Response to reviewers, PCI Registered Reports #493
We thank the recommender and reviewers for their comments. 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 your Stage 1 Registered Report “Sugary drinks devaluation with executive control training helps to resist their consumption” to PCI Registered Reports.

I have now received comments from three expert reviewers in this field. As you will see, while we all agree that this RR addresses worthwhile and relevant questions, several aspects of the manuscript can be further improved. All three reviewers have provided valuable feedback on how you may do that. Based on the reviews and my own reading, I would therefore like to invite you to submit a revised version of the manuscript.

We thank you for your and the reviewers thorough and constructive comments. We have now addressed each of them and modified the manuscript accordingly.

1. For the introduction, I agree with Reviewer 1 (Dr. Van Dessel) that it is important to not make too strong and too general (and thus sometimes inaccurate) claims. Similarly, I think 'executive control training' is not the best term here, as neither the go/no-go training nor the cue-approach training trains 'executive functions' or 'cognitive control' per se. It is also unclear to me whether the 'executive control training' in the second paragraph refers specifically to the go/no-go training, or a combination of both training tasks. If the former, then the claim that "instruction to withhold responses to cues may reduce their hedonic value by developing attentional biases away from them" does not seem entirely accurate (and one of the two references for this claim is actually about the cue-approach training). If it refers to both training tasks, then it is also a bit odd, as different underlying mechanisms have been proposed for these two tasks. For these reasons, I think it would be clearer if you would start directly with the two specific tasks, explain what they entail (for readers who are not familiar with them), the general findings and the proposed underlying mechanisms. Note that should you adopt this advice, the title of this RR will need to be adjusted accordingly.

In the light of this comment and of the reviewer, mentions of “executive control training” have been replaced by “food Go/NoGo and cue-approach training” or simply “food response training” as a general term. The introduction of the tasks has been re-structured and clarified according to the suggestion of the recommender and reviewers. Detailed responses and excerpts of the text that has been modified are reported in our replies to the reviewers.

2. One of the proposed interventions does not look like an attentional bias modification task to me, but more like the cue-approach training (see comment by Reviewer 2, Dr. Aulbach). Although the effects of the cue-approach training have indeed been explained via attentional mechanisms, it would be more in line with previous work and less confusing if you would refer to this task as the cue-approach training.

We initially shifted from CAT to ABM because during the assessment of our RR Najberg et al., 2023: *Sci Rep,* a reviewer argued that our task included external reinforcements contrary to the traditional CAT, and thus that we should change the name of our task. Yet, we now came back to our initial choice, and followed the reviewer suggestion to use the term CAT.

3. One critique of previous work is that they employed self-reports that might be susceptible to "memory and social confounds". This is very true. However, I do not think that the main dependent variable here is completely free from these confounds. After all, it is also a self-report, and will likely also suffer from memory biases and social desirability issues.

Our view is that the self-report of the first transgression in the diet of easily recognizable items would be less biased by memory than those of common consumption level of food self-reports. Breaking a diet is indeed a clearer, more easily identifiable one-time event than just modifying a level of consumption over the course of days. Yet, we agree that our claim should be tempered down, and now wrote p3: “it is easier to report and less biased by memory or the relationship with the experimenter”.

4. Sampling plan: All reviewers had concerns about the sample size. First of all, What are the "principled grounds" to determine the potential effect size of interest? Obtaining an effect size as large as a Cohen's d of 0.7 (7 divided by 10) for H1 seems highly unlikely. A recent meta-analysis on the effects of food-specific go/no-go training on explicit food liking revealed an effect size of Hedge's g of 0.285, much smaller than 0.7 (Yang, Y., Qi, L., Morys, F., Wu, Q., & Chen, H. (2022). Food-specific inhibition training for food devaluation: a meta-analysis. Nutrients, 14(7), 1363).

Our effect size and sample size of interest were determined based on discussions with board certified dieticians on the minimal number of dieting susceptible to be of daily life or clinical significance, as well as on internal reflections on how the effect of our intervention observed on items’ valuation would translate into our present primary outcome of interest (the change in survival to restriction). We also had to take into account our experimental resources (time, human resources, budget for participants’ fees) to determine the maximal number of participants we could record for this study.

That said, due to the concern of the reviewer and recommender, our estimation of the smallest clinically relevant effect size for H1 has been reduced to result in a larger sample size. It now reads p4: “For H1, the estimated smallest effect size of interest that would be relevant to an applied setting is a difference in means of 5 days more of restrictive dieting in the experimental than control training groups, with an estimated standard-deviation of 10 (Cohen’s d = 0.5). A-priori power analysis using G*Power shows that a sample size of 140 (70 per group) is needed to reach 90% power with an alpha of .05 for a one-sided independent t-test and this effect size. Any smaller effect will not be interpreted as relevant even if significant.”

