studying the loot box implementations in these two games in-depth is complex and is probably a project on its own. I realise that what I have proposed and adopted in the UK is arguably an imperfect solution to a very unique situation that would arise only in relation to very few games, so my alternative proposal, if the Recommender and the Reviewers are unsatisfied with the above, would be to **exclude such games from the sample and replace them with the next highest-grossing game**.

8. *On page 13, you have the ethics statement "No ethics approval will be required because the present study examines and records publicly available information." Please elaborate, according to what university/country. Different universities and countries have difference ethics policies (e.g., according to the ethics policy of IT University of Copenhagen and the Danish Code of Conduct for Research Integrity, the study did not require ethics assessment).*

Response 9: I have amended as follows (p.17):

> "In accordance with the *Danish Code of Conduct for Research Integrity*[75], as adopted by the IT University of Copenhagen, the present study did not require research ethics assessment and approval because no human participants or personal data are involved and only publicly available information is examined and recorded."

Methods

a) I notice that the prevalence of 0.77 comes from a preprint that has not been reviewed yet. I am flagging it because if the paper remains un-reviewed at Stage 2, or if its peer review ends up affecting the result 0.77, there is a chance that this RR would have to be rejected based on criterion 2B (changes in hypothesis). As it is your own co-authored paper, we can proceed without changes; however, you should be aware that at Stage 2, if the results of the cited paper are still pending, we possibly cannot provide IPA.

b) Related to the above, I can see that in a previous study Zendle et al. (2020) found 0.59 in the UK, and this number is from 2019. Although I understand that you prefer to use a more recent prevalence rate, we do not have evidence that the change is due to time alone, i.e., there can be variation in samples for other reasons, too. Considering that your to-be RQs are interested in whether Belgium has a lower rate vs other countries, and previous studies have found UK 0.59/0.77, Australia 0.62, and China 0.91, a bit more justification is needed why 0.77 has been selected, and as one reviewer points out, why would 0.4 be a remarkably low prevalence.

Response 10: Thanks to Dr Karhulahti for bringing up these two points, which I will address together. For context, I note that we have rapidly communicated the 0.77 UK prevalence rate here in a Letter to the Editor (which I believe was refereed, although I have not confirmed this) in the journal where the Zendle et al. (2020) results were originally published:

> Xiao, L.Y., Henderson, L.L., & Newall, P.W.S. (2022). Loot Boxes are more prevalent in United Kingdom video games than previously considered: Updating Zendle et al. (2020). *Addiction*, _(_), __. [Advance online publication]. https://doi.org/10.1111/add.15829.

We wrote that Letter to the Editor following peer reviews of the original preprint that did not raise concerns regarding the methodology or the data.

The preprint itself is, of course, under review, but I do not know when it would be published. I am unsure whether that letter to the editor could be deemed as sufficient prior publication as to justify the use of the 0.77 result for the purposes of Stage 2 IPA, hypothetically.

With the said, I note the general concern (aside from whether the data comes from a refereed publication) about using the 0.77 result, which appears comparatively quite high when the other studies are considered and may render it relatively 'easy' for a Belgian result to be significantly lower.

I personally think that the 0.59 UK result was lower than the then true value due to potential errors (or at least conceptual disagreements): in the abovementioned letter to the editor, we noted the potential reasons why the differing 0.59 and 0.77 results might have occurred. This includes that, in our view, (a) a few false negatives (at least 3% of the sample because Zendle *et al.* used a different methodology for detecting paid loot boxes) and (b) a certain type of paid loot boxes (social casino games) representing a non-trivial percentage of the sample (5% of the sample) was not recognised as paid loot boxes by Zendle *et al.* (although Belgian law would likely view them as paid loot boxes).

The 0.62 Australian result was based on all video games and not games on the mobile platform specifically. p.28 of that Australian report (https://doi.org/10.25946/5ef151ac1ce6f) states: 'A total of 82 video game titles were selected for the environmental scan, which was conducted between August 2019 to October 2019. Video games were selected based on several data sources to identify which were the best selling amongst Australians in 2019. The final list of 82 games reviewed in the environmental scan are provided in Appendix B, along with the data sources.' We know from prior research, specifically, Zendle et al. (2020) that loot boxes are more prevalent on mobile platforms, so the 0.62 does not appear to be the best comparator either as the true value on the Australia mobile platform will likely be higher.

