Dear Prof Dr Malte Elson,

Thank you very much for the opportunity to submit a revised version of this stage 1 registered report proposal to *Peer Community In Registered Reports* (*PCI RR*).

I include the relevant text of the recommender comments and the three reviews in *black italics*; reviewer instructions in ***bold italics***; my point-by-point responses in purple; and amended or newly added text in blue below.

I must note that I understand that *PCI RR* will be closed for resubmissions in December, which is a policy that I fully support. However, due to the time restraints placed on the data collection period, this may well mean that I would need to proceed with data collection prior to potentially receiving acceptance in principle. I appreciate that the bias control levels may need to be reduced or that I may even need to withdraw this submission as a registered report as a result. Regardless, I will conduct the study using a revised version of the manuscript that has been improved by the comments from the recommender and reviewers and will be preregistered.

In the interest of open science, I intend to publicly archive the review history for this manuscript regardless of publication outcome.

Yours sincerely,
Leon Y. Xiao

--
Recommender Comments from Prof Dr Malte Elson
26 Oct 2023 19:22

Invitation to revise

Dear Dr. Xiao,

Thank you for your submission to PCI-RR. I now had the opportunity to read the paper in-depth, and the excellent reviews provided to evaluate the merit of your research proposal.

All three reviewers – Dr. Przybylski, Dr. Chambers, and Dr. Gunschera – mention they found your proposed study timely, and the research question worthy of an in-depth investigation as suggested. I too, share this view: Lootbox regulation, and industry compliance with it, are a topic of increasing attention within the gaming community and the public sphere. As such, a study as the one proposed could easily become a material piece of evidence in the evaluation of policy effectiveness, and perhaps even affect compliance with regulation itself.

However, all three reviewers raised important concerns with the proposed design of the study. I fully concur with them, and will add a few of my own observations below. Some of these points you might disagree with, and I invite you to provide counterarguments in a response letter. Others might be addressed by providing more details and improving clarify of the manuscript. And yet others, I believe, will require changes to your study protocol. The reviewers have offered guidance how the study design and the writing in the manuscript might be improved – please consider these points as you prepare a revision of your research protocol.

Response 1: I am very grateful to Prof Dr Elson for arranging for these three helpful reviews from a range of different perspectives. I address all issues raised in detail below.

STUDY SAMPLE AND GENERALISABILITY

Dr. Przybylski has remarked on the choice to include games not represented by Ukie, and that this weakens the severity of your test. I agree with this point: If this study is designed to test compliance with self-regulation principles by an industry trade body, then it does not seem ideal to include studies that do not fall under this self-regulation, and whose developers are not represented by Ukie. Whether there is a difference in regulation compliance between Ukie and non-Ukie games may itself be an interesting empirical question. I will leave it up to you to decide whether to pursue this or not, but if you do, then you need to account for this in your sample size and sampling strategy somehow. For example, if only 10% of the top 100 games are actually represented by Ukie (or vice versa), a serious empirical estimate of the difference would probably not be within reach. If resources are an issue, as you state, then it may be advisable to only include those games that are represented by Ukie, at the price of narrowing generalisability of your findings.

On this point, I also agree with the reviewer that the focus on the UK market should be represented in the title and conclusions of the paper. Going further, I believe it would also be appropriate to highlight the focus on mobile games, as the sample is restricted to games in the Apple store.

Response 2: I address the Ukie point in detail under Response 8 to Prof Przybylski. But in short, Ukie itself has publicly said that these principles *do* apply to non-Ukie games and that the trade body will seek to punish any non-compliance irrespective of whether it originated from a Ukie-represented company or not. Indeed, two of the principles being tested are already required through either platform rules (from Apple) or advertising regulations, regardless of whether a game is operated by a Ukie-member company or not. On that basis, I hope that it is evident that the original proposal of testing the 100 highest-grossing games regardless of their provenance and Ukie-representation status is sound.

I originally entitled the paper: 'Assessing compliance with ***UK*** loot box industry self-regulation' (emphasis added). I hope this would be sufficient in highlighting the UK point.

As to the Apple App Store point, I agree. I have therefore amended the title to: 'Assessing compliance with UK loot box industry self-regulation ***on the Apple App Store***' (emphasis added).

