**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).

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

I always sign my reviews,

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