My proposal therefore is to use a hypothetical 0.65 value as a comparator. This approach would avoid issue (a) that Dr Karhulahti identified in relation to needing to change the hypothesis if the preprint is yet to be published, which might lead to a stage 2 rejection. This approach would also more fairly represent the various lower loot box prevalence rates found by the prior literature.

I justify this 0.65 value as follows (pp.13–14):

> "The hypothetical 65.0% figure is derived from a holistic consideration of historical loot box prevalence rates in other countries found by the prior literature. Zendle *et al.* (2020) found the UK iPhone loot box prevalence rate amongst the 100 highest-grossing games in February 2019 to be 59.0%[4]; Rockloff *et al.* (2020) found the Australia loot box prevalence rate amongst the 82 'best selling' games on various platforms (*e.g.*, PC, console, and mobile) between August and October 2019 to be 62.0%[2]; Xiao *et al.* (2021) found the Chinese iPhone loot box prevalence rate amongst the 100 highest-grossing games in June 2020 to be 91.0% [3]; and Xiao et al. (2021) found the UK iPhone loot box prevalence rate amongst the 100 highest-grossing games in June 2021 to be 77.0% [5]. The comparatively high Chinese 91.0% prevalence rate appears to be an outlier that has been influenced by Far East Asian cultural factors that would not affect a hypothetical Western country that has not regulated paid loot boxes; therefore, little reliance is placed on that datum. The Rockloff *et al*. Australian 62.0% is derived from games on various consoles, whilst it is known that games on mobile platforms (*e.g.*, the iPhone platform which the present study will assess) tend to contain more loot boxes[4]; therefore, the 62.0% value might not reflect the contemporaneous Australian loot box prevalence rate amongst mobile games specifically, which likely would have been higher. A comparison of Zendle et al.'s 2019 UK data with Xiao et al.'s

> 2021 UK data suggest that the loot box prevalence rate have increased due to a variety of reasons, including that the 2019 59.0% datum might have been an underestimation, due to certain paid loot box implementations not having been recorded [75]. Xiao et al.'s 2021 77.0% figure is the closest comparator for the present study, in terms of data collection time; however, in context, it is comparatively higher than other values previously observed in Western countries. Accordingly, a hypothetical value of 65.0%, which is slightly higher than the previously observed Zendle et al. UK 59.0% and Rockloff et al. Australian 62.0% values (which were likely slight underestimations), but which is lower than the comparatively high Xiao et al. UK 77.0% value, will be used. This 65.0% value errs on the side of caution and avoid potentially overestimating the effect of the Belgian ban, although unavoidably it is possible that the effect might consequently be underestimated."

c) There in issue with the method for testing H3. You are planning to use the binomial test, which assumes that the Belgian Top 100 list provides a random variable, but the compared UK Top 100 list is fixed. Yet since both lists produce outcomes as similar random variables, what you seem to need for testing is a 2x2 contingency table.

Response 11: In light of the revised decision to use a hypothetical value of 0.65 as the comparator, I believe the binomial test will now be appropriate. Please do correct me, if I am wrong.

d) I also encourage thinking about the comment from one reviewer regarding the effect, i.e. what effect would be societally beneficial for a regulation like this to be useful in practice. Depending on how you proceed, you will then also need to recalculate power, depending on what your final RQ + hypotheses + method is. Please note that PCI RR does not demand any particular power as long as it is justified; however, some journals do, so you should double check that if you have a specific journal in mind (see the next point).

Response 12: Recognising also Dr Chin's comments below on this point, which I address in Response 4 to Dr Chin, I have decided to amend the test that will be used (p.13; pp.14–15). A two-sided test would allow me to draw conclusions both ways.

