Dear Dr. Pennington,

Thank you for the extensive feedback from yourself and the Reviewers, and the opportunity to submit a revised manuscript to PCI Registered Reports. In the below we include the complete text of the feedback that we were provided, and our point-by-point responses.

It took some time for us to complete these revisions, due to the range of suggested improvements to the methodology and statistical analysis. To this end, we have added an additional author at this stage of the work (Richard James), who has helped with some of the requested additions. We have also obtained an updated Ethics Committee approval for the updated experiment.

We look forward to hearing from you again in due course.

Sincerely,
Dr. Philip Newall

Decision on your Stage 1 Registered Report: Revisions Required

Dear Philip Newall and Olivia Maynard,

Thank you for submitting your Stage 1 Registered Report “How does the phrasing of house edge information affect gamblers’ perceptions and level of understanding? A registered report” for consideration by PCI Registered Reports.

I have now received comments from three expert reviewers in this field. As you will see, these reviews are overall positive and based on these reviews and my own assessment, I would like you to revise your manuscript accordingly. You will see that the majority of the reviewer comments are minor, but I would like you to pay attention to the following, which will be a particular focus on re-review. These two criteria are essential for Recommendation at PCI Registered Reports.

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

1. Please ensure that the power analysis is consistent with the analysis being used: as one Reviewer points out, it seems as though the power analysis is based on M/SD and an effect size of Cohen’s d, but the main planned analysis is a regression which is inconsistent with this. [but also, see Point 3 below which needs to be considered in addressing this]. Importantly, your plan is to compare two groups based on the phrasing of the message, but the justification of your power analysis is not currently based on this and is instead based on an average point on a Likert scale. Perhaps I am misunderstanding something here, which could simply be clarified, but again the power analysis should be based on the specific analysis you plan to conduct.

Response_1: Thank you for this observation. Based on some of the other below comments (which will be explained further lower down), we have decided to switch the analysis for this outcome to OLS. This switch has the side-effect of now making the power and the planned analyses consistent.

2. You state that the sample size would be able to “detect even relatively small effects”, which is ambiguous – what specific small effect could be detected? Have you considered a Smallest Effect Size of Interest (SESOI) and powered your study according to this?

Response_2: Consistent with some of the other below comments, we have now added a power analysis for the second outcome, and also planned an equivalence test in the event of either H1 or H2 yielding a non-significant p-value. For the equivalence testing, we have proposed a smallest effect size of interest of d = 0.133, and performed a power analysis based on this figure.

This has been reflected in the following addition to the manuscript (pp.11-12):

“In the event that the tests for either H1 or H2 are not significant (*p*’s ≥ .05), equivalence tests will be conducted using the two one sided t-test (TOST) procedure. Whereas standard null hypothesis significance procedures test the hypothesis that the difference between groups, or the association between variables is significantly different from zero, equivalence testing allows effects below a given interval to be rejected as “too small” to be of practical significance, which is referred to as the “smallest effect size of interest” (Lakens, 2017). Power analysis was conducted to test whether the proposed analyses were appropriately powered given the sample sizes proposed using the ‘power_t_TOST’ function in the TOSTER package (Lakens, 2017) in R (R Core Team, 2020). For this power analyses, the smallest of the two effect sizes from the previous power analysis was used as the smallest effect size of interest (*d* = 0.133), and this suggested a required sample size of 969 participants per-condition to achieve 80% power. This final power analysis supports our intention to collect 1,000 usable responses per-condition.”

3. As per the Reviewer comment, the study is not sufficiently powered for the ‘understanding’ variable. An explicit reason should be given for this, which can also include resource limitations (i.e., funding/time), but must be apparent.

Response_3: A power analysis for this outcome has now been added (p.11):

“For H2, we chose to explore a change in accuracy of 6% (accuracy moving from 50% to 56% or 44%, OR = 1.27, *d* = 0.133), again with 95% power and an alpha of 0.05. This was calculated via G Power for a logistic regression model as requiring 771 participants per-condition (Faul et al., 2009).”

