Dear Dr. Pennington,

Thank you for an opportunity to submit a revised version of the manuscript. In the below we detail our point-by-point responses to each comment raised.

Sincerely,

Dr. Philip Newall

Dear Philip Newall and collaborators,

Thank you for submitting your Stage 2 Registered Report. I have now received reviews from two experts in the field of behavioural addictions who also reviewed your Stage 1 RR. You will see that one is very positive with only minor suggested changes and the other is more detailed with more substantial recommendations. I agree with the evaluation of Reviewer 2 (Chen) and would like you to carefully consider these, outlining your responses along with your revision. I have also undertaken a full review of your manuscript and detail my recommendations below – the main one being the reporting of effect sizes and associated confidence intervals for the t-tests and equivalent tests. I therefore offered a decision of: Invitation to revise.

Some additional comments from me regarding my own review and information about formatting:

1. Thank you for including the Stage 1 accepted protocol in the Abstract section. You may want to adjust its positioning, however, to ensure it reads exactly how you'd like to see it when it's in print. For example, I personally think this would read better in the Methods section of the Abstract. Please use the date format DD/MM/YYYY. This same information should also be outlined in your Methods section, in the first paragraph (also see Point 2).

Response: We have moved the reference to the Stage 1 protocol in the Abstract and also added this text to the first paragraph of the Method.

2. The first paragraph of the Methods section is still written in present tense and from my review, the link to the OSF project page (https://osf.io/6hbyp/) does not include the data and

materials. Please also carefully consider the layout of your OSF project page – it is clearer for users if you use sub-folders such as 'Data', 'Materials' etc with accompanying read.me files or data descriptions. PCI RR is a signatory of the Transparency and Openness Promotion (TOP) guidelines, which describe a series of modular standards for transparency and reproducibility in published research. In general, authors are required to make all study data, digital materials, and computer code publicly available (at Stage 2 submission) to the maximum extent permissible by relevant legal or ethical restrictions.

Response: We have added the materials to the OSF project page and also created those two sub-folders as suggested.

3. The Abstract states: "The optimal communication of risk information can act as an input to a public health approach to reducing gambling-related harm" but it is not clear exactly what this optimal communication of risk information might be from your findings.

Response: The findings were mixed across the two dependent measures. We have rewritten this last sentence of Abstract to try and provide a more clear summary of the pattern of findings as follows:

"These results replicated prior work on the advantages of the original house edge phrasing over return-to-player information, while showing that the alternative house edge phrasing has advantageous properties for gamblers' perceived chances of winning only."

4. Please check your manuscript for grammar. E.g., there are full stops missing in the Participants section of the Methods.

Response: We have added the missing full stop and also each re-checked the manuscript for spelling/grammar issues.

5. The Method section seems to read confusingly without the information about the final sample size. Can the following sentence be taken from the Results and added to the Participants section in the Methods: "Overall, 3,453 participants completed the experiment as planned. Of these, 62 (1.8%) participants completed the study in under one minute, and a further 58 participants (1.7%) indicated that we should not use their data, resulting in a final sample size of 3,333 participants". You also need to provide a breakdown of how many participants were in each condition.

Response: This has been moved as requested and the following breakdown added:

"(1,099 in return-to-player condition; 1,118 in original house edge condition; 1,116 in alternative house edge condition)."

6. The study has now been completed so I ask you to pay close attention to the phrasing from your Stage 1 Report – it is fine to change such phrasing/tense (only) in the Stage 2 report. For example, the following sentence reads strangely: "We now consider some power analyses to support our plan to collect 1,000 usable responses per-condition. For H1 and H2, it is impossible to know at this stage what magnitude of change on the dependent measures would lead to meaningful differences in actual gambling environments." This could be "Within our preregistered protocol (see LINK), we considered some power analyses…"

Response: We have reworded this to:

"The preregistered protocol (https://osf.io/5npy9) contained some power analyses to support our plan to collect 1,000 usable responses per-condition."

And we have also changed the rest of this section accordingly.

7. In your Methods section you state that "We aimed for an average sample size of 1,000 participants passing data quality checks per-condition, as this is was the closest round number which exceeded the required sample size in each of the below power analyses", but this is inconsistent with a later sentence that states "Therefore, with these two data quality checks in mind, we planned to collect data from 1,151 participants per-condition in order to reach our planned sample size." Can you explain this discrepancy and fix it?

Response: Yes, the 1,151 was planned during the previous stage of review given the data quality checks that were added. This has now been clarified:

"Therefore, with these two data quality checks in mind, we planned to collect data from 1,151 participants per-condition in order to reach our planned sample size (of 1,000 usable responses per-condition)."

8. The section on Statistical Analysis is still written in the present tense – please ensure consistent tense is used in the final Stage 2 report.

Response: Apologies. We have now changed the whole Methods section to past tense.