Second, for H2 and 3, please specify whether they will be tested in the experimental group only (which makes more sense to me), or with both groups combined.

For H2 and H3, only the experimental group will be investigated. This is now clarified in the revised manuscript and reads p4: “For H2 and H3, which only consider the experimental group [...]”.

Similarly, please provide more justification for why r = 0.4 is a reasonable effect size of interest. Note that the "Rationale for deciding the sensitivity of the test" column in Table 1 does not provide sufficient justification. Rather, it merely re-states that the effects need to be this big otherwise it is not relevant (but why?).

This column has now been deleted as it does not seem to fit our method of estimation. Please see our reply to the point above for the factors having contributed to determining our smallest effect size of interest and the sample size.

Third, given the many exclusions planned, please make clear whether the excluded participants will be replaced until the planned sample size is reached. If not, then the planned sample size will need to be larger still to leave room for potential exclusion.

With regards to the replacement of excluded participants, dropouts and participants with missing data will not be accounted for in their corresponding sampling plan, and thus replaced. Distribution outliers will not be replaced, because 1) it would create a circularity where the thresholds for exclusion would change each time we replace participants, eventually creating a new threshold that could have included past outliers, and 2) this procedure improves the reliability of the mean and its variance, thus increasing the quality of the results and our power. For an exclusion to comply with the positive control, participants will be replaced if this results in a sample size smaller than the planned sample size. This is now clarified in the analysis plan section p9: “Dropouts and participants with missing data will not be accounted for in their respective analyses.” and concerning the positive controls: “Participants excluded this way will be replaced only if their exclusions result in a sample size below the planned threshold.”

5. As Reviewer 3 (Dr. Tzavella) suggested, please share all experimental materials (e.g., the custom-made health questionnaire for assessing eligibility, the exact questions in the weekly questionnaires and the final debriefing questionnaire; these can be shared in e.g. Supplemental Materials). This will allow the reviewers to better assess the experimental procedure, and also facilitate the (re-)use of these materials in future work.

We now have uploaded the above-mentioned questionnaire translated from French in our OSF project, under the "PROTOCOL" folder (https://osf.io/s4trh/?view_only=4934c0215f2943cfb42e019792a30b53) and have referenced it in the manuscript.

Another major issue is that the two intervention tasks need to be described in much more detail, so that readers will understand the interventions without having to go to a previous publication.

The tasks' specifics are now reported in the Methodology section.

6. There is some ambiguity in exactly how "the end of the training phase" is defined, and accordingly the main dependent variable "days of successful restraint". I imagine that you will plan two weeks as the training phase, and administer weekly questionnaires after those two weeks. If a participant finished 7 days of training and then stopped, is the end of training for this particular participant the day they finished the last training, or the end of the two-week training phase, the same as with all other participants?

Indeed, the training phase ends whenever the participants stop the training sessions, with an enforced minimum and maximum number of sessions. The number of days of successful restraint is counted from the moment the participants stop their training.

After the minimal 7th day, an "End training" button appears on the menu screen of the software. If pressed, participants are instructed that they will not be able to come back to the intervention. If they confirm they want to stop the training, the post-training questionnaires and dieting phase start. If they never press the button, the software will present the post-training questionnaires and dieting phase at the maximum number of days. It now reads, p5: "Participants have the option to stop the study at any time through an "End training" button appearing in the software after the minimum 7 days of training, which in turn blocks the game and triggers the post-training measures."

7. The analysis plan section is a bit difficult to follow. I think it will be clearer if for each hypothesis, you would start with the raw data, explain the data exclusion and aggregation methods step by step, and then specify the eventual confirmatory analysis plan. At the moment, the information is scattered around in different sections and not always in the order of how you would process the data. Some more specific comments about data analysis:

We agree and reorganized the analysis plan section by hypotheses from raw data processing to confirmatory analyses. The positive controls section has been kept in a separate subsection at the end of the analysis plan to avoid repetition.

7.1 I think the independent t test function in base R uses Welch's t test by default, which handles unequal variances between groups better than Student's t test (Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. International Review of Social Psychology, 30(1), 92-101).

The default t.test() function indeed returns a Welch t test, which is the test we planned to use.

7.2 For H1, it should be "one-sided independent t-test" rather than "one-sided dependent t-test"?

This is now corrected.