I have also changed Research Question 1 from being more general to now only asking:

> Are the 100 highest-grossing iPhone games complying with the UK loot box industry self-regulation?

As to the conclusions, these will of course duly highlight the focus on the UK market and on the Apple App Store as limitations.

WHAT IS A LOOTBOX?

Dr. Chambers and Dr. Gunschera both raised aspects that regard the definition of lootboxes in your study. Whereas Dr. Chambers asks whether one hour is enough to "encounter" a lootbox, Dr. Gunschera raises concerns regarding the focus on lootboxes that can be bought with real currency rather than in-game currency. Both of these points are important, and I believe they concern a mutual point: What is a lootbox, empirically, in your study? Surely it is not the virtual representation as a box, nor can it be any in-game purchase, nor any chance-based event. As such, I invite you to provide further details how you define and identify lootboxes in games, and by which means: You mention each game will be played for an hour. Does that mean "typical" game actions will be performed (as if you were a regular player), or will you just have the app open for this time? I am asking because it is conceivable that certain in-game actions are linked to lootbox drops. Overall, the manuscript lacks procedural and methodological details that the readers of the paper would surely appreciate.

Response 3: As to the definition for a 'loot box,' additional details have now been provided as explained in Response 23 to Dr Gunschera. Basically, I adopt the definition used by the UK industry self-regulation compliance with which the proposed study is seeking to test.

As to 'playing' the games for one hour, I now further clarify what this means by adding the following to the Method section:

> One hour of 'playing' the game will mean that, from downloading and starting the software, I will use my best endeavours for 60 minutes to unlock as many aspects of the game and gain access to as many in-game purchasing offers as possible: for example, I will choose to access the in-game store where loot boxes are presumably sold as soon as able, including by skipping unnecessary story elements. Our previous research using this methodology has acknowledged that the detection rate of loot boxes is not 100% because there are likely games that only begin to sell loot boxes many hours after the player starts playing and because loot boxes might simply be missed by the researcher[19(p. 12)]. This one-hour time limit is justified on resource constraints on my time. In addition, based on previous research, this method should be sufficient to detect at least 80% games with loot boxes (assuming that every game contains loot boxes, which is most likely untrue, so the true detection rate is higher)[13]. The percentage rate of games found to contain paid loot boxes within one hour of examination will be referred to as the 'prevalence rate' of loot boxes (as has been done in the past), even though more accurately, it would be the prevalence rate when only one hour has been spent examining the game and the true prevalence rate is therefore likely higher.

I also argue that this imperfect detection rate would not affect the testing of the compliance rates in Response 16 to Prof Chambers.

GAMES VS GAMERS

There is another important point by Dr. Przsybylski regarding the sampling framework as it affects the conclusions from your observations: Are you studying games or gamers? That is, if only games that with a small following are noncompliant, then surely we would have to conclude that the problem is smaller than if the top games (by number of "encounters" with lootboxes) were noncompliant. I think this is a conceptual problem that deserves further

attention, and that may not be easily "fixed" given that even obtaining reliable numbers on the games' market share might be difficult to obtain.

Response 4: As addressed in detail under Response 9 to Prof Przsybylski, this is a valid criticism and a limitation that I will duly acknowledge. I appreciate that a game that is played by 10,000 players being non-compliant is going to be practically different from a game that is played by 10 players being non-complaint (in terms of how many people are potentially negatively affected). But the present methodology would treat that non-compliance as equal. Unfortunately, I do not think this can be addressed with the data that are publicly available to us. With that said, I do believe treating each title as the most basic unit of measurement is justifiable, particularly in relation to the 100 highest-grossing games.

CUTOFFS

Dr. Przybylski and Dr. Gunschera have both remarked on the somewhat arbitrary choice of cutoffs to determine the compliance level. I, too, was confused where they came from, and to be honest I was wondering about the utility of defining cutoffs for the purpose of making a dichotomous decision in a hypothesis framework when just knowing about the empirical rate itself is of great interest (though I am happy to be convinced otherwise, maybe this just needs some justification). Further exacerbating, the point estimates you propose using will suffer from substantial uncertainty. For example, an incidence rate of 95 in a sample of 100 games has a 95% confidence interval of 76.861 to 116.133, the lower bound being below your cutoff for "inadequate compliance". Of course, I understand this is not a random sample of an unknown population of games: the top 100 are the top 100. Then again, I am sure you would prefer generalising your findings to games not included in the sample.