> "Hypothesis 3 will be tested using a binomial test (two-sided test, $p = .05$) to identify whether the percentage of the 100 highest-grossing iPhone games containing loot boxes in Belgium that will be found by the present study will be significantly different from a hypothetical loot box prevalence rate of 65.0%, which a Western country that has not restricted loot box sales is assumed to have. Then, if a significant difference is found and the Belgian loot box prevalence rate is numerically lower than 65.0%, a binomial test (one-sided test, $p = .05$) will be used to identify whether the Belgian rate is significantly lower than 65.0%. Alternatively, if a significant difference is found and the Belgian loot box prevalence rate is numerically higher than

> 65.0%, a binomial test (one-sided test, p = .05) will be used to identify whether the Belgian rate is significantly higher than 65.0%."
>
> …
>
> "If a statistically significant difference is found, then Hypothesis 3 is confirmed. As to interpretation, if the Belgian value is significantly different from and significantly lower than 65.0%, then the present study will conclude that it is *possible* that the Belgian 'ban' may have been effective at reducing paid loot box prevalence in Belgium and that this measure could be considered for adoption in other countries, although it must also be recognised that national differences between Belgium and the previously assessed Western countries (*i.e.*, the UK and Australia), and the passage of time between the data collection points, may also have contributed to the results. The present study will then recommend other countries' policymakers and regulators to consider adopting a similar measure if they desire to reduce paid loot box prevalence rates in their country: how strongly this recommendation will be put by the present study depends on the results of Hypotheses 4 and 5, as detailed below. In contrast, if the Belgian value is significantly different from and significantly higher than 65.0%, then the present study will conclude that the Belgian ban has been ineffective, noting the same abovementioned limitations. The present study will then caution against other countries' policymakers and regulators from making the assumption that a loot box ban will necessarily be effective, and conclude that the Belgian measure was likely ineffective and should not be adopted by other countries unless effective enforcement can be guaranteed or some other improvements are made. Further, reasoned criticism of the apparent lack of enforcement actions by the Belgian Gaming Commission will also be made."

I will consider a loot box prevalence rate reduction from 65% to 50% as indicating some benefit and have accordingly reconducted a prior power analysis (p.14).

> "In the absence of any prior guidance on what effect size would constitute a 'legally meaningful' and 'socially beneficial' regulatory measure, based on intuition, it is proposed that a reduction from the abovementioned hypothetical 65.0% loot box prevalence rate to 50.0% or lower in Belgium would justify a law researcher to argue in favour of the Belgian Gaming Commission's regulatory enforcement actions as having been effective at providing improved consumer protection in comparison to other countries that have taken no regulatory actions. Accordingly, setting the Hedges' g at −.15, a priori power analysis using G*Power has determined, given an α value of .05: the present sample of 100 games would achieve .86 power in a two-sided test for finding a statistically significant difference between the Belgian and the hypothetical 65.0% prevalence rates (see Fig. A1) [73]."

Further to what Dr Karhulahti and Dr Chin say regarding what reduction will be viewed as 'effective,' I have decided to add a new research question to compare the loot box prevalence rate that will be found in Belgium against 50% and 25% respectively. These values are derived based purely on intuition as, to my knowledge, there is no guidance.

The tests will only be run in the event that the Belgian loot box prevalence rate is below 50%. This would help me to preregister whether and how I will interpret, as a legal researcher, whether the measure as either 'effective' or 'very effective' and how I strongly I can suggest the adoption of this measure in other countries. This is to prevent me from mentally 'shifting the goal post' when writing the Discussion section based on the found results and thereby biasing the eventual presentation of the results.

> "Research Question 4: How effective has the Belgian ban on paid loot boxes been at reducing the prevalence rate of paid loot boxes in Belgium?"

The following hypotheses have been added (p.7):

> "Hypothesis 4: The paid loot box prevalence rate amongst the 100 highest-grossing iPhone games in Belgium will be lower than 50%.
>
> Hypothesis 5: The paid loot box prevalence rate amongst the 100 highest-grossing iPhone games in Belgium will be lower than 25%."

And paragraphs added to the Method section (pp.15–16):

> Hypothesis 4 will be tested using a binomial test (one-sided test, $p = .05$) to identify whether the Belgian loot box prevalence rate that will be found by the present study will be significantly lower than 50.0%.
>
> Hypothesis 5 will be tested using a binomial test (one-sided test, $p = .05$) to identify whether the Belgian loot box prevalence rate that will be found by the present study will be significantly lower than 25.0%.
>
> One-sided tests are appropriate for Hypotheses 4 and 5 because Research Question 4 is only concerned with the possibility of the Belgian loot box prevalence rate having been reduced by the ban and to what degree that reduction has been. Assuming the loot box prevalence rate in Belgium to be 2%, in line with Hypotheses 1 and 2, a priori power analyses using G*Power have determined, given an α value of .05 and setting the Hedges' g to $-.48$ for Hypothesis 4 and $-.23$ for Hypothesis 5: the present sample of 100 games would achieve .99 power in a one-sided test for finding a statistically significant difference for both tests (see Figs. A2 and A3) [73]. These two tests will only be run and reported if the Belgian loot box prevalence rate that will be found by the present study is significantly lower than 65.0% (as will be