4. You need to ensure that your data will still be informative if you arrive at a null result (e.g., p > .05) rather than stating that it is not “statistically significant/different”. This requires either Bayesian analyses or equivalence tests. Please see our Stage 1 criteria for guidance on this,

which includes supporting references that will be of use. Note that this will have implications for your power analysis too: a power analysis should be based on the analysis tests you will conduct.

Response_4: See our Response_2.

Review criteria 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).

1. You will see from the Reviewers that there are different perceptions of including an attention/manipulation check in this study. On one hand, the study is short which allows the Authors to speculate that participants will be attentive. On the other hand, we know that data quality from online crowdsourcing platforms can be impacted by careless responders (see Jones et al., 2022). This should be carefully considered. One reviewer suggests adding an attention check at the end of the experiment which asks participants to select the statement they have just seen and to analyse the data with those that fail this included and excluded. I agree with this, but also highlight that you will need to consider such exclusions in your sample size planning: in Jones et al. (2022), we estimated that around ~12% of participants were careless responders through failed attention checks or implausible response times, which can give you a base estimate for your own study.

Reference:

Jones et al. (2022). Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. Experimental & Clinical Psychopharmacology, 30, 381-399. Link: https://pubmed.ncbi.nlm.nih.gov/35130007/ OA: https://psyarxiv.com/rs9xh

Response_5: Based on this feedback and some comments below we have decided to add two data quality checks, and have increased the planned number of collected responses in order to factor in for an anticipated rate of responses failing these checks. One thing is that, if we are to go down this route, we would prefer to only analyze data from participants passing the various checks. This is in order to keep the length of the manuscript manageable given all of the added material.

We have added the following to the manuscript in line of this feedback (p.9):

“Two data quality checks will be performed. First, we will exclude data from participants completing the experiment in under one minute. Based on data from a similar previous study, we expect this to lead to around 3.5% of all data collected being excluded (Newall, Walasek, & Ludvig, 2020b). Second, methodologists have recommended the use of self-reported carelessness checks, such as, “In your honest opinion, should we use your data in our analyses in this study? (Do not worry, this will not affect your payment, you will receive the payment code either way.)" (Brühlmann et al., 2020). This text will be included after the rest of the experiment, and all participants responding with "no, please do not use my data" excluded. Previous data have suggested that up to 11.7% of crowdsourced responses might be careless (Jones et al., 2022), although previous data with that exact item suggest a lower rate of 5.6% self-reported careless responses (Brühlmann et al., 2020). For the present research, we will plan for a rate of 10% self-reported careless responses. Therefore, with these two data quality checks in mind, we will plan to collect data from 1,151 participants per-condition in order to reach our planned sample size.

2. The third Reviewer also makes an important point regarding the replication of the original framing difference. Whilst acknowledging that this will increase the sample size considerably, a justification of the exclusion of this control condition is required. The Authors may want to list this as a limitation in their Discussion section at Stage 1, showing that they considered this now and not later in the review process.

Response_6: We have added a condition to replicate this original framing effect, and have added justification as follows:

"However, seeing as how replication is an important aspect of gambling psychology research (Heirene, 2021), a secondary aim of the present research is to attempt to replicate findings on rates of understanding and perceived chances of winning from the original studies on this topic (Newall, Walasek, & Ludvig, 2020a, 2020b)." (p.5)

"Furthermore, in following our research aim to replicate previous findings, we make a secondarily hypothesise that:
H3. Both house edge conditions will result in lower perceived chances of winning and higher rates of understanding than equivalent return-to-player information." (p.6)

3. I agree with the Reviewer that you should run some simulated data to check your analysis pipeline and to document the analysis syntax. This avoids problems later down the line at Stage 2 if a planned analysis doesn't seem appropriate and/or reduces any errors at the planning stage.

Response_7: We have added some simulated data and an analysis script to the OSF repo.

Other minor points:

- I am unsure whether you've considered the journal you'd like to submit this too given a positive Stage 2 acceptance but wanted to flag that your manuscript is not consistent with APA 7th edition style. For example, anything over 2 authors can now be stated as 'et al.' if you are planning to use this formatting style.

Response_8: The references have been compiled via APA 7 in Zotero. I believe the issue is caused by multiple papers published in the same year with the same first author but then different authorship lists. Zotero is only creating an “et al” after the first name if the entire authorship list is identical. We would prefer to leave this issue as-is without manual fixing, which could introduce some errors, and allow the production assistants at a journal to fix this if this if they use an alternative referencing style.

