Dear Dr Leganes-Fonteneau,

We submit a revised Stage 1 submission of the following article:

"*The effects of isolated game elements on adherence rates in food-based response inhibition training*"

We thank you and the reviewers for your constructive comments on this manuscript. We have now addressed all of the comments, providing our responses to each below. We believe this has resulted in a higher quality and more consistent article as a result.

We have also added a line after our hypotheses explaining extra variables are being included at baseline for undergraduate dissertation projects. However, these do not form part of any hypotheses tests or exploratory analyses we will be running, we have not given further details.

We hope this now meets the standards required for in principle acceptance as a Stage 1 Registered Report.

Kind Regards,

Alexander MacLellan and co-authors.

**Reviewer 1**

**Comment**: Before I start my review, I must disclose that although I am not completely unfamiliar with this literature, I do not consider myself an expert in cognitive training programs for unhealthy eating and I can't judge the more conceptual and theoretical aspects of the proposal. As far as I can see, the protocol addresses an interesting topic and in general the proposal seems well designed. Perhaps my most important concerns are related to sample size and power analysis. In fact, I wasn't able to understand what's the sample size the authors are planning to recruit. On pages 7-8 the authors state that "Detecting effect sizes of f=0.23 … would require 80 participants per group… which was deemed achievable with our resources… this was selected as our target sample size". But just a few lines below… "Our total sample size is therefore set at 150 to detect effects in our primary and secondary hypothesis". So, is it 50 per group or 80? To make things more complicated, in the final table in the supplementary material states "we propose to recruit 51 participants per group" and in the next row "we propose to recruit 30 participants per group". Maybe I missed something, but I wasn't able to follow any of this.

***Response:*** We thank the reviewer for this important comment as it pointed out an inconsistency in our initial report, which was a remnant of the multiple iterations of sample size calculations we conducted for this study. These have now been revised throughout the document for consistency, and for clarity here we declare that our target sample size is set at 80 participants per group, with 240 in total across the three training groups.

**Comment**: Also, in general, the structure of the power analysis section looks a bit awkward to me. The authors begin by discussing reasonable effect sizes that could be expected based on previous studies, but in truth their sample size is not based (a priori) on any of those estimates. Instead, it looks like sample size will be mainly determined by the availability of resources. Therefore, perhaps this section would be much clearer changing the other of ideas. Perhaps along the lines "With our resources we can afford to test X participants. With this sample size we can detect an effect size of X with 90% power. This effect size is reasonable based on previous evidence." In other words, if sample size is based on resource

availability, it doesn't make sense (I think) to present the power analysis section as an a priori power analysis (i.e., based on effect size X we need Y participants); a sensitivity analysis is possibly more appropriate and clear (i.e., with the Y participants we can afford, we have reasonable power to detect X).

*Response*: We thank the reviewer for this detailed consideration, and have incorporated the feedback into our draft, on page 7. As suggested, we have clarified that our sample size estimation is based on striking a balance between available resources and plausible effect sizes of interest, we have amended, as follows:

"With our current resources, we estimate it is possible to recruit 80 participants per group, for a sample size of 240 in total. This would allow us to detect an effect size of f = 0.23 with 90% power. Given previous literature finding a large effect of gamification on task engagement, g = 0.72 (which we approximate to a Cohens f value of 0.36), with no evidence of publication bias (Vermeir et al., 2020), we believe this to be an appropriate target sample size that would yield informative results." – Page 7

**Comment**: Another detail of this section (and other bits of the text) is the constant change from f-units to d-units. If the final sample size is going to be, say, 50 per group, then it would be nice to have a simple sentence explaining what's the smallest f that can be detected across 3 groups with 50 x 3 participants and what's the smallest d that can be detected with in a pairwise comparison with 50 x 2 participants.

*Response*: We thank the reviewer for these comments; they have helped us clarify our thinking and we hope to have addressed these in turn. In brief, all effect sizes discussed are now clearly signposted and other than a discussion of the effect sizes and power for our TOSTER analysis, we refer to *f* values in the first instance when discussing power, for example:

"an effect size of f = 0.24 was estimated for devaluation scores based on the previous work of the authors (from a d = 0.48, Lawrence et al., 2015)." – Page 8

**Comment**: Also, wouldn't it be more appropriate to explain the power for the TOSTER equivalence test on this same section instead of presenting it on p. 14? I must confess that a minimal effect size of d = .6 doesn't sound terribly convincing for TOSTER. Technically, this choice means that the authors consider d = .6 too small to matter, but the average observed effect size in psychological research is around d = .5 (see, e.g., Bakker et al. 2012 PPS). So, in principle this logic implies that the authors consider that most effect sizes reported in psychology are irrelevant!

*Response*: We agree with your comment and have amended our analysis plan accordingly. We have now determined that we can detect a minimal effect of *d* = 0.46 with 80 people in each group at 80% power from a power analysis using the TOSTER package and have stated this in the report on page 8 and in the supplementary table, as follows:

"Finally, from a power analysis using the TOSTER R package (Lakens & Caldwell, 2021), we would be able to detect equivalence within the parameters d = -0.46 and d = 0.46 at 80% power with a sample size of 80 per group. We have been more lenient with our target power in this analysis to target relevant effect sizes which correspond to our previously stated effect size of interest (converting from f values of 0.23)." – Page 8

**Comment**: I understand that this is a somewhat biased comment, but the authors might want to mention that one of the shortcomings of this area of research is, precisely, that statistical power is often too low (Navas et al., 2021, Obesity Reviews).