> determined through Hypothesis 3) and numerically lower than 50.0% and 25.0%, respectively. The 50.0% and 25.0% values were chosen based on intuition, due to the absence of any guidance on what reduction would objectively be deemed in law as 'effective' or 'particularly effective.'
>
> Hypothesis 4 is confirmed, if a statistically significant difference is found. The interpretation will be that the measure has been effective at reducing paid loot box prevalence in Belgium. If no significant result is found, then the interpretation will proceed on the basis that the loot box prevalence rate was significantly lower than 65.0%.
>
> Hypothesis 5 is confirmed, if a statistically significant difference is found. The interpretation will be that the measure has been *very* effective at reducing paid loot box prevalence in Belgium. If no significant result is found, then the interpretation will proceed on the basis that the loot box prevalence rate was significantly lower than 50.0%.

e) I would also like to highlight some parts of the conclusions. On page 12, it says: "if no significant difference is found, then the present study will conclude that the Belgian ban did not appear to affect paid loot box prevalence in Belgium, thus disconfirming Hypothesis 3. The present study will then conclude that the measure is likely ineffective and should not be adopted by other countries." However, not being able to find effect is not the same thing as finding evidence for no effect. You cannot conclude ineffectiveness based on non-significance alone. If you want to obtain evidence for no effect, see e.g., Dienes (2021)

Dienes, Z. (2021). Obtaining evidence for no effect. Collabra: Psychology, 7(1), 28202. https://doi.org/10.1525/collabra.28202

Response 13: Dr Karhulahti is correct to point this out. I have amended as follows (p.15; the 'alternative research methodologies' refers to what is discussed in Response 7 to Dr Chin):

> "However, if no significant difference is found, then the present study will state that no sufficient evidence that the Belgian ban affected paid loot box prevalence in Belgium has been found, thus Hypothesis 3 can be neither confirmed nor disconfirmed. Alternative research methodologies for future studies will be discussed."

f) This is a small issue, but in H1 and H2, one reviewer notes how absolute null might not be optimal, and I agree some type 1 error control would be appropriate here. I would suggest keeping it simple, e.g., considering that Zendle et al. (2020) found 1 false positive, you could just double that to be safe and corroborate H1/H2 if more than 2 cases occur (more than 2% prevalence). Although confirming the positives should be rather easy due to the sample size, some control seems reasonable because many things can affect the obtained the sample. If you *wish, you may add that in case H1 or H2 is not corroborated but 1 or 2 instances are found, these games will be investigated in-depth as an exploratory analysis. I also encourage you to consider setting alternative/competing hypotheses, as suggested by one reviewer (but you can choose not to, if you so prefer).*

Response 14: Thank you for highlighting this point, which I appreciate that Dr Moshirnia has also raised. I have amended H1 and H2, as suggested, to include some type 1 error control. Specifically, my wording of 'any one' (*i.e.*, > 0%) has been changed to 'more than two' (*i.e.*, >2%) (p.12). In addition, 'None of' has been changed to 'More than two of' (p.7).

Some amendments have also been made to the abstract and main text to reflect this change.

Finally, I must ask you to revise the table at the end of the MS. Please include all tested hypotheses, and carefully think in each case what can and cannot be deduced from their outcomes. In addition to all the above, please see and respond to the reviewers' respective feedback. Needless to say, if you disagree with some the requested revisions, you are free to justify alternative choices. Do not hesitate to contact me if something is unclear. I look forward to reading the next version, based on which I will see if another external review round is needed.

Response 15: I have revised the table accordingly and added this as Appendix 2. Given that I have made quite substantial changes to the methods in relation to H3–5, I welcome and would indeed appreciate further scrutiny through another round of external reviews.

Sincerely,
Veli-Matti Karhulahti

--

Review 1 by Dr Andrew Moshirnia

Thank you for the opportunity to review this paper at stage 1. I am familiar with Dr. Xiao's work in this area. Overall, I would heartily recommend this experiment, with some slight revisions. I note these below:

Response 1: I thank Dr Moshirnia for his time and attention in reviewing this manuscript. I note merely for the public record (appreciating that these reviews and responses might be published, if the manuscript is eventually accepted) that I do not yet have a PhD and am pursuing one presently.