- Under design, the instructions to participants state: ““Imagine that you are a member of an online casino. You have played many of this online casino’s games over the last year.” Shouldn’t this be ‘these’ (online casino games)?

Response_9: A given online casino can have many games. The wording is meant to refer to a gambler who uses a single online casino, but plays many different games within that casino. We would prefer to keep the current wording to maintain consistency with previous studies.

- Please refer to Table 1, the design table, within the manuscript itself and also provide a title for this. Previous Stage 1 accepted Registered Reports can help with guide you with this.

Response_10: The design table has been titled and is now referred to in the first paragraph of the Methods section, “The PCI RR study design template is shown in Table 1.”

- This links back to what I state under Review criteria 1C, point 1, but in the design table I do not understand the following sentence: “In order to detect a reduction on this outcome from 4.1 (see main Methods) to 3.8 (SD = 1.6), with 95% power and an alpha of 0.05 requires 741 participants in each condition”. This reads as though this is a within participant design where the phrasing of the sentence reduces a Likert point average from 4.1 to 3.8, but this is not the case given your design is between-participants with no baseline measures. Can you clarify this throughout the manuscript, please? It may be that you change your power analysis given the points above, which would mitigate this.

Response_11: We have reworded this in order to better emphasize that this is a between-participants comparison:

“In order to detect a change on this outcome from 4.1 in the original condition to 3.8 or 4.4 in the alternative condition (SD = 1.6), with 95% power and an alpha of 0.05, we would require 741 participants in each condition.”

In addition, we have fully updated the study design template in order to reflect all of the changes incurred during this round of peer review.

- A minor note is that the term ‘Registered Report’ is usually written in capitals and there are numerous times that you use this term (in lowercase) in your manuscript.

Response_12: Capitalisation has now been used throughout.

Please note, you will need to download some of the reviewer's comments from the PCI-RR portal; others will be shown without the need to download a file.

I look forward to receiving your revision in due course.

Yours sincerely,

Dr Charlotte Pennington

Reviews
Reviewed by Graeme Knibb, 29 Jul 2022 11:47
This study aims to compare two different 'house-edge' messages on gamblers understanding of loss.

Overall, I am enthusiastic about this report. Throughout, the writing and methodology was clear and precise. The design and methodology have been well thought out and are well considered. However. there are some aspects that I think require further clarity. This includes further information regarding the chosen phrasing and the power analysis (please see below).

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

I have no concerns regarding the scientific validity of the research question. The study aims to assess the phrasing of 'house-edge' information. This is a valid question and I have no doubt that the proposed method will address this question effectively.

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

The hypothesis seems reasonable and logical. No specific direction is predicted, which is fine given that this is the first study assessing wording variants. However, some further information regarding this phrase could be included. For example, why was this particular phrasing chosen? If this was based on psychological theory, then this could be outlined. Or perhaps this phrasing was based on previous research in other domains which have assessed the effect of such phrases? Or the work of the Victorian Responsible Gambling Foundation?

Response_13: Our aim was to compare the exact wording used in previous experimental research (original phrasing: "This game keeps 10% of all money bet on average"), and the wording proposed by Livingstone et al. (alternative phrasing: "on average this game is programmed to cost you 10% of your stake on each bet").

This choice has been justified via additional explanation in the Introduction, for example (pp.5-6):

“In order to maximize the present research’s usefulness to policymakers, an experimental comparison will therefore be made between these two exact phrasings of house edge information from the previous literature. We are aware that they differ across several dimensions, which means that any significant differences found here should be subject to follow-on work exploring precise mechanisms. While there is some reason to think that the longer alternative phrasing may be more effective, we do not believe that there is sufficient evidence to support a strong directional prediction at this time.”

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

The methodology and analysis plan are generally clear and appropriate. There are some aspects I would like clarification on (even just to satisfy my own curiosity). The use of an ordered logistic regression is interesting- what is the reason for using this approach over a simple t-test (or non-parametric equivalent)?