7.3 "The mean explicit liking of each participant will be trimmed of their 20% highest and 20% lowest rated items at pre-intervention". If I understood it correctly, only the 8 most drunk items will be selected for each participant. In that case, what does trimming the 20% highest and 20% lowest rated items mean (20% * 8 = 1.6 items)? Also, how would this influence H2? The consumption behaviour is still about the 8 selected items I assume, but the explicit liking is only about a subset of these items.

We thank the recommender for pointing out this inaccuracy. This rule came from our previous studies where we aggregated the scores of more items. This processing step should not be applied in the present study because it focuses only on 8 items. We have now removed it.

7.4 Some of the positive control criteria seem better suited for exploratory analysis. For instance, it would be informative to examine dieting compensatory strategy, but why condition the analysis of H1 on this variable? There seems to be much flexibility in exactly which participants will be excluded, so it is not really a 'confirmatory' analysis anymore.

Our rationale was that differences in compensatory strategies between the experimental and control groups would bias the interpretation of H1, because it would be easier for the group with more frequent compensatory strategy to maintain the restrictive diet. We however agree with this comment and have thus removed this positive control. It now reads in the General procedure section p5: "A debriefing questionnaire will assess whether they consumed other types of sugary drinks as a compensatory strategy for exploratory purposes".

The idea behind data exclusion following this positive control was to select participants at random using a sample() function in our R script, with the sharing of the seed for reproducibility.

7.5 Related, to control "Baseline reported consumption", wouldn't including it as a covariable be a better approach than removing participants from data analysis?

We are reluctant to use baseline variables as covariates because it is generally not recommended to include in a model variables orthogonal to the hypothesis. If both groups initially differ in how they consume sugary drinks, then the bias in our random (but limited) sampling should be corrected from the raw dataset. By correcting this imbalance in a reproducible way, we ensure that the confirmatory model stays simple and free of any coincidental biases, and in turn can be interpreted unequivocally. Introducing covariables in the confirmatory model might complicate the interpretation of the model while keeping biases into the dataset.

7.6 For "Pre-post explicit liking reduction", I agree with Reviewer 2 that you can run the correlation analysis even if the overall effect is not statistically significant. Again, I feel removing participants to reach a certain criterion does not sound like a good approach here. Removing participants will further limit the range of this variable, which may make it even more difficult to detect the correlation, should it exist.

Our rationale for this positive control was that if no pre-post effects were found, then the training did not impact the participants the way we expected. From our own previous RR (doi.org/10.1098/rsos.191288 & doi.org/10.1038/s41598-023-36859-x) and the extensive literature on this outcome, we are confident that we will find at least a medium pre-post training effect in the explicit liking scores. On this basis, and according to the advice of the recommender and reviewers, we have removed this positive control for H2.

7.7 For both H2 and H3, I think it's a good idea to make e.g. scatter plots and visually inspect the underlying distributions. However, the analysis plan for H3 seems rather flexible to me because it involves visual inspection of a distribution (and thus much room for subjective judgement). If there is previous data showing that the distribution is likely to be uniform, I think that's a good working assumption for the confirmatory analysis. If that's not the case, you may adopt alternative statistical methods but those can then be clearly labeled as exploratory.

We agree with the proposed solution for H3's subjective visual inspection. It now reads p10: "Based on previous data showing a uniform distribution of the number of training days across participants, we expect a one-sided correlation between the number of successful days of diet and the number of days of training to be applicable as our confirmatory test".

The reviewers made other excellent points that I will not reiterate here, but it is important to carefully consider and respond to each of their comments. I wish you good luck with revising the manuscript, and I look forward to seeing the revision.

Thank you for your encouraging comments.

Kind regards,

Zhang Chen

by ***Zhang Chen****, 21 Aug 2023 13:49*
Manuscript: **https://osf.io/qe6j7?view_only=4934c0215f2943cfb42e019792a30b53**
version: 1

Review by Loukia Tzavella, 20 Aug 2023 15:55

Review for Stage 1 RR:

“Sugary drinks devaluation with executive control training helps to resist to their consumption”

Sampling plan

If 50 participants are needed in the experimental group for H2 and H3 why is the target sample size based on H1 which requires a smaller number of participants in each group (N= 36)?

This was a mistake on our part. The sample size has been corrected.

Also, I would be more inclined to use a d of 0.4 for the power analysis and not the chosen mean/sd differences that result in a Cohen’s d of 0.7, as in the rest of the manuscript you mention a d of 0.4 as your benchmark (smallest effect size of interest). In any case please make this clearer as reading this the first time I thought that the 7-day difference resulted in a d of 0.4. If you apply this change, your total sample size would be 216. A greater sample size may also allow for more power for H2 and H3 analyses (under 50 participants in the experimental group - which will further be reduced if you apply data exclusions as per my comment for the Pre-post explicit liking reduction section).