In reviewing the document, I flagged the assertion that a value out mechanism would constitute gambling under most laws/national codes. This statement should be softened, as the value out mechanism would not become gambling provided that there is guarenteed value of the purchase in any event (this is the legal manuever that renders collectibe card cards legal and non-gambling, even if there is a secondary market in which cards may be exchanged for value).

Response 2: I take Dr Moshirnia's point, and I have amended my wording of 'most, if not all, countries' to 'many countries' instead. I also clarified that the 'national gambling regulators' are 'European' and gave some examples: ', including in the UK, Denmark, and Belgium.' Each of the reference following that assertion is to a European gambling regulator asserting the same position that a value out mechanism would constitute gambling under their public gambling laws.

I believe what Dr Moshirnia highlights as a legal element of 'gambling,' 'injury to property' is a relatively unique legal hurdle in US law, specifically under RICO, to recover civil damage.

I note that in the process of revising this manuscript, Dr Moshirnia has been proven right: the Dutch court has determined that cash-out loot boxes are not gambling, but for a different reason. Portions of the manuscript has been amended (quite significantly) accordingly.

The research question makes sense: an interpretation of law has been announced and in theory compliance should be absolute or near absolute. The hypotheses are perhaps too strict to test this, however, as near perfect compliance (presence of 2% of top 100 containing loot boxes) would return the same hypothesis rejection as complete refusal to comply (presence of 98% of top 100 containing loot boxes). In light of this it may be useful to insert an alternate hypothesis (with a cut off of 2% or 5%) rather than an abolute (as currently stated), because the rejection of the null may lead to less meaningful post-hoc if these alternates are not established before-hand.

Response 3: This makes sense. I have amended H1 and H2, as suggested, to include some type 1 error control. Specifically, my wording of 'any one' (*i.e.*, > 0%) has been changed to 'more than two' (*i.e.*, >2%) (p.12). In addition, 'None of' has been changed to 'More than two of' (p.7).

Hypothesis #3 is a simple comparative and I agree with the approach (one-tailed) and the method.

Response 4: I note that H3 has since been changed and H4 and H5 newly added.

Hypothesis #4 is an interesting question based primarily on what terms of service will control (locality of play or locality of installation). The method may be improved by also setting the locality of the phone to Belgium through the OS, but this may not be needed.

Response 5: This is sensible. I have amended the relevant section to clarify exactly what locality the phone's setting will have at each stage of the study. Indeed, inspired by Dr Moshirnia's suggestion, I will attempt both possibilities (p.13).

> "… With the Country/Region setting of the Apple ID initially set to the UK, the coder will then attempt to access the paid loot box purchase screen and the Apple App Store payment pop-up screen and record their experience from within Belgium. Then, the Country/Region setting of the Apple ID will be changed from the UK to Belgium, and the coding process will be repeated. Thusly, the two potential possibilities of setting the phone's geographic location to either Belgium or a non-Belgian country will be tested. …"

Relevant to this. I also added the following to account for the unlikely event that one of these three games will be removed from the market prior to data collection (p.13):

> "… In the unlikely event that any of these three games becomes unavailable for download and online gameplay (e.g., removed from the Apple App Store), another popular game developed by a company from the same region as the unavailable game; known to contain paid loot boxes in the UK; and in which paid loot boxes represent a fundamental aspect of the game's gameplay and monetisation will be chosen to replace that game. …"

The sample size is sufficient to provide meaningful results.

The use of 1 hour of game play is a reasonable means of arriving at a result, but the author is surely aware that some games' loot boxes function primarily as forgiveness or pity counters (that is, the player is losing frequently so loot box item is offered to increase victory chances). The 1 hour of play should then include deliberate losses by player to solicit these offers (if present). This may account for prior interrater disagreements.

Response 6: Thanks. A false negative is unfortunately never avoidable but acceptable in this context, as explained and justified in the manuscript. With that said, I will ensure that the 1 hour of gameplay will be used efficiently to explore all possibilities.