Response_14: Given this comment, and also feedback from both the Editor and Reviewer 3, we have decided to switch this analysis to one based on ordinary least squares (equivalent to a t-test).

I think there could be more clarification regarding the power calculation. Firstly, the power calculation is based around a reduction of a response on a 7-point Likert scale from 4.1(SD= 1.6) to 3.8 (SD= 1.6), why is this? Was this reduction based on any previous research or deemed to be meaningful in some way? This power analysis is said to require 741 participants, so why are significantly more participants being recruited?

Response_15: See our Response_2. The sensitivity analysis requires a sample size of just under 1,000 participants per-condition.

Further information from other sections could also be included within the power analysis discussion. For example, the fact that the study is not powered for the ‘understanding’ variable (which I think is understandable) could be included here rather than within the measures section. A clear statement regarding why the study was not powered for this dependent variable could be included (the same information presented in the table at the end of the manuscript perhaps).

Response_16: A power analysis has now been added for this dependent variable; see Response_3 for further details.

Finally, the authors propose some exploratory analyses. These are fine and will be interesting to assess. Can the authors clearly state within the registered report that these analyses will be labelled as exploratory in any future publication?

Response_17: The exploratory analyses proposed in the initial submission have now been introduced as follows (p.12):

“Finally, we plan some exploratory analyses, investigating H1 and H2, which are the most novel aspects of the present research. These will be marked as exploratory in any future publication.”

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.

There is sufficient detail to replicate the study. Although, as highlighted above, it wuld be beneficial to state that exploratory analyses will be clearly stated as such in any future publication.

Response_18. See Response_17.

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 authors have considered these aspects well. There is discussion of potential ceiling effects regarding the ‘understanding’ outcome and this is being addressed. They have considered the use of attention checks, and I agree with them that this is not necessary. They provided an example experiment for review. The study was short and to the point which should mitigate issues regarding online recruitment.

Finally, I want to commend the authors on what is a well-considered and produced registered report. This work is strong and, in my opinion, only requires minor clarifications. Thank you for the opportunity to review this piece of work.

Best wishes,

Graeme Knibb

Response_19. Thank you.

Reviewed by Zhang Chen, 01 Aug 2022 14:54

Thank you for the opportunity to read this interesting registered report. This RR aims to address an important question, with sound and feasible methodology. The planned analyses are described clearly, to ensure that the results can address the question of interest. Below I list some detailed comments based on the Stage 1 RR criteria. Hope my comments will be useful to the authors.
1. The scientific validity of the research question(s).

This registered report aims to examine whether different ways of communicating the house edge in gambling would influence gamblers' perception and understanding of risk. This research question follows directly from the previous work on comparing the house edge format versus the return-to-player format, and has clear practical implications. The research question is therefore scientifically valid, practically relevant and important.

2. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.
As I understand it, the authors do not have a strong directional hypothesis for which format will be more effective. The untested format has more words than the original one, but it "contains additional words which might either increase the perceived severity of the resulting average gambling losses, or improve gamblers' comprehension of this information." (Page 4). I think a discussion in the introduction on how these factors might influence the effectiveness of a message will be useful. For instance, has previous work (maybe from other domains, such as food or alcohol consumption, or the literature on persuasion more broadly) examined how the length of a message may influence its effectiveness? And what are these 'additional words' exactly that may increase the perceived severity of gambling losses, and why?
Discussing these issues will have implications for the interpretation of the results. In case the results show that one format is more effective than the other, it would directly support the use of the superior format in practice (Page 10). However, both researchers and practitioners will probably want to know what are the elements that make one format more effective than the other. This is a valid question, as it may serve as a starting point to further optimize such messages. As such, a somewhat more extensive discussion on the potential differences between the two formats and why these differences may matter, even though speculative at this stage, will still be useful.

