Response to reviewers Round 2, PCI Registered Reports #493
Please find below our reply (in blue) and the main part modified in the manuscript (in green). All changes are also reported in blue in the new version of the manuscript.

Dear Dr. Najberg,

Thank you for submitting the revised version of your Stage 1 RR, now titled "Sugary drinks devaluation with response training helps to resist their consumption" to PCI RR. Two expert reviewers who have assessed the initial version have now re-reviewed the revised manuscript. Most of the previous comments have been addressed satisfactorily. However, there are still some remaining issues that are not fully resolved yet. I would therefore like to invite you to further revise the manuscript to address these issues.

We thank you for your and the reviewers' review of our manuscript. We have addressed each of the comments bellow.

1. One main issue is that sample size justification is not entirely clear. I appreciate the increased sample size, which will for sure make the results more informative. However, this does not really answer the question of how the smallest effect sizes are determined. For H1, a difference of 5 days (with an estimated standard deviation of 10) is said to be relevant in an applied setting. But what are the exact reasons and/or justifications behind this choice? In the response letter, you mentioned that this decision was made based on discussions with board certified dieticians and your own previous studies on item valuation. I think the exact content of the discussions, and the previous effect sizes, ought to be mentioned in the manuscript. For instance, is this effect size determined by comparing it to other existing interventions, or based on certain guidelines in the field? Such information will allow readers to be better informed on what this effect size exactly entails in the current context. Similar issues exist for H2 and H3, where r = 0.4 is the smallest effect size of interest. At the moment, this is said to be "based on clinical subjectivity", but it is unclear what this means.

To reduce context-dependency and subjectivity in the choice of our SESOI, we have now opted to rely on a sensitivity power approach (cf our replies below). The main aspects that were reported by our consultants as benefiting of extending the diets for 5+ days were the following. These points remain, however, highly context-dependent and a general-purpose justification can thus not be fully backed up by evidence.

- Physiological Changes: Achieving specific metabolic states like ketosis or autophagy usually takes five days.
- Detoxification/Medication Protocols: Some diets align with detox or medication five days timelines.
- Safety Buffer: The added five days can allow the body to benefit after initial adjustment to the diet.
- Patient Feedback: Patients might report better outcomes with an additional week.

- Biomarker Changes: Certain biomarkers might show significant change only after five days.

In the response letter, you additionally mentioned that you had to take into account the resources available for this project, which is of course a constraint that we often face. Basing the sample size on resources available is completely legitimate (resource constraints, see https://doi.org/10.1525/collabra.33267). In this case, an alternative approach may be to start with a maximum sample size allowed by resources, and then conduct a power sensitivity analysis, to see what is the smallest effect size that can reasonably be detected. Again, it is important to put these smallest detectable effect sizes into context, e.g. explain what they entail in clinical and applied settings, and/or in relation to previously observed effect sizes.

Following this comment, we have modified our a-priori power analyses to power sensitivity analyses, before discussing the relevance of the smallest detectable effect sizes. This new sampling plan justification can now be read p4.

Related, I think the "Rationale for deciding the sensitivity of the test" in the design table is relevant, and should contain the justifications that are currently missing. I would therefore suggest putting this column back.

A column has been added in Table 1 to address this comment, now titled "Interpretation of the smallest detectable effect size".

2. Some aspects of data analysis are a bit vague to me. To make it more concrete, can you please generate some 'fake' data and use it to write down the R code that you would use to analyze the real data?

We agree and now give a script going through the different analysis steps. We included randomly generated data to facilitate the comprehension of the script. This script can now be found in our OSF page under the "SCRIPT" folder: (https://osf.io/s4trh/?view_only=4934c0215f2943cfb42e019792a30b53).

By generating this script, we realized that the only positive controls that demanded eventual exclusions were specific to H1. As such, the analysis plan section was slightly restructured for a better flow.

Some more specific questions include:

2.1 This may be due to my own lack of understanding. For H1, you wrote that you would apply the Greenhouse-Geisser correction. However, I've only seen the GG correction for repeated measures, whereas H1 involves two independent groups. Also, it is unclear to me how the GG correction would then be combined with an independent t-test.

The recommender is correct. The GG correction is an error from an earlier draft that we forgot to update. The Welch t-test is robust to inequal variances between groups (doi: 10.5334/irsp.82), so no correction will be made in case of heteroskedasticity. This has now been removed from the manuscript.

2.2 If I understood it correctly, participants that were excluded based on 2.5 MAD range (i.e., 'distribution outliers') would not be replaced. Please mention this explicitly in the manuscript.