Response 5: These cut-offs are being proposed purely for the purposes of my commentary. Depending on what the compliance rate is going to be, I am going to express an opinion. I simply wish to preregister now how I will interpret the results. For example, one might subjectively interpret a 60% compliance rate as either poor or satisfactory: an industry representative might say it is good, whilst an advocacy group in favour of banning loot boxes might view it as terrible. I want to eliminate that flexibility from my future interpretation. This is why I am also seeking for the stakeholders to preregister their potential interpretation of the results.

Perhaps this is a slight misuse of the hypothesis testing framework, but I am unsure how else to present these cut-offs. In a previous *PCI RR* submission that has since been published, I used the same cut-offs for the same reason: https://doi.org/10.1098/rsos.230270. I am open to the Recommender's and the reviewers' suggestions on any potential alternatives.

I am actually of the view that the finding should *not* be generalised beyond the two present samples. It is my belief that the compliance rate amongst *all* games (or even a selection of the 500 highest-grossing games) would be significantly lower (particularly given previous external interventions with the most popular games). There are so many games on the platform that are never downloaded or played. I will take care not to do so in the reporting of the results.

Indeed, although it appears that I am taking two 'samples,' I intend to treat them as the population for the purposes of the research questions and hypotheses, which is

why these have been amended as set out in Response 2 above and Response 19 to Prof Chambers.

PROGRAMMATIC RR
Dr. Chambers raises a concern regarding your proposal to register this study as a programmatic RR. To be honest, I overlooked this point until I read his review, but I tentatively agree that I currently do not see the value or necessity to have two separate publications rather than one comprehensive paper that encompasses all research questions and data. Of course, I cannot stop you from writing two papers rather than one, but if you do insist on submitting this as a programmatic RR rather than a single RR, please consider the guidance offered by the reviewer, and highlight the different contributions of each paper, and why it is important or sensible to treat these differently.

Response 6: The original plan was to propose a programmatic RR that would not only look at the (self-)regulations in the UK, but also in South Korea, Taiwan, and the Netherlands. This is why one might have spotted a remnant of an older draft saying 'UK Component' that I originally forgot to delete but have since deleted. However, the exact regulations in some of those regions have not yet been set out in detail and/or I did not have confirmed funding to conduct that component (although I now do), so it was not possible to propose other components.

I then submitted this as a programmatic RR that will look at the UK situation at two different points in time, with the intention of producing a separate paper for each point in time. I considered whether this was appropriate prior to submission and have considered it again since. My original ideal plan was to publish this in the same journal as Part 1 and Part 2.

The relevant considerations in favour of having two papers were:

1. I want the first study to be published as a preprint separately. Having the results early can help to inform the implementation process (and perhaps enhance compliance before the second study).
2. I also want the first study to undergo the stage 2 peer review process before conducting the second study, so that any issues that could be addressed or improved upon in the second study might be identified (*e.g.*, interesting additional matters that could be assessed during the second study, which might be worth additionally preregistering).
3. From my experience conducting similar studies in the past, I expect that publishing both of these two studies may well take the word count to over 20,000 words.
4. I do believe that the amount of work hours involved with each study would justify 'a paper,' and I admit that having two papers on the CV instead of one might be beneficial…

However, having further considered the various arguments and your (the Recommender's) views and that of Prof Chambers, I propose to proceed with this as a regular, non-programmatic RR. I do hope the reviewers might be willing to have an (informal) look at the study 1 results prior to study 2 being conducted. I will publish study 1's results on its own and inform all stakeholders in any case. The ultimate aim is of course to inform the implementation process so that it might hopefully lead to better compliance and consumer protection. I want to conduct the best possible study 2.

With kind regards
Malte Elson
--
Review by Prof Andy Przybylski
26 Oct 2023 13:30

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

Reply 1A. The question of whether companies comply with statutory or suggested regulatory initiative is an interesting one to me. I approach reading this believing that there is very low compliance, the report suggests I should expect 1 in 3 games might comply if the UK is like the US. I am not quite sure that they research questions that are research questions in the classic academic sense. It is some form of policy or programme evaluation to my reading. I will defer to the editor on this point but note that the UK focus should be consistent from title to interpretation.