*Response*: We have decided not to include this reference in our methods section because our paper does not focus on the methodological issues present in the prior literature, but rather on the seemingly robust effectiveness of response inhibition training on eating behaviour. Low statistical power has been a problem inherent in many previous studies across research disciplines; a problem that the Registered Reports publication format itself aims to mitigate. However, this will be a relevant citation for when we discuss the strengths and limitations of our study in the stage 2 report.

**Comment**: In RQ1 and RQ2, the authors plan to test their hypothesis with one-way ANOVAs, but I imagine that if they find a significant result they will want to follow up on this with pairwise analyses. Shouldn't these be mentioned in the analysis plan?

*Response*: The reviewer is right to point out this oversight, and we have now specified this in the corresponding sections of the paper (pages 13 and 14). However, RQ2 will be tested with a mixed 3x2 ANOVA so we have specified that follow up analyses will be conducted if the interaction term, and main effects are significant, for example:

"Should this result be significant, we will follow this up with independent-samples t-tests to investigate the direction of effect found." – Page 13

**Comment**: In RQ3, the authors plan to test if motivation/adherence mediate the effects of the intervention on food evaluations, but wouldn't it be even more interesting to test the mediating effect of these variables on actual snacking?

In the same vein, in RQ4 the authors only plan to test whether both interventions are comparable in terms of motivation/adherence, but wouldn't it be more interesting to test whether they are similar in terms of effectiveness? (i.e., in terms of food evaluations and snacking?)

*Response*: We take the reviewer's comments and have included their suggestion to explore the potential mediating effect of adherence on snacking frequency in our plans, with these changes included on pages 6 (including them in the hypotheses) and 14 (including snacking frequency in the analysis) as follows:

"H3c – Pre- to Post-intervention differences in snacking frequency will be mediated by training adherence.

H3d - Pre- to Post-intervention differences in snacking frequency will be mediated by training motivation." – Page 5

"… change in food item evaluations, and in snacking frequency, as the outcome. Secondly the direct effect of intervention group on change in training engagement or motivation will be established, followed by establishing the indirect effect with both intervention group and change in the mediator as our predictor variables and change in food evaluation score, and snacking frequency, as our outcome variable." – Page 14

We did not however include any confirmatory hypothesis testing of equivalence between the gamification groups for snacking frequency or food evaluation scores given there is a lack of previous work investigating the effect of single element gamification on training effectiveness in this area. We feel it would therefore be inappropriate to specify a confirmatory hypothesis at this stage, but will include this as an exploratory test to inform future research.

Minor comments

**Comment:** The authors will run frequentist and Bayesian analysis, which I think is great. But what will they conclude if different analyses lead to different conclusions? In the same vein, the authors state that they will run all the analyses both including and excluding participants who fail the attention check. But what will they conclude if the results are not identical? In general, it is not a good idea to have multiple confirmatory tests for the same hypotheses in Registered Reports, as this leaves too much analytical flexibility and provides more opportunities for biases in the interpretation of results. If the authors think that excluding participants is best (or that frequentists statistics are more appropriate) they should probably stick to those analyses in the pre-registered protocol. This doesn't prevent them from presenting additional analyses in the exploratory section. But ideally the authors should state a priori what are the analyses that in their opinion provide the strongest test for their hypothesis.

*Response*: We thank the reviewer for this very helpful comment, and in response to both reviewers, we have now removed our intention to perform a Bayesian analysis because we understand that the Two One-Sided Tests of Equivalence (TOSTER) analyses will allow us to state whether effects are within the equivalence bounds and therefore not meaningful or inconclusive and worthy of further investigation. Whilst the intention was to include both methods of analysis to lend additional confidence to any findings, we believe that such frequentist methods will be sufficient. We have revised the manuscript and Table is S1i accordingly. As for excluding participants who fail the attention check, our primary analysis will include those who fail the check to include our full sample and protect against issues with generalisability, though we will also check if excluding those who fail the manipulation alters the results, in line with suggestions from previous work looking at the prevalence and effect of careless responding (e.g. Jones et al., 2022).

**Comment**: The second paragraph on page 8 mentions for the first time "secondary" analysis. Although these hypotheses are not of primary interest, maybe something about them should be explained at the end of the introduction, so that the reader knows that further tests will be run before they reach this paragraph. This will also help the reader understand the "exploratory outcome variables" section on p. 11.