9. For the equivalence tests, I don't necessarily agree with Reviewer 2 that the t-tests don't add anything different. The t-tests show you whether the effect was statistically significant to zero whereas the equivalence tests show you whether or not you have a meaningful, equivalent, or inconclusive effect. There is missing information here, however, which is required to the evaluation of equivalence tests – the 90%/95% confidence intervals (whichever one used) of the effect size. As per Zhang Chen's comments, please ensure that your evaluation of these is correct – if the equivalence tests are significant but include effect sizes at either end of the CIs which overlap with both a practically meaningful and practically equivalent effect, then the data is best described as inconclusive (the ES could be meaningful, or it could be not meaningful… there is not enough data to conclude). Please add the effect sizes and associated CIs to the reporting of the equivalence tests and check your interpretation of these with regards to meaningful, equivalent or inconclusive.

Response: We thank the editor for their comments on the equivalence tests. We agree with the suggestion regarding the confidence intervals, and have amended the manuscript to incorporate these:

"Moreover, the 95% confidence intervals for the tests fell outside the lower bounds for both sets of equivalence bounds (H1 – raw difference = -0.713, 95% C.I. = -0.820 – -0.605, H2 – raw difference = -0.187, 95% C.I. = -0.221 – -0.153). Therefore, we can conclude these differences are both statistically and practically different."

The CIs do not overlap, and so we interpret these findings as having a practically meaningful difference.

10. The discussion states: "Overall, this supports the use of the original phrasing instead of return-to-player information (Newall, Walasek, & Ludvig, 2020a)" – is this finding in line with your 2020a research? The placement of the reference made me question this so you might want to say "in line with Newall et al. (2020a)" (if this indeed was the case).

Response: This rewording has been used.

11. On Page 18, the first and second paragraphs should be combined together – they don't read well separately.

Response: These paragraphs have now been combined.

I wish you the best of luck with your revision and look forward to receiving it for further consideration.

Dr Charlotte Pennington

PCI Recommender

Reviews

Reviewed by Zhang Chen, 17 Jan 2023 15:09

Thank you for the opportunity to review this Stage 2 manuscript. I have received the Stage 1 proposal of this paper, and thus mostly focused on the results and the discussion section. The authors adhered precisely to the registered study procedures, and the data are able to address the proposed hypotheses. The conclusions are also justified given the evidence. Below I list a few comments/suggestions, mostly quite minor, to hopefully make the results more informative and easier to follow. To be clear, the page numbers refer to the version with tracked changes. Hope my comments are useful to the authors.

1. Page 9: I think it would be useful to have more information on the participants in terms of their (self-reported) gambling behavior. For instance, from Prolific, it should be possible to download the participants' responses to the prescreen question "What types of online gambling / casino games have you played?". Providing some descriptive information on e.g. the number of gambling games played by each participant, and the most popular gambling games in this sample would be informative.

Response: We thank the reviewer for their excellent suggestion on this, and have incorporated this into the participants section. On average participants engaged with between 3 and 4 different types of activity, with around half having played on a slot machine, which is of particular interest to this study.

2. Page 9: It would be useful to mention the scores on the PGSI used to create the four groups of gamblers.

Response: These have been added: recreational gamblers (score 0; 39.3%), low-risk gamblers (score 1-2; 31.7%), moderate-risk gamblers (score 3-7; 21.0%), high-risk gamblers (score 8+; 8.0%).

3. Page 12: It would help me as a reader if the authors could specify how the seven options in the Likert scale are coded. For instance, when first encountering the scale on page 11 “My chances of winning are... Very high / High / Somewhat high / Neither high nor low / Somewhat low / Low / Very low chance of coming out ahead”, I assumed that "Very high" = 1, "High" = 2 etc. However, after reading the results and the conclusions, I think "Very high" = 7, "High" = 6 … Thus, a higher score means a higher perceived chance of winning. The authors may explicitly state this in the methods and results section.

Response: This has been added to the statistical analysis section (7 = “very high”, 1 = “very low”).

4. Page 12: I find the part on equivalence tests a bit difficult to follow. First, I think the first two t-tests are redundant because they essentially test the same hypotheses as the two regression analyses reported above? Or are they actually inferiority tests (see below)? The authors may need to explain the rationale behind equivalence tests a bit more to help readers better interpret and understand the statistical results. E.g., one of the two one-sided tests tests when the effect is larger than the lower bound of the region of practical equivalence (ROPE). However, since the effect size in this case is smaller than the lower bound of the ROPE (thus in the opposite direction as tested by the one-sided test), the one-sided t test yields a p value of 1. From Figures 1 and 2, it is clear that the effect sizes are outside of the ROPE, and the authors conclude that the differences are "practically different". However, they may strengthen this conclusion by formally testing it, by e.g. using an inferiority test. Lakens et al. explained these different tests in their tutorial paper Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259-269.