Response 7: Thanks to Prof Przybylski for taking the time to review this manuscript.

The previous Western results that we have regarding compliance with probability disclosure requirements are from the UK and would suggest that 2 in 3 games would comply and 1 in 3 would not (https://doi.org/10.1371/journal.pone.0286681).

The UK focus was included in the originally proposed title, which has since been amended also to highlight the focus on the Apple App Store in particular, as detailed in Response 2 to the Recommender.

Question 1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

Reply 1A. Given the UK-specific focus that justifies the research questions (and by extension the hypotheses) I am concerned by the framing of the research questions and how they're translated into testable hypotheses. If this is indeed a study of industry practices in the UK and premised on principles articulated by UKIE, shouldn't these hypotheses be focused on paid loot boxes in games that are represented by UKIE?

I do not believe it is a fair test of the principles if they don't only apply to companies who are represented by UKIE. Like social media, and online safety conversation more generally, this is a thorny problem. How might we regulate global tech industries (e.g. porn, social media, games) when these firms and the decisions they take are determined in Beijing or Palo Alto? I think the VGRF and these principles are very good ideas but I don't think it's a fair test of their local effectiveness to examine top grossing games in the UK if they're creators are based in the USA (ESA), or EU (VGE), or other jurisdictions. I believe the UK has many smaller developers, but I am not sure if they're represented in the top 100 or more likely to be on mobile or console/pc platforms. Is this the case?

Response 8: Thanks for raising this important point. I refer to this news article by Neil Long dated 31 July 2023: https://mobilegamer.biz/ukie-threatens-severe-fines-and-delisting-for-ignoring-new-loot-box-guidelines/ and preserved via the Internet

Archive Wayback Machine at: https://web.archive.org/web/20230801011753/https://mobilegamer.biz/ukie-threatens-severe-fines-and-delisting-for-ignoring-new-loot-box-guidelines/. I have also uploaded a copy to OSF believing that the preservation of this article for academic research, criticism, etc. is fair dealing within the meaning of UK copyright law: https://osf.io/6dmqn.

In this article published after the principles were first announced, a Ukie spokesperson was asked to clarify certain point and was on the record saying: 'The principles and guidance are there for industry to adhere to, and we expect the ***entire industry*** to adopt the principles' (emphasis added). In addition, the spokesperson said: "It is worth noting the working group membership is wider than just Ukie members, therefore ***the principles and guidance apply to all those in the video games ecosystem***" (emphasis added).

Indeed, the same Ukie spokesperson was further quoted: "More widely, members of this working group use a range of enforcement measures to ensure games are correctly labelled and carry an appropriate age rating." and "Remedial measures include delisting, relabelling and in some cases, severe fines."

Further, Ukie co-CEO Daniel Wood has said on the record: "We will also be working with the wider industry to urge implementation and track effectiveness."

Given the aforementioned statements made by Ukie, it is fair to say that the principles do indeed apply to non-Ukie games. Ukie understands this and expects this. Indeed, the UK Government and other stakeholders similarly expect compliance by the entire industry and not just Ukie members. Just having Ukie members complying would not be a satisfactory regulatory solution to the issue.

I hope that the above has satisfactorily demonstrated that testing compliance with the principles amongst all games, rather than just Ukie member games, is sound.

Similarly, I'm not sure that 100 top grossing makes sense given that I doubt these are equally profitable or popular games in the UK. For example, it might be the case that the top 4 or 5 games accounts for 80% of the play volume and spending. And the remaining 95% of the top 100 are just 20% of the market. If these 5 games were 100% compliant with the principles would you count this as 80% compliance or 5%? I think this materially effects all of the research question including the incidence/prevalence of probability disclosures.

Response 9: This is a good point, but I do not believe we have access to any objective measures that can address this point (*e.g.*, what percentage the spending in one specific game represents in terms of all spending on the app store). I can only offer to note this as a limitation that each game is treated as an equal. I believe this is justifiable because stakeholders are still interested in what percentage of the most popular titles (as the most basic unit for counting purposes) are complying.