The resulting Cohen’s d is now reported alongside the estimated population parameters for H1. The rationale behind using population parameters instead of a relative effect size is also clarified. Furthermore, our estimation of the smallest clinically relevant effect size for H1 has been reduced to result in a larger sample size. This now reads p4: “Our rationale for the sampling plan is to try and detect at minimum the smallest effect that would be relevant to a daily living or clinical setting, instead of searching for the minimal effect expected with the current literature. As such, when possible (i.e., H1), the population parameters (e.g., differences in means) are used as effect sizes instead of relative indexes (e.g., Cohen’s d).

For H1, the estimated smallest effect size of interest that would be relevant to an applied setting is a difference in means of 5 days more of restrictive dieting in the experimental than control training groups, with an estimated standard-deviation of 10 (Cohen’s d = 0.5). A-priori power analysis using G*Power[31] shows that a sample size of 140 (70 per group) is needed to reach 90% power with an alpha of .05 for a one-sided independent t-test and this effect size. Any smaller effect will not be interpreted as relevant even if significant.”

Recruitment and screening

As an inclusion criterion, willingness to follow a restrictive diet is important but it is also worth recording participants’ baseline consumption behaviour. Are you including everyone from people who rarely drink sugary drinks to people who drink more than a few sodas a day? In this section it is important to also add any methodological details or at least point to where the reader can access them (what are the exact questions for your screening- e.g. how do you define ‘healthy’ individuals in this context).

We did not plan screen participants for baseline consumption because we indeed wanted to include everyone. Another reason for this choice is that participants may estimate their consumption based on unpredictable criteria; for example, they might not consider orange juice or iced coffee as sugary drinks whereas we (and other participants) would consider them both as sugary drinks.

The questionnaires and its criteria can now be read via our OSF page, under the PROTOCOL folder (https://osf.io/s4trh/?view_only=4934c0215f2943cfb42e019792a30b53).

Training tasks

I understand that the tasks have been used in previous studies but for this Registered Report I think you should not omit the methodological details and specific parameters of the tasks being administered (contingencies, time limits, number of trials, feedback element etc.). They are central to the study and should be presented as part of the Stage 1 proposal for further evaluation - even if the app cannot be changed at this time. The video for the app demonstration was really helpful - you could add a figure with screenshots from the game in the main text for convenience as well.

We have now included detailed tasks parameters alongside a summary of the training gamified features in the Method section, p6 to p8.

Questionnaires

Please add a reference to supplementary material or an online repository where the full questionnaires can be found. In the text you can add more details about the 10 items being included in the health questionnaire - what are you measuring with regards to participants' health?

We now provide a link to the screening questionnaire in the OSF page in the Questionnaires section, p9: "All questionnaires translated from French can be read via our OSF page under the "PROTOCOL" folder: https://osf.io/s4trh/?view_only=4934c0215f2943cfb42e019792a30b53."

In the recruitment and screening section, we have specified the definition of healthy participants. This now reads p5: "Unhealthy participants include self-report of past or current eating disorders, any visual or hearing disability preventing gamified training, and any olfactory or gustative impairment (including smokers consuming ≥10 cigarettes daily)."

Analysis plan

In this section you need to specify all the analyses that will be run and treated as 'confirmatory' with details so you can move the paragraphs from the Statistical contrasts section here. For Bayesian analyses, what priors will you be using for the t-test and correlations? Also, while BFs can be very informative in the case of inconclusive results it would be preferable in my opinion to report them for all results irrespective of significance. Also, although it may seem obvious please state on which statistics you will base your conclusions on (e.g. frequentist but BFs reported in a supplementary manner?).

The Analysis Plan section has been restructured following comments from the recommender and reviewers.

For the Bayes Factors, the t-test and correlations will follow the noninformative priors computed by default with the BayesFactor R package, namely for the t-test: "a noninformative Jeffreys prior is placed on the variance of the normal population, while a Cauchy prior is placed on the standardized effect size" and for the correlation: "noninformative priors are assumed for the population means and variances of the two population; a shifted, scaled beta(1/rscale,1/rscale) prior distribution is assumed for ρ".

We could report the BFs for all analyses, but at the cost of (over)complicating the interpretation of the results. As mentioned by the reviewer, we rather propose to interpret significant results using the frequentists values, and the BFs to support the null results in case of frequentist non-significance, so reporting BFs for all tests might be superfluous. Additionally, we would like to stress more on the effect sizes than on the inferential statistics for interpretation, and thus limiting as much as possible the quantity of inferential statistics that we report.

