

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.

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.

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.

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.

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.

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.