Finally, without knowing the base rate of "ask to buy" I find it difficult to assess how well-justified disregarding this feature is. I know I use this feature with our under18s and I would not allow our children to use the app store at all without it. I think that this would introduce an unknown source of error or uncertainty in any of the point estimates which would be

reported in the work.

Response 10: I have justified why the Ask to Buy feature should be disregarded in my opinion, including, importantly, that it fails to directly address the loot box issue: the request that a parent receives would not even mention loot boxes as it is allegedly the industry standard for this request to be of purchasing premium in-game currency only. Therefore, one cannot claim that through the Ask to Buy feature, explicit parental consent/knowledge to purchase loot boxes has been obtained. In the discussion section, I will note in due course that a high rate of adoption with an improved version of the Ask to Buy feature that actually sends a request for the purchasing of loot boxes with paid premium virtual currency could be viewed as compliance with the relevant principle.

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

I think this is a feasible way to approach it if the above base rate and geography issues are tackled. I am not really sure where the 95, 80, and below 80 levels come from though. Reading earlier in the report, I might expect the rate to be 35%. I could envision this being the standard and movement starting from this level (to I would hope something much higher) being the standard. My sense is that the author is interested in improvement, so how much improvement would be needed to know if progress is being made in the UK?

The author might also consider starting with a prior belief there is a 50/50 chance that a UK game creator is getting things right is the correct starting point and seeing if this is true at the start of the data collection and if this has improved at the 6 month mark.

Response 11: I justify why these cut-offs are being proposed in Response 5 to the Recommender. These are purely for the purposes of my expressing an opinion on the compliance rates that will be found.

The present sample size is far too small to test any increase in compliance (of maybe 10%), so I propose simply to do the following, which I have added to the Method section:

> If the compliance rate with a specific measure improves from one band into the next (*e.g.*, from $< 80\%$ to $\geq 80\%$) when the 18 January 2024 sample is compared with the 18 July 2024 sample, then I will comment positively on how compliance has improved.

I do not understand how (or who) at DCMS or UKIE would preregister their hypotheses (lines 473 and 474) or what the value of this would be. I don't think most video game researchers would be able to do this.

Response 12: The intention here is to send an email informing the relevant civil servants at DCMS and staff members at Ukie of this study and asking them to preregister what results they would consider satisfactory. I can ensure that this request is delivered, but Prof Przybylski is right that I cannot guarantee that they will respond. It is understood that both are interested in tracking compliance, and

asking them to preregister their potential interpretations would help to provide more accountability to the public. For example, if a stakeholder preregisters that only compliance over 60% would be deemed satisfactory, then it would be obliged to admit that compliance has not been satisfactory if it turns out to be below 60%. More action is then needed to improve that compliance rate. The public would benefit from knowing each stakeholder's expectation. Therefore, I do believe this is a worthwhile endeavour. I will send this email alongside the first revision of this manuscript.

Question 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 do not believe so. I think a detailed protocol with its own figure would be helpful.

Response 13: I now provide more information on how the game analysis will be conducted as explained in Response 3 to the Recommender. I would be willing to consider making a figure if Prof Przybylski may be able to provide me with an example. I am unsure what a figure should focus on detailing and highlighting.

Question 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).

I do not think so, but I am not sure that is a problem. As the study is framed currently, I do not see a situation where the hypotheses won't be confirmed.

Response 14: Haha! I would like to think that companies are following platform rules and advertising regulations, especially after both were further clarified through academic research, industry and popular media reporting, and decisions being made by the Advertising Standards Authority. I am personally of the hope that compliance with probability disclosures and loot box presence disclosure might actually achieve the highest band of compliance this time.

--
Review by Prof Chris Chambers
20 Oct 2023 12:52

I enjoyed reviewing this Stage 1 RR – it tackles a timely and important research question and clearly spells out the rationale, hypotheses and proposed methodology. I am not a researcher in this area and will defer to experts for specialist assessments. Instead I focus my evaluation on issues that are generally relevant across most Stage 1 RRs. I hope my comments are helpful.

1. On the issue of sampling bias, I think you make a good point that we cannot know whether compliance is driven by the current changes or prior external intervention (pp8-9); and that consequently this makes it difficult to generalise the eventual results to compliance rates more broadly. To address this point specifically, could it be useful to include an exploratory analysis at Stage 2 within the subset top-100 games for which no previous intervention was known? Could a comparison be useful (even descriptively) between games subject to prior intervention vs no prior intervention?