This is now clarified in the manuscript, p9: "Excluded participants (i.e., dropouts, distribution outliers, positive controls exclusion) will not be replaced because of resource constraints […]".

2.3 Positive controls. The exact steps for removing participants are a bit vague, so writing down the R code would really help. One reason for not replacing distribution outliers (see 2.2) is that the thresholds may change each time when participants were replaced. I wonder whether the positive controls would not potentially create a similar 'circularity' problem? After all, Cohen's d between groups is computed on all data points, which is essentially the same issue that one may face when computing e.g. MAD? Also, data collection for this project is very time-consuming (up to about 3 months), so in case there is a need to repeatedly replace participants (e.g., if by replacing some participants, new participants will need to be replaced), the data collection phase might be very long.

In Najberg et al., 2021 and 2023, we had to exclude little participants to respect the positive controls (5 and 4 participants out of 95 and 185 respectively), and our positive controls were more stringent than this current study. On the other hand, distribution outliers were more frequent. This is why we decided to treat the replacement of positive controls and distribution outliers differently. The former seems rarer and affects the interpretation of the results, while the latter is more frequent but only improves the centrality of the mean. So we are confident in our ability to replace positive controls, by simply recruiting a few more participants.

However, according to the recommender's suggestion of adopting a resource-based decision on sample size, we have decided to stop the recruiting after having recorded n=140 participants instead of going for a potentially difficult to follow replacement strategy. This can be read p9: "Excluded participants (i.e., dropouts, distribution outliers, positive controls exclusion) will not be replaced because of resource constraints (see Sampling plan section). The study will stop recruiting after having 140 participants with complete data (i.e., all questionnaires filled)".

2.4 For Bayes factors, please add the priors that you are going to use into the manuscript itself. I also agree with some previous comments from reviewers, that you should report Bayes factors for all results, not just for null results. You can still specify that the statistical inference will be based on p values. However, adding Bayes factors to all results does not seem to complicate the results too much (it's just BF = x for each effect), but does provide a more complete picture of the results.

According to the suggestion of the reviewer, we will now report BF01 for all analyses, as well as the default priors. It now reads p9: "All results will be interpreted using frequentist statistics, with Bayes Factors against the null hypothesis (BF01) reported as a supplementary information to support non-significant results. The BFs will be computed using the BayesFactor R package with default priors. Please refer to the package manual for details on the priors (https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf)".

3. One reviewer (Dr. Van Dessel) questioned the use of a strict cutoff value for determining the 'relevance' of a finding. I have a related question, which is about how this inference will be made. Will you formally test this, for instance by comparing the 95% confidence interval of the estimate, and declare an effect as relevant only when the lower bound of the CI exceeds the cutoff value? Or would you simply look at the point estimate itself and see if it exceeds the cutoff value or not? I think the former approach is the better one, but it will likely require a much larger sample size (the CI will need to be rather narrow). The second approach is not principled, because there is uncertainty in the point estimate. As such, reporting the estimates and the associated variations, and putting these estimates in the current context (i.e. what they mean in the current setting, see Point 1 above) seem like a more nuanced approach.

As we have modified the effect size justification to be based on resource constraints, this relevance cutoffs are no longer discussed as such.

The goal of this relevance thresholds was to supersede the p-values. As such, the observed effect size (point estimate) would have been used instead of the lower bounds of the 95%CI. This is because the p-value is derived from the 95%CI and would thus have run against our aim to emphasize effect size threshold over an alpha threshold.

Some more minor issues:

1. Title: This is a matter of personal taste. At the moment, the title assumes that the training will be effective, which may not be the case after the data is in. You may want to pose the title as a question, and also add 'Registered Report' to highlight that this is a RR.

We propose to adopt the optimistic choice of leaving the title as it currently is, and to adjust it to the "Sugary drinks devaluation with response training does not help to resist their consumption" at stage two if our hypothesis is not confirmed. Regarding the inclusion of RR, we feel it would unnecessarily lengthen the title since this information is usually reported just before the title on the editors' websites (e.g. https://royalsocietypublishing.org/doi/full/10.1098/rsos.191288 or here https://www.nature.com/articles/s41598-023-36859-x)

2. Page 2: "There is, however, little evidence supporting real-life effects of cognitive bias modification". Cognitive bias modification sounds like a very general term. Since this paragraph is about the cue-approach training, you may want to modify the statement to be more specifically about the CAT.

Following the suggestion of Dr. Van Dessel, this sentence has been removed from the manuscript to allow a stronger focus on the tasks' observed findings instead of their mechanisms of action.