Response: We thank the reviewer for their suggestion on this. Based on the recommenders comments, and the advice from the literature, we feel sticking with the equivalence tests is the best course of action. There are two reasons for doing so. First, the hypotheses made were non-directional, so the equivalence testing is keeping with the overall spirit of the hypotheses made. Second, because inferiority testing is one directional, the equivalence tests are more conservative in the inferences they make. Moreover, as the Lakens et al paper points out, this ought to be a decision made before initial analysis (although repeating them with a Bonferroni correct is unlikely to change the overall findings). For future studies, we agree that the use of inferiority testing is advised as there is a much clearer idea of the direction of the effect.

5. Table 3: This may be a matter of personal taste. Personally I find the table easier to read and understand if the coefficient and the 95% CI would be presented together in one column (since the 95% CI reflects uncertainty in the estimation of the coefficient), and t value and p value would be in one column (since they often have a one-to-one mapping).

Response: This change has been made.

6. Table 3 shows that the coefficient for the interaction effect on perceived chances of winning is positive, with the confidence interval just including 0. Seeing this effect, I think the conclusion that "PGSI levels did not influence either outcome measure (p's ≥ .078)" (page 15) may be premature. Apart from the usual problem that one cannot support the null hypothesis based on p values larger than .05, this interaction effect shows that the effect on perceived chance of winning between alternative versus original phrasing of the house-edge information might be reduced among problem gamblers. While acknowledging that this is exploratory in nature, I would not immediately dismiss this potential effect.

Response: We are also wary of interpreting p's > .05 as evidence for an effect. We have added the following extra sentences to this point:

"The closest effect to the preregistered level of significance was the interaction effect between PGSI and house edge condition for the perceived chances of winning measure. If further replications find that these effect exceeds the preregistered significance threshold, then this suggests that the difference between the two house edge phrasings might be lessened among gamblers suffering from higher levels of problem gambling severity."

7. This is a minor point - sometimes the text and the statistical result are not consistent with each other. E.g., on page 15, "which according to an ordinary least squares regression was higher than both the original (B = -0.95, t = -14.1, p < .001, 95% CI [-1.08, -0.82]), and alternative (B = -1.66, t = -24.6, p < .001, 95% CI [-1.80, -1.53]) house edge conditions", the text says "higher" but the coefficients are all negative. While I am aware that this has to do with which group is used as the reference category in the analysis, presenting the results consistently may nevertheless facilitate understanding.

Response: The return-to-player condition was used as the baseline, so it having higher perceived chances of winning (7 = highest chances of winning on response scale, see response above) is fully consistent with the house edge conditions having negative coefficients.

8. Similarly, for Figures 1 and 2, to facilitate understanding, the authors may consider adding an arrow pointing from 0 to the left, and a label such as "the alternative group had lower perceived chances of winning/had a lower accuracy", and an arrow pointing to the right and saying "the original group had lower perceived chances of winning/had a lower accuracy".

Response: We have added a further sentence of explanation to the legend of each Figure, and hope that this change would be sufficient to satisfy the Reviewer.

9. Page 17: The reduced correct understanding in the alternative group might be explained by one of the two added incorrect answers. While the authors argued that this made the measurement more sensitive, it also made me wonder whether there is research on what kind of interpretations gamblers spontaneously generate when they are presented with different messages. If they do not spontaneously generate interpretations as the added incorrect option, adding this option may make the measurement less valid? It may also be interesting to test understanding in such a way, with open-ended questions - after all, in real-life gambling situations, gamblers are also not presented with options to choose from.

Response: That sounds like an interesting area for future research. We have added the following to the limitations section:

“Finally, future studies may want to use other methodologies to measure gamblers’ understanding of relevant information, for example by asking them to explain their interpretation via an open-ended text box, rather than by providing a list of potential answers via a multiple-choice question.”

10. The discussion mostly focused on the issue of correct understanding. I wonder how the authors may explain the lower perceived chance of winning in the alternative phrasing of house-edge condition.

Response: We have added the following speculation to the Discussion:

“On the other hand, this specific wording may have resulted in that phrasing’s low perceived chances of winning, as it emphasises the likelihood with which gamblers will lose over time in electronic gambling formats.”

Reviewed by Graeme Knibb, 15 Feb 2023 14:00

This study investigated the effect of two different ‘house-edge’ messages on gamblers understanding of loss. In addition these two messages were compared to return-to-player information.

Since the stage1 review the manuscript has only minimal changes (related to tense). There is also a clear link to the approved protocol. The analysis does not deviate from those which were initially planned.

The overall conclusions, as presented within the discussion, are clear and based on the evidence.

As stated in my previous review I think this is a strong and well considered RR and I am very happy to see a stage 2 manuscript. I have no issues with the study or the conclusions that have been drawn and I am very happy to recommend this stage 2 manuscript for publication.

One minor wording that may need changing does appear on page 8 in which the open materials are discussed. The language here is still in future tense and should be altered accordingly. I.e. ‘will be placed’ could become ‘can be found’.

Response: We have made that change in-line with the Editor’s recommendation.