Response 15: I thank Prof Chambers for providing helpful feedback.

Of course, I would be happy to do so in the Stage 2 submission. I expect we will see between 10–20 games that have not previously been studied/externally intervened with. I will also note that we cannot say whether any potential differences between the compliance rates are caused by the lack of external intervention. This is because the games that have not been studied before are likely newly released games. These games are probably more likely to be compliant as the rules have become clearer and industry self-regulators seem to be monitoring those games more actively. Therefore, the compliance rates amongst the subsamples might appear very similar but for different reasons. I added the following to the Method section:

> To further address the issue of how the compliance rates amongst the highest-grossing games may have been affected by previous external intervention, the compliance rates for each loot box self-regulatory measure will also be separately reported for games that have previously been studied and those that have not been.

2. You have allocated 1 hour per game to detect loot boxes. How confident are you that this is long enough to detect loot boxes where they exist? I would recommend including some justification for this specific period. Ideally, the sensitivity of this test could be confirmed through evidence rather than intuition: e.g. the strongest case would be previous data confirming that in cases where loot boxes are known to exist, 1 hour is sufficient time to always detect them (and if the detection rate is less than 100%, then what consequence will this have on the sensitivity of the current design to test the hypotheses).

Response 16: We actually do not have data to directly answer that particular question. Our previous research would suggest that loot boxes can be found in one hour in about 75–80% of games (and we tended to have just referred to that percentage rate as the loot box prevalence rate). We suspected in previous studies that some games contained loot boxes that were only accessible after dozens of hours of gameplay, but we treated those games as not having contained loot boxes. The loot box prevalence rate that will be found can only be lower than the true value but

not higher. We have always recognised this as a limitation, as it is always possible to fail to detect loot boxes but any loot box that we do find would be evidenced with screenshots that can be objectively verified. I will again note this limitation with the results.

As to any potential effect on the compliance rates due to the detection rate being admittedly less than 100%, I do not believe there will be any because games deemed to not contain loot boxes will be excluded when testing the hypothesis. We do not know whether games whose loot boxes are only accessible after many hours of gameplay are more or less likely to comply, and I do not think any prediction could reasonably be made. I will also not seek to overinterpret the results beyond the actual sample.

I further clarified what it means to 'play' a game for an hour under Response 3 to the Recommender, including adding specific justifications regarding the time limit:

> This one-hour time limit is justified on resource constraints on my time. In addition, based on previous research, this method should be sufficient to detect at least 80% games with loot boxes (assuming that every game contains loot boxes, which is most likely untrue, so the true detection rate is higher)[13].

I also added the following to the Method section to explain how an imperfect detection rate should not affect the testing of the hypotheses regarding compliance:

> Even though some games might be inaccurately marked as not containing loot boxes even though they do using the present methodology of examining the game for one hour only, the compliance rates with various regulatory measures will not be affected because games assumed to not contain loot boxes will be excluded. The relevant compliance rates will reflect the true situation amongst the games containing loot boxes that were actually tested.

3. p15: "Stakeholders (specifically, the DCMS and Ukie) will be invited to preregister how they will interpret different potential results that may be found by the present study." If possible, I would suggest inviting them to do this now and then including this pre-specification in the revised Stage 1 RR – that way they will be as bound by their prospective interpretation of the findings as you are.

Response 17: As I also discussed in Response 12 to Prof Przybylski, I will proceed to do this now with the revised version of the manuscript. I did not want to reach out yet at the time of the initial submission because I thought the research methodology might be subject to major changes.

4. On the issue of delisting resulting in loss of apps: To ensure an adequate sample size, I suggest anticipating the likely delisting rate and overrecruiting in Jan 2024 by that amount to maximise the probability that the July 2024 sample still includes the top 100 at that time (e.g. if a 5% delisting rate were to be expected then take top 105 games in Jan 2024).

Response 18: The July 2024 sample will consist of a new list of the 100 highest-grossing games at the point in time (which will be different from the January 2024 sample), so any delisted games would simply have been replaced with lower

ranking games. This means that the July 2024 sample is guaranteed to contain 100 games.