--

Review 2 by Dr Joseph Macey

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

- No research question is explicitly presented by the author(s), instead the aims of the research are presented in the body of the text, for example:

"Given that there is significant interest in emulating this regulatory approach, it is important to assess whether this Belgian 'ban' on loot boxes has been effective."

And

"… a survey replicating the methodology of previous loot box prevalence studies [3–5] will be conducted in Belgium to assess: (i) the effectiveness of the Belgian Gaming Commission's threat to criminally prosecute video game companies for implementing paid loot boxes without a gambling licence (i.e., the Belgian 'ban') [44] and (ii) whether the loot box prevalence rate in Belgium is consequently lower than in other Western countries where no loot box regulation has been enforced, e.g., the UK. Doing so sheds light on whether the Belgian ban has effectively changed video gaming companies' behaviour."

Whilst the research is both well-designed and justified, it would benefit from the research question(s) being clearly presented, in lines with the requirements of PCI RR.

Response 1: I am grateful to Dr Macey for taking the time to review this manuscript. As I detailed in Response 4 to Dr Karhulahti's Recommender Comments, the research questions are now individually stated.

Although not directly connected to the validity of research questions, I would urge the author(s) to revise the following content:

"The restrictive course of action taken by … who would never have been harmed [58]."

In its current form, the language used clearly reflects the author(s) opinion rather than a neutral assessment of arguments supporting or opposing the Belgian approach (further emphasised by the use of 2 prior papers by the same author to support the statements).

Response 2: Yes, Dr Macey is right in saying that this is my counterargument to a ban on loot boxes that I have not seen expressed elsewhere by others. I think it is fair to make this point in the context of a law paper, especially as this is one of the primary motivations as to why I personally think it would be sensible to conduct this study. However, I appreciate Dr Macey's point. I now provide a more neutral assessment of the situation by adding the following description of the more widely accepted opposing position that the measure is beneficial before my own counterargument is presented (p.4):

> "… Therefore, Belgian players will likely find it more difficult to purchase loot boxes (if they are able to do so at all) than players from other countries who continue to have unrestricted access. Belgian consumers are thereby likely better protected from the potential harms of loot boxes: players who cannot spend any money at all on loot boxes could not 'overspend' and would not suffer potential financial harms…"

For transparency, I would be happy to add that "I have argued elsewhere…" before the section on overregulation (*i.e.*, immediately following the section quoted above), if Dr Macey may think that would improve the fair presentation of the arguments.

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

- The 4 hypotheses are appropriate and the rationale for their development is logical and easy to follow. However, H3 would benefit from some minor revision as the way it is presented may cause confusion to some readers. I would suggest something along the lines of:

"Of the highest-grossing iPhone games, fewer will contain paid loot boxes in Belgium than in the UK."

Of course, the author(s) are free to make any changes, or not, as they see fit.

Response 3: This makes sense. I note that this hypothesis was changed to reflect a change in study design, as detailed in Response 10 to Dr Karhulahti. I have changed this as follows (p.7):

> "Hypothesis 3: Of the 100 highest-grossing iPhone games, the percentage that will contain paid loot boxes in Belgium will be significantly different from the percentage of a hypothetical Western country that has not restricted loot box sales."

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

- The described methodology, including sampling procedures, variables recorded, and analytical approach appears feasible and well-planned. The sample sizes used to address the aims of the research appear to be more than sufficient. However, the justification for selecting the 3 games referenced in H4 could be further expanded. The fact that they represent offerings from game companies in 3 different regions (US, Europe, and China) is appreciated, but the reader would benefit from a more detailed explanation of why these particular games were chosen; are they the highest-ranked examples from each chosen reason (either in terms of number of players, or of revenue raised) or did other considerations guide the author(s)?

Response 4: I have added the following to justify the choices made (p.16):

> "These three popular games were chosen because they have been widely published across the world (including in both the UK and China) and have consistently performed well financially. Importantly, engagement with loot boxes is a fundamental and arguably unavoidable and inalienable aspect of all three games' gameplay and monetisation because the vast majority of in-game content (*e.g.*, playable characters) *requires* engagement with loot boxes to unlock (at least in the UK version of the games)."