Response_20: This is a good point, which have introduced extra material to the introduction to cover. We are not aware of any relevant research from other public health domains that could speak to this issue. Our own previous gambling research on "volatility warnings" does suggest some credible possibilities of increased effectiveness from longer warnings. However, we believe that that finding is rather circumstantial in this instance, and is not sufficient to make a strong directional prediction. This has been reflected in the introduction as follows (pp.5-6):

"One limitation of this literature is that previous experimental research on the house edge format uses the same way of phrasing this information. This issue is important, as at least one alternative phrasing has been proposed: "on average this game is programmed to cost you [10]% of your stake on each bet" (Livingstone et al., 2019; p.3). This phrasing is longer, at 16 words compared to nine words, and contains additional words which might either increase the perceived severity of the resulting average gambling losses, or improve gamblers' comprehension of this information. Previous work suggests that added explanation can alter how gamblers evaluate this information. For example, the addition of a 32-word "volatility warning" significantly decreased gamblers' perceived chances of winning with both return-to-player and house edge information (Newall, Walasek, & Ludvig, 2020b). In order to maximize the present research's usefulness to policymakers, an experimental comparison will therefore be made between these two exact phrasings of house edge information from the previous literature. We are aware that they differ across several dimensions, which means that any significant differences found here should be subject to follow-on work exploring precise mechanisms. While there is some reason to think that the longer alternative phrasing may be more effective, we do not believe that there is sufficient evidence to support a strong directional prediction at this time."

3. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).
The methodology is sound and feasible. For the power analysis, it is not entirely clear to me what the planned analysis is. Since the authors mentioned means and standard deviations of the responses on the Likert scale (Page 5), I suppose the authors computed Cohen's d as the effect size, and used a t test as the planned analysis to determine the required sample size. However, the planned analysis for the Likert scale data is ordered logistic regression. Thus, strictly speaking, the power analysis will also need to use ordered logistic regression as the planned analysis.

Resonse_21: We have now switched this analysis to OLS, as explained in Response_2 and also consistent with some feedback below from Reviewer 3. The power analysis and planned model are now consistent.

For the exploratory analyses, AIC values will be computed for the models with and without the interaction term, and models with the lowest AIC will be considered to provide the best fit to the data. I wonder if there are cutoff values for how large the difference between two AIC values needs to be for one to select one model over another.

Response_22: First, the switch from ordered logistic to OLS obviates this problem for hypothesis 1. Second, another recommended way of dealing with this issue with interactions in nonlinear models, is to switch to OLS (Ai & Norton, 2003). In order to simplify things, we have decided to therefore do both exploratory interaction models in OLS. This is explained in the text as follows (p.12):

"Two exploratory analyses will be run to see if there are any interaction effects between the phrasing of house edge information and PGSI, in order to detect whether the optimal phrasing of house-edge information might depend on gamblers' level of problem gambling severity. Two extra regression models will be run, adding a main effect of PGSI and an interaction between PGSI and experimental condition. Since p-values on interaction terms in non-linear models are not always interpretable (McCabe et al., 2020), the model for hypothesis 2 will also use ordinary least squares, as this has been recommended as a way of counteracting this issue (Ai & Norton, 2003)."

4. 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 authors provided a link to the proposed study, which I have tried out. The method is described in detail and accurately in the manuscript, and is sufficient to closely replicate the proposed procedures.
The proposed analyses are also described clearly. However, to further increase the computational reproducibility and prevent undisclosed flexibility, I would recommend mentioning the statistical software the authors intend to use and the version number in the manuscript. The analysis syntax for the confirmatory analyses should also be included as part of the pre-registration to further reduce undisclosed flexibility.

Response_23: This is a helpful suggestion, which we have incorporated into the revised report. The analysis script has now been included as part of the registered report.

5. 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 authors have carefully considered a potential ceiling effect in the understanding of the message, by including more options than in previous work. For the responses on the Likert scale, the floor or the ceiling effect does not seem to be of concern based on previous data. I was a bit surprised to see that no quality control or manipulation check is included (Page 7). The manipulation between the conditions is rather subtle (i.e., difference in one sentence), and it is possible that some participants may not read the text very carefully. Setting cutoff values for response time may indeed be arbitrary, but it can allow the authors to filter out some low quality responses. A memory test at the end of the experiment, asking participants to select the statement they have just seen, may also provide some useful information. For instance, this will allow the authors to explore whether there is a difference between the participants who correctly remember the message, versus those who do not. Or whether participants may differ in how well they remember the two statements.

Response_24: See our Response_5 for the data quality checks that are now planned.