5. Precision of hypotheses. The hypotheses are generally clear but I would recommend two changes. First, they should make explicit mention of the two time periods and whether the same predictions are made at each point. Second, even though there is no inferential statistical analysis, this is still quantitative hypothesis testing so the manuscript should include a study design template.

Response 19: I have amended Hypotheses 1–3 to explicitly refer to the time periods and read as follows:

Hypothesis 1: All highest-grossing iPhone games containing paid loot boxes in the 18 January 2024 sample and the 18 July 2024 sample will prevent loot box purchasing by under-18s unless parental consent has been provided.

Hypothesis 2: All highest-grossing iPhone games containing paid loot boxes in the 18 January 2024 sample and the 18 July 2024 sample will disclose loot box presence.

Hypothesis 3: All highest-grossing iPhone games containing paid loot boxes in the 18 January 2024 sample and the 18 July 2024 sample will make loot box probability disclosures.

Hypothesis 4 already referred to the time periods, hence no changes were made.

I now include a study design template; this omission was unintended.

6. The Jul 2024 period is very reasonably at the conclusion of the implementation period. I am wondering however if there would be any value in pushing this back to Aug 2024 to capture any possible delays in compliance? I don't know enough about this area or the regulatory frameworks that operate, but is there any possibility that a company could intend to comply but just be a few weeks late? By allowing a post-implementation "grace period" of e.g. 1 month (from Jul to Aug), would the demonstration of low compliance rates be a more powerful signal to stakeholders and provide less wriggle room for non-compliant companies to plead minor delays? I will defer to the author's judgment on this point and note it for consideration only.

Response 20: Prof Chambers makes a good point here, but I think I would stick to the July 2024 data collection point because it is when companies are expected to comply, and they have already been given 12 months to implement the measures (which I view as the 'grace period'). Attempts to plead minor delay (if made) would actually reveal that companies are not taking the issue sufficiently seriously. A further study could always be conducted a few months following implementation to test the compliance rates again (although perhaps not by me).

7. My final comment is about the programmatic nature of the submission. I can certainly see the value of separately evaluating compliance during and following the implementation period. However, it also seems to me that the final results will be more coherent as a single encapsulated Stage 2 RR rather than two RRs. I am also not sure that the pre vs post

implementation components are sufficiently substantive to justify 2 x Stage 2 outputs under Stage 1 criterion 1C (though I concede I am viewing this through a non-specialist lens and do not intend to devalue the amount of labour involved). Also: A programmatic Stage 1 RR typically includes separate sections to explain which specific parts of the proposal will be presented in the different outputs, sometimes going as far as to indicate different font colours to show which text will go in which manuscripts, and these details are always specified in advance (e.g. see here and here for examples). So, in the event that the submission ends up being programmatic, some similar structural work will be needed here.

Response 21: Thank you for pointing this out. I have set out my detailed thoughts on this in detail under Response 6 to the Recommender. In short, I will proceed with this as a regular RR. However, I will publish the results of Study 1 as soon as possible to inform all stakeholders and hopefully assist in better implementation.

Minor:

Lines 152-155: I struggled to parse this sentence.

Response 22: As addressed under Response 26 to Dr Gunschera, this has been fixed.

--
Review by Lukas J. Gunschera
16 Oct 2023 11:27

The manuscript at hand addresses an important issue, the compliance of the mobile game industry with UK self-regulation loot box measures. This work is timely and will make a great contribution to literature and policy concerning gaming consumer protection. That being said, I have found that the manuscript may be improved in the following areas.

1) The scope of the present manuscript concerns loot boxes purchased with real currencies as opposed to in-game obtained currencies. Although this distinction is common in the literature, I believe it warrants elaboration and think the proposed work would benefit from recording data on all possible avenues of purchasing loot boxes (i.e., whether players have the option to purchase the loot box with in-game currencies in addition to real currencies). I believe this is informative due to the fact that the gambling-like characteristics of loot boxes persist irrespective of the currency used to obtain them. The value of any currency, whether real or virtual, is learned. Therefore, beyond the concerns for parents' wallets, the psychological effects of loot box purchasing may span across the currencies used to purchase them.