- The authors describe the analytical approach and the conditions under which the different hypotheses will be considered to be a) met, or b) rejected. In reference to H3 the authors are frank in their presentation when they discuss how the presented methods cannot offer a clear assessment of a), they state that any conclusion will be discussed in terms of "possibility" that Belgian legislation affected paid loot box offerings. While it is likely impossible that author(s) will be able to access earlier versions, it may be worth supplementing the game analysis with additional analysis of company statements (if any) regarding their reaction to the Belgian legislation.

Response 5: This is a really good idea, and I will keep this in mind. I do not think that I would necessarily have space/word count to include this exploratory analysis in this particular registered report.

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.

- The method of data collection and analysis is clearly presented and comprehensible, allowing replication.

Response 6: Thanks.

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).

- The author(s) state that only a single researcher will be coding the data sample, using prior studies to justify the approach. Given the nature of the analysis required, and the conditions under which coding will be conducted, this is likely to be acceptable.

Response 7: Thanks very much to Dr Macey's for his time and attention.

--

Review 3 by Dr Jason Chin

Title: *Breaking Ban: Assessing the effectiveness of Belgium's gambling law regulation of loot boxes*

Recommendation: *Revise and resubmit*

This stage 1 manuscript seeks to measure the effectiveness of Belgium's rare legal position that makes it illegal to sell loot boxes in online games, even if the loot cannot be resold. This is a significant issue because, among other reasons, loot boxes seem to be a major way in which modern game developers attain revenue and it is generally of interest to know if regulations are having their desired effect.

Overall, I think the proposed approach is a logical one to address the stated questions. However, I think it could be improved and clarified in some ways (or at least these approaches should be considered).

Response 1: I would like to thank Dr Chin for his detailed review.

1. *The methods might be laid out a bit more fully and perhaps with an appendix with all the measured variables. For instance, I didn't see a variable for the time and date (but maybe I missed it) the loot box was searched for. This would seem to be important if it's ever matched up with the gambling license because this would establish the breach. Would the information provided the Belgian Gaming Commission provide the dates the license is/was active?*

Response 2: Dr Chin is correct to point this out. I did not note the data collection date and time as a variable and now I include this (p.12):

> *"Date and time of data collection*
> The date and time, based on Central European Summer Time (or Central European Time, depending on which will be used by Belgium at the data collection period), on and at which paid loot boxes were searched for will be recorded."

I cannot say whether the Belgian Gaming Commission would even respond, but I will seek to ask for the dates at which the licences (if any) were active (with permission, the response will be uploaded to OSF).

2. *Not knowing much about these types of games, I didn't have a great sense of how meaningful the top 100 sample is (I note the author(s) provide some other reasonable justifications beyond this point). Is this a big chunk of the market? Or, is market share pretty spread out in this area? This would help me get a better sense of the logic behind the sampling. Along those lines, if a large portion of games in the top 100 just are game styles that wouldn't have loot boxes, then it's not a very large sample at all.*

Response 3: With the limited, non-academic information (data from credible private analytics companies) that we have access to, yes, the video game market appears to be top-heavy and so selecting the highest-grossing games as the sample is justified. I have added the following to clarify (p.8):

> "Globally, the 100 highest-grossing mobile games reportedly accounted for 53.5% of all player spending on those platforms in 2020 [73]."
>
> Reference: 73. Chapple, C. (2021, June). The Top 100 Mobile Games Accounted For 64% of U.S. Player Spending in 2020. https://sensortower.com/blog/mobile-game-revenue-share-analysis-2021

3. *Fig 1a sets Hedges' g at .27 because of a previous study. First, I think this should be in the text and not relegated to a figure caption. Second, I'd like to see some consideration of why a previous study's effect size should guide the sample size here. Couldn't a smaller effect size be legally meaningful? If I was using this study to determine whether to adopt a certain regulatory regime and was quite concerned about gambling in video games – and if I thought it was not that costly to enact that regime – I might think a much smaller effect size would still be societally beneficial. Relatedly, why is a one-tailed test appropriate?*

Response 4: I have reconsidered this as set out in the first half of Response 12 to Dr Karhulahti. The Hedges' g is now −0.15 and is mentioned in the main text (p.14).