*Response*: We have removed reference to primary and secondary hypotheses because such phrasing may inadvertently imply that some of our hypotheses were not an initial consideration of the study. Given we have made reference to the effect of gamification on both adherence (e.g. Najberg et al., 2021, Aulbach et al., 2021) and food evaluation and snacking (Forman et al., 2019) in our introduction (namely on page 3), we hope this is not the case. We have however added a sentence specifying our intention to conduct exploratory analysis in line with your suggestions on page 6, shown below:

"Given the lack of previous work on the effect of gamification on specific components of motivation, and potential equivalence of training effectiveness between single task gamification groups, we do not propose to test any hypotheses, however, we do state our intention to explore the effects of gamification here to inform future research."

**Comment**: P. 10. Participants will be asked to report their confidence in their food evaluations. Is it possible that participants prefer one food to another but with little confidence?

*Response*: This is possible, though the evaluation task refers to each food image individually, rather than choosing one food over another. The confidence ratings therefore refer to their judgement of their evaluation of each specific food, rather than about whether they preferred one food to another. Further to this, the confidence rating does not form part of our hypothesis tests, but rather will be reported descriptively.

**Comment**: p. 12. Isn't it weird to remove participants who perform the task too well? (2 SDs above the mean?)

*Response:* We agree with the reviewer comment here and so have removed this data exclusion criteria.

**Comment**: Appendix B. What's the effect size unit in the power curve?

*Response:* The effect size unit (f) has now been added to the graphs.

**Comment**: Final table, first row. In the sampling plan the author present a g = .72 as reference but the corresponding analysis is a one-way anova with 3 groups, for which cohen's g is undefined (to the best of my knowledge). Note also my previous concerns about sample size and power analysis, as they apply to this table as well.

*Response:* We thank the reviewer for pointing out this inconsistency, which we hope to have now clarified. You can see this in the supplementary table on pages 26-29 and as below:

"Based on an effect size of f = 0.23 we propose to recruit 80 participants per group, 240 in total, to detect between group differences on adherence and engagement rates."

**Reviewer 2**

This registered report tests several response inhibition techniques with varying forms of gamification. I find the overall research topic to be valuable and interesting, and the manuscript thus far to be well written and informative. My comments on the manuscript are as follows:

Main Comments

**Comment**: The authors might consider explicitly adhering to a reporting guideline for trials, such as SPIRIT or similar. It appears the content of such checklists is largely covered in the manuscript, but an explicit report of a checklist may add value to the already strong open science basis of this trial.

*Response*: First of all, thank you for your positive appraisal of our article and for your constructive and helpful comments. To clarify, this study is not a 'trial' but rather a randomised study and, and the relevant items on the checklist we believe have been

adhered to in this manuscript. Furthermore, we feel that submitting this research study via the Registered Report route mitigates any potential bias, such as undisclosed flexibility in analysis, which the SPIRIT guidelines also aim to achieve.

Having reviewed the 31 SPIRIT checklist, we feel this is a positive guideline for this project and will produce a checklist for this study to publish alongside this as a stage 1 RR after IPA on the associated Open Science Framework repository. Given the time sensitive nature of this project, it is not practicable to create this ahead of resubmission.

**Comment**: For power analysis, please include the specified alpha value (I assume .05?). Otherwise, I accept the authors explanation. I also suggest for the less informed reader that the authors note .23 constitutes a medium effect size in Cohen effect size taxonomy, give using this taxonomy is another common method of arriving at power estimates.

*Response*: The alpha value of .05 has now been specified at the beginning of this section on page 7 and the clarification that an f value of .23 would constitute a medium effect size has now been added on page 8.

**Comment**: I agree with the authors choice of measures. Another potentially valuable addition here might be some measure of automatic or implicit attitude towards target foods such as the implicit association test or affect misattribution procedure. I believe templates for these measures are available on Gorilla already if the authors choose to make use of this suggestion.

*Response:* We thank you for this comment, though after consideration we have not included an implicit attitude test in our revised submission based on previous literature (including our own) that has found limited support for the proposition that implicit attitudes are sensitive to this training (e.g. Yang et al., 2022). Given our desire to keep this protocol as short as possible to make best use of our resources, we do not believe adding the measure would be practical.

**Comment**: Will training and data collection be restricted to any particular type of device? i.e., will training be required to be conducted on a computer or is a touch screen version for tablets or phones available? If there is a restriction I suggest noting this.

*Response*: Data collection is restricted to those with a stable internet connection and a personal computer with a physical keyboard. This has been added to the study inclusion criteria on page 7.

**Comment**: I have not used the particular Bayes package in question, but details on the default priors is somewhat essential here, especially as many default priors are uninformed. If this is the case you would expect almost identical results using Bayes, so I am unsure of what additional value this analysis adds. Give the authors cite several previous tests and analyses, could these be used as priors?

*Response:* After consideration of reviewer comments, we have now removed our plan of carrying out Bayesian analysis.

**Comment**: Minor Issues

- Some minor typos throughout, e.g., in H2c “the” is missing

· Given there are so many forms of food frequency questionnaire out there, I would suggest referring to the measure here as “a unhealthy snacking based food frequency questionnaire” or something similar, rather than “the food frequency questionnaire”

· RQ3 ANOVA says 3x2, but lists 4 groups. I assume intervention is the duplicate as its covered later by the actual group names.

*Response:* Thank you for highlighting these errors and suggestions in the paper, all of them have now been corrected.