Furthermore, the psychological effects of loot boxes may even be strengthened for purchases with in-game obtained currencies, as opposed to money. Players who have invested many hours into obtaining the said in-game currency may perceive this to be a much larger investment than money, especially when the money comes from their parent's wallet. While I understand that the distinction between real-world and in-game currencies is common, I believe it would be worthwhile collecting information on the currencies that can be used to obtain loot boxes (money, in-game, both) for each of the 100 mobile games (ll. 87-93, 364-368).

Response 23: I am grateful to Dr Gunschera for reviewing this submission.

I agree with Dr Gunschera that loot boxes bought with virtual currency that has been 'earned' *purely* through gameplay without the involvement of real-world money should be further studied. A relevant article is: https://doi.org/10.1007/s10899-022-10127-5, which supports some of the points Dr Gunschera has made.

However, such loot boxes are not within the ambit of the Ukie self-regulations. In Annex B to the Principles, the following definitions were provided:

> *"Loot Box" means a video game mechanic that provides random in-game virtual items to players in exchange for real-world money or in-game virtual currency. This document does not apply to a loot box that is purely earned through gameplay.*
>
> *"Paid Loot Box" means a Loot Box that is either purchased using real-world money or acquired using virtual currency that itself has been purchased.*

For the purposes of the present study, I therefore use the same definition adopted by the UK self-regulation. As Prof Przsybylski correctly pointed out (albeit in a different context), the tests should only be conducted on products against whom the regulations actually apply.

I added more details about this definition to the introduction:

> However, the present study focuses on paid loot boxes that players spent real-world money to purchase either directly or indirectly by spending money to purchase 'premium' in-game currency that can then be used to purchase loot boxes.
>
> …
>
> Importantly, a 'loot box' needs not be visually portrayed as a box: any in-game purchase involving real-world money with any randomised elements satisfies the definition[2].

For further clarity, I also added the following in relation to the variable concerning the *Presence of paid loot boxes*:

> … paid loot boxes (as defined in Annex B of the Ukie self-regulation, which aligns with the present study's definition as set out in the introduction section)…

2) Despite resource constraints and stakeholders' heightened interest in the highest-grossing mobile games, the sample size rationale is insufficient. A power analysis/simulation would help determine which effects the study would be sensitive to, especially in consideration of the fact that precise decision cut-offs are given for all hypotheses (ll. 245-254).

Response 24: As addressed under Response 5 to the Recommender and Response 11 to Prof Przybylski, I am not intending to test a 'sample,' but instead I am testing the population, which is the 100 highest-grossing games. I will not attempt to generalise the results more broadly than from where they were derived.

3) For Hypothesis 4 the decision criterion is different to the preceding hypotheses. Please add a brief explanation for this change (ll. 227-230).

Response 25: I have added the following:

> The expectation that 100% (rather than 95%) of games will either become compliant or be delisted is justified on the basis that a list containing all relevant games will have been provided to the stakeholders to take enforcement actions. Any potential Type 1 error will be eliminated by how the Apple App Store and/or the relevant video game companies will be given the opportunity to provide evidence that the game does not contain loot boxes or have already made the relevant disclosures, so a further 5% of leeway (given to Hypotheses 1–3) is not appropriate for Hypothesis 4.

Above this section, in relation to Hypotheses 1–3, I also added this clarification:

> This 5% of leeway will be permitted as a type 1 error control measure to account for potential false positives.

4) Overall, the manuscript would benefit from some type-editing. This includes breaking up long and convoluted sentences; using accessible language as opposed to unnecessarily complex words; and using precise and objective wording. Some examples below:

ll. 152-155 Convoluted sentence structure

ll. 171-174 Complicated wording

ll. 196-198 Subjective/moral wording

Response 26: The sentence that was at lines 152–155 has been broken up and changed to:

> Prior research has demonstrated that loot box regulations, particularly industry self-regulatory ones, were poorly complied with in the past. Accordingly, reasonable doubt can, and ought to, be cast on whether companies will comply with the newly proposed UK loot box industry self-regulation.

As to potentially subjective/moral wording, I am of the view that researchers should not attempt to hide their opinions under a veneer of apparent 'objectivity.' I have to express an opinion on whether compliance will have been satisfactory or not, and I have set my expectations at a certain level (*i.e.*, the cut-offs), so it would be more accurate to share any preconceptions that I may have through my writing rather than to attempt to conceal them. I appreciate that others might disagree.