> "In the absence of any prior guidance on what effect size would constitute a 'legally meaningful' and 'socially beneficial' regulatory measure, based on intuition, it is proposed that a reduction from the abovementioned hypothetical 65.0% loot box prevalence rate to 50.0% or lower in Belgium would justify a law researcher to argue in favour of the Belgian Gaming Commission's regulatory enforcement actions as having been effective at providing improved consumer protection in comparison to other countries that have taken no regulatory actions. Accordingly, setting the Hedges' g at −.15, a priori power analysis using G*Power has determined, given an α value of .05: the present sample of 100 games would achieve .86 power in a two-sided test for finding a statistically significant difference between the Belgian and the hypothetical 65.0% prevalence rates (see Fig. A1) [73]."

I have changed the test to a two-sided test for H3, which is then followed up with one-sided tests. However, please note that one-tailed tests will be used for the new H4 and H5. I explain as follows (p.15):

> "One-sided tests are appropriate for Hypotheses 4 and 5 because Research Question 4 is only concerned with the possibility of the Belgian loot box prevalence rate having been reduced by the ban and to what degree that reduction has been."

4. *What was the interrater reliability in the two previous studies?*

Response 5: 93.3% for loot box presence (one disagreement; namely, a false negative) and 100.0% for Apple age rating in our 2020 study that dual-coded 15 games.

100% agreement for both variables in our 2021 study that dual-coded 20 games.

I have added this detail to p.12: "… (near-perfect or perfect agreement was achieved)."

5. *As to Hypothesis 3, what does it mean to say "If the Belgian paid loot box prevalence rate that will be found is statistically significant…". Should this be rephrased to statistically significantly lower than the UK rate, if that's what the author(s) mean? What's the logic behind the 40% figure? Should it be statistically significantly lower than 40% rather than just numerically lower?*

Response 6: This has been amended to Research Question 4, and H4 and H5, as detailed in the second half of Response 12. What I meant to do with the 40% was preregister how I will interpret the results as a legal scholar. I think the presently proposed 25% and 50% are more sensible.

6. *For Hypothesis 4, I was also wondering if these are the types of games known to have loot boxes. If not, it might be a waste of time, but I truly don't know enough to say. Would another approach (and this also applies to the study more broadly) be to pick games that are known to include loot boxes in other jurisdictions? This might be a stronger test of the hypothesis.*

Response 7: Yes, these three games mentioned in H4 are known to contain loot boxes in China and/or the UK. Loot boxes are also very important for these three games. I have emphasised this by adding further detail:

> '… (known to contain paid loot boxes in the UK)…' and

> 'Importantly, engagement with loot boxes is a fundamental and arguably unavoidable and inalienable aspect of all three games' gameplay and monetisation because the vast majority of in-game content (*e.g.*, playable characters) *requires* engagement with loot boxes to unlock (at least in the UK version of the games).' (p.16).

I think Dr Chin has proposed a different approach (*i.e.*, specifically only examining whether games known to contain paid loot boxes in other countries also contain them in Belgium, or whether they are even available for download in Belgium) that is capable of answering a different research question. Specifically, if the approach suggested by Dr Chin is adopted, which I have considered before, then it is possible to show that a certain percentage of games are simply not available in the Belgian market (a compliance action that some well-known companies, e.g., Nintendo, are known to have taken in Belgium).

However, adopting that approach instead of the one proposed here would mean that the loot box prevalence rate in Belgium could not be assessed. I personally think that the loot box prevalence rate in Belgium is more important to know than what percentage of a list of games known to contain loot boxes in other countries have been removed from Belgium. This list of games known to contain loot boxes in other countries is also difficult to objectively capture because games popular and highest-grossing in China/the UK might not be/have been so in Belgium (and may not have been published at all, rather than subsequently removed). Accordingly, I think that the approach that I have proposed is better for now, although I do not dismiss the possibility of returning to Belgium at a later date to use the second approach, especially if a non-significant result is found and further explanation is worthwhile.

Indeed, I note here and I will note in the limitation section in due course that it is possible that the overall loot box prevalence rate of ***all*** games in Belgium has indeed been lowered since the ban (considering that many reputable companies have reportedly removed their games or the loot box mechanic); however, the removed games are thereby removed from the highest-grossing list (which possibly now consist mostly of less legally compliant companies). This shortcoming cannot be addressed by the present study and would have to be addressed using the alternative methodology suggested by Dr Chin.

I do not think that I want to combine both approaches for this study because of logistical reasons: I would need to examine too many games during this visit to Belgium.

I always sign my reviews,

Jason Chin (ORCID: 0000-0002-6573-2670)

Response 8: Thanks again to Dr Chin.