3. Page 3: 'it is easier to report and less biased by memory …' than food frequency questionnaires and food journals?

This sentence indeed referred to FFQ and food journals. This was however revised to address a comment of Dr. Van Dessel.

4. Page 3: "letting the participant stop their training whenever they want in a two-weeks window enables to investigate the link of the intervention's length on its real-world effect size, thereby allowing to formulate recommendations for its use in applied settings." I believe the training window is now 20 days?

Indeed, the minimum required is 7 days, and the maximum 20 days. This corresponds to a window of 13 days (ca. two weeks).

Furthermore, the reviewers correctly pointed out that H3 cannot be explained as causal effects, but the implication (i.e., formulating recommendations on how long the training would be) still implies a causal interpretation?

We agree and now have removed in the manuscript such implications.

5. Page 6: There is a mention of ECT.

This has now been corrected.

6. Page 7: The section heading for the CAT says 'Attentional bias modification'.

This has now been corrected.

7. Page 8: Table 2 and 3 should be Table 3 and 4.

This has now been corrected.

8. Page 12: The content of Table 1 is not updated. The 'Interpretation given different outcomes' only depicts one possible outcome, but I think you really ought to list all possible outcomes and explain how you would interpret each of them.

The different outcomes for non-significant results supported or not by the BF01 are now discussed in this table.

The interpretation for H3 implies a causal interpretation, but I do not think the current data can support that.

We agree and now have removed in the manuscript such implications.

*by **Zhang Chen**, 20 Sep 2023 08:52*
Manuscript: **https://osf.io/e68ja?view_only=4934c0215f2943cfb42e019792a30b53**
version: 2

Review by Pieter Van Dessel, 08 Sep 2023 11:37

I extend my appreciation to the authors for their receptive approach to the suggestions provided. Overall, the authors have made substantial improvements to the paper, resulting in a notably enhanced manuscript.

Here are several observations I made while reading the revision of this manuscript:

1. Introduction Clarity:

The introduction section contains several areas that could benefit from further clarity. It would be advisable to take out any reference to cognitive processes, as the primary focus appears to be on evaluating effectiveness rather than delving into cognitive explanations.

Specific points of concern include:

Page 2, paragraph 1: The statement, "Interestingly, recent evidence indicates that the practice of tasks involving the execution or inhibition of motor responses to food cues modulates their self-reported value, and their consumption," could be nuanced to indicate that the evidence **suggests** these practices **can** modulate these variables. In general, it is best to avoid making strong claims or clearly outline evidence supporting such claims in the event that there would be strong enough evidence.

We have modified the manuscript according to the comment of the reviewer. It now reads p2: "Interestingly, recent evidence suggests the practice of tasks involving the execution or inhibition of motor responses to food cues can modulate their self-reported value and their consumption".

Page 2, paragraph 2: "The repeated inhibition of motor response to unhealthy cues is thought to reduce their reward value to solve the conflict between the task demand for response withholding and their tendency to respond to palatable cues". This sentence is very complex and it is not well explained (what is this "reward value" or this "tendency"). It would also be crucial to specify the source of these theoretical explanations as there are many theoretical explanations of observed effects. It might be better if the authors omit this sentence and instead of talking about these cognitive explanations simply explain the procedure and the findings more. The same holds for the next paragraph on the cue-approach-task.

Following the reviewer comment, we have omitted the problematic sentences and explained the procedure and finding more. It now reads p2: "In the food Go/NoGo (GNG) task, participants have to respond as fast as possible to healthy food cues, while withholding their responses to target unhealthy food cues. The practice of these tasks have been shown to reduce the self-reported valuation of the target NoGo unhealthy items, as well as their in-lab and self-reported consumption (see 14,15 for discussions on the underlying cognitive mechanisms of action).

In the Cue-Approach Training (CAT), participants have to respond to items when a Go-cue is displayed. Importantly, the Go-cue appears after the item, and the item disappears rapidly after the presentation of the Go-cue. The practice of this task has been shown to increase the self-reported value of the

trained Go items through preference tasks, snack auctions, as well as their consumption during bogus taste tests (see 20 for a discussion on the supporting cognitive mechanisms)."

The fourth paragraph is relevant as it goes into evidence of effectiveness. However, the next paragraph again goes into the cognitive processes (making some strong claims without presenting evidence and without explaining the cognitive constructs well). Consider omitting paragraph 4 as this would streamline the introduction and allow it to flow more smoothly into the next paragraph, which is well-written and clear.

The paragraph mentioned by the reviewer has now been removed.

Page 3: There is again a strong claim, this time the claim is that adherence to a restrictive diet is valuable due to it being easier to report and less biased. This claim requires substantiation with references or additional context or should be omitted.

The claim in question has now been omitted.

2. Sampling Plan and Effect Size:
The section regarding the sampling plan raises questions about the choice of stringent cutoff values for effect sizes. For instance, the authors note: "any smaller effect than r =.4 will not be interpreted as relevant even if significant." This seems very arbitrary. Why would there be such an important difference between an effect of r=0.40 versus r =0.39. It's important to consider that researchers are increasingly avoiding strong cut-off points and instead reporting all values clearly (every piece of evidence is relevant), allowing readers to assess the findings in a nuanced manner.

We actually agree with the reviewer on the problems related to cut-offs. Our reason to give an effect-size threshold for interpretation above the alpha is to minimize the emphasis put on the p-values (which are also based on arbitrary cut-offs) while still giving a clear method to avoid over-interpretating the results. We will certainly pay attention to not interpret the results as binary; we already avoided binary interpretations in our previously published papers (Najberg RR RSOS & SciRep). Furthermore, the effect sizes will still be given as they are in the Stage 2 Results section, which will enable the reader to interpret the results with nuance.

As we are now justifying the sampling plan with resources constraints/ power sensitivity tests (as suggested by the recommender), these cutoffs are no longer used as such in the sampling plan section.

Authors note that "For H2 and H3, which only consider the experimental group, the smallest effect size of interest was estimated to be small (r = 0.4) based on clinical subjectivity". I'm not sure what they mean with clinical subjectivity.

We have reformulated this section. It now reads p4: "We consider that the coefficient should be of at least $r \geq 0.4$ to consider the association between the decrease in explicit liking and dieting behavior (H2) or between the length of the intervention and its effect (H3) as non-negligible".

Additionally, it's worth noting that "r = 0.4" is generally considered a moderate to large effect size, rather than a small one, which should be accurately reflected.

The reviewer may be referring to the R-squared. A r of 0.4 is usually considered as a weak to moderate correlation (e.g., https://doi.org/10.1213/ANE.0000000000002864).

3. Demand Compliance Measure:

I also still wonder why the authors did not include a demand compliance measure. The authors note in their response to the suggestion to include this measure: “Concerning the demand compliance question, the experimental group should not have a larger response bias than the control group. Contrasting experimental vs. control should thus isolate any effect of this potential bias.” I’m not sure if the authors have data to support this claim. It seems there is contrasting evidence that, in the experimental group, participants are often much more likely to figure out the purpose of the training and are more likely to become demand aware and give demand compliant answers (and sometimes reactant answers as well).

We have added two questions in the debriefing questionnaire to address this comment. We ask to the participants: “Do you think the researchers of this study expect that your maintenance of the diet has been improved because of the training?” and “Do you think your maintenance of the diet has been improved because of the training?”. These questions will allow us to identify potential differences in demand compliance between the two groups.

Of note, when we measured the participants’ expectations in our previous RR on the topic (i.e., figuring out the purpose of the training) using the same experimental and control training groups, we observed phis ranging from 0.21 and 0.26 depending on the measures (weak effect sizes; see supplementary material of doi.org/10.1098/rsos.191288 & doi.org/10.1038/s41598-023-36859-x). This provides good reason to think that using our procedure, both training groups reach corresponding levels of expectations.

Review by Matthias Aulbach, 18 Sep 2023 06:43

The authors have replied to all my comments in a satisfactory manner. There remains, however, one point to be clarified, as I keep thinking about the personalized item set: what happens if a participant reports drinking less than eight of the drinks at a value above 0? Will a random selection of zero-value items be included in the training? I might be worrying about this too much as I don't have experience studying sugary drinks consumption, so maybe the authors could clarify this point.

In case of a tie in the consumption self-report, the tied trained items would be chosen randomly. In the edge-case of a participant reporting not drinking any sugary drinks, they would be excluded from the study. This is now clarified p 5: "Ties during the personalization process will be broken by choosing at random" and p 9: "Only participants who completed at least 7 sessions of training and reported non-zero values on the trained items consumption analogue scales will be considered".

Relatedly, will the study be advertised as relating to the reduction of sugary drinks consumption? That, of course, would lead to a selective sample in which consumption is probably rather common.

We indeed plan to advertise the study as a training aiming to reduce sugary drinks consumption.

Apart from this, I have no further points and wish the authors best of luck with their study.

We thank the reviewer for their kind wishes.