Related, one possible result is that the two formats do not significantly differ on the dependent measures. I think this is still an informative result, but $p > .05$ does not provide support for the null hypothesis. To do that, alternative statistical approaches are needed, such as Bayesian statistics or equivalence tests against the smallest effect size of interest. I wonder whether the authors have considered this scenario (i.e., no difference or very small effects), and have planned the study to also be able to draw this conclusion. This may also have implications for how many participants to recruit in the power analysis above.

Response_25: See our Response_2 which describes the addition of equivalence testing to the present work in the event that p-values above .05 are observed for either H1 or H2.

Reviewed by Luke Clark, 15 Aug 2022 20:43

I appreciate the opportunity to review this Stage 1 RR and I co-reviewed this paper with one of my graduate students, who found it a useful training exercise.

The authors present a study to compare the effects of two different ‘house edge’ labels on online gamblers’ responses to a hypothetical gambling scenario. The study builds on recent work from Newall that found a single, specific ‘house edge’ label (“this game keeps 10% of all money bet on average”) was associated with superior performance relative to a ‘return to player’ label (“returns 90% of your money”). A natural question, addressed in this study, is whether the effect generalizes across different house edge phrasings. Indeed the alternative format presented here (“costs you 10%”) could lead to even better outcomes. This is an important research question that will inform gambling policy. The design, sampling, and hypotheses are clearly specified.

One difficult decision that the authors must have faced is whether to include the original ‘return to player’ label as a third condition. They elect not to do that, and recruit only two groups (original house edge, alternative house edge). I can see that the third cell would add another ~1000 participants to the study, with cost implications, and that the Prolific platform may not have the capacity to recruit so many experienced online gamblers. At the same time, the third ‘return to player’ label is currently the industry standard that the authors are looking to challenge. In the eventuality that they see no difference between the two house edge labels, I feel there would be much value gained from the basic replication of the original framing difference (i.e. 2=3>1). (Conversely, if the original label performs better, is the alternative label still sufficient to generate the framing difference?). I would be interested to hear the authors’ justification for the exclusion of the third group.

Response_26: See our Response_6 for our decision to include this format as an extra condition.

Other points

- The authors explicitly state that they will not apply any data cleaning e.g. attention checks to the Prolific data. This is a bold decision. I agree with their logic that the insertion of additional ‘attention check’ may do more harm than good (although my perspective is that the academic wind is blowing in the opposite direction). Applying data cleaning for fast completion times is, in my view, a rather different point; could there be bots in the data, or some participants who do not read the materials at all, and submit the entire survey in ~30 seconds? (Data quality on Prolific is higher than MTurk, I agree, but I’m not convinced it is “high” – pg 7). My own recommendation would be to run a sensitivity analysis using a pre-registered threshold for completion time. (I recognize any such threshold is arbitrary)

Response_27: See our Response_5 for the data quality checks that are now planned, which include a criteria based on speed as well as seriousness.

- The authors propose to use the Prolific data balancing function to balance gender. The analysis plan does not mention any gender-based analyses. While under-powered, it would be useful to test whether any observed differences are separately robust in both men and women.

Response_28: We would prefer not to include these analyses at the present time. The reason is that we are not aware of any relevant theory that would lead us to anticipate a difference between males and females in this instance. The data will be made openly-available, so that other researchers would be able to investigate these issues if they would like to.

- Statistical analysis. Given the possibility (high possibility, in my view!) that the two house edge labels will not differ, I note the analysis plan does not mention any Bayesian testing of support for the null.

Response_29: This has been amended in the revised report. Instead of a Bayesian analysis, we have remained within a frequentist framework, and adopted equivalence tests as a method for testing whether the house edge labels do not differ.

- The analysis plan on the 7 point rating proposes an ordered logistic regression. I have not encountered this technique before, and we have been discussing different approaches to analysing ordinal ratings in my lab meetings recently, so I will look into this further (note of thanks!). I raise it because the Newall 2020 Addiction paper applied a simple ANOVA to the ratings data, thus treating the rating as a fully continuous variable. If the authors have not used the ordered logistic technique elsewhere, a methods reference would be helpful.

Response_30: See our Response_1 above about our switch to OLS.