We have clarified the interpretation of the results. This now reads p9: "All results will be interpreted using the frequentist statistics, with Bayes Factors reported as a supplementary manner to support null results."

Data exclusions

Could you add details here about potential missing data and related exclusions with regards to your questionnaires? For example, what if participants don't complete the weekly questionnaires, or if information is missing (e.g. exact dates of first consumption etc.)?

Participants with missing data will be removed from their corresponding analyses. For instance, if a participant does not complete the weekly questionnaire, they will not be considered in any confirmatory hypothesis testing on this dependent variable. We have now clarified this point in the Analysis plan section p9: "Dropouts and participants with missing data will not be accounted for in their respective analyses."

Given that reaction times from the analogue scales are recorded I presume that you can also have access to training performance data. While technically a day of training can be counted as successful if completed, I believe you should mention whether potential exclusions can apply to adherence. For example, does training proceed if you miss the reaction time window (if there is one, not currently known based on the details presented) or if participants simply skip trials and not interact with the game?

Based on our two previous RR with the same gamified software (https://doi.org/10.1093/cercor/bhaa259, doi.org/10.1098/rsos.191288, doi.org/10.1038/s41598-023-36859-x), participants completing the minimum required sessions all have satisfactory performance data. In addition, as now included in the manuscript, the tasks control for most of 'bad' behavior (too slow RT, missed trials, etc) because only a limited number of speed and accuracy errors are allowed, and the participants have to complete enough trials to finish their daily training.

Training rules are now specified in the "Training tasks" sub-section. This now reads p6: "In both tasks, the participants must complete as many trials as they can in one block. Each correct response awards points to the participant. After five correct responses, the reaction time threshold (RTT) is increased of a level (Table 2). After making a certain number of accuracy or speed errors (5 without powerups), as indicated by two distinct life gauges, the run is over. This process is repeated until the participants reach 10 minutes of training for each task."

Since the proportion of successful inhibitions in such training tasks may be a moderator of training effects (see meta-analysis by Jones et al. 2016), it is worth considering a performance benchmark for data exclusions - e.g. if one day of training is completed but participants fail to stop for more than half of the trials. As this may be a conservative criterion for data exclusions given the sample size you propose, it would be interesting to add training performance as a secondary outcome or consider certain exploratory analyses at Stage 2 to look at learning effects and inhibition success.

To ensure a minimal amount of successful inhibition trials, the tasks were designed to have a slow increase in difficulty during a training session (see Table 2 in the revised manuscript), and a maximal number of inhibition errors is allowed for the run to continue to higher difficulty levels. According to our previous data using the same task parameters, the accuracy is high and consistent across participants (mean = 79%, sd = 7%; doi.org/10.1098/rsos.191288, doi.org/10.1038/s41598-023-36859-x).

We will however surely analyse these aspects and report them as exploratory analyses at stage 2.

Statistical contrasts

Please state why have you chosen this criterion for your effect interpretation - i.e. why is a min d of 0.4 required to consider the result 'relevant' - I have found this confusing as mentioned in the Sampling plan. You mention that the result will only be relevant if the difference is at least 7 days or more of successful dieting, but in your power analysis this corresponds to a Cohen's d of 0.7 and yet in the text a Cohen's d ≥ 0.4 is treated as relevant.

The minimal effect size for each hypothesis has been corrected. It now reads for H1 p9: "This result will be interpreted as relevant only if the difference between both conditions is at least of 5 more days of successful dieting, even with a p-value below 0.05".

Please refer to our reply to the recommender's point on this question for details on how the sample size / smallest effect size of interest were determined.

Baseline reported consumption

If I understood this correctly, will you inspect the data, run the analyses and if for H1 you get a Cohen's d greater than 0.4 you will exclude participants and report the results with the reduced sample size? For other data exclusions that do not require statistical analyses I assume that recruitment will continue until the sample size target is met, but for this exclusion criterion please add more details regarding the sampling plan - that is, how your sample size may be affected given that the target is based on an a priori power analysis.

The exclusion plan related to this positive controls has now been clarified. It now reads p9: "All positive controls are checked from the raw data before any processing (see "positive controls" section), including the potential exclusion of participants to respect them. Participants excluded this way will be replaced only if their exclusions result in a sample size below the planned threshold".

The relationship between exclusion related to positive controls and the target sample size has been clarified. It now reads in the Sampling plan section p4: "Overall, a total of 140 participants will be needed for the analysis of H1, and 50 participants will be needed for the analyses of H2 and H3. If exclusions to comply with the positive controls reduce the sample size below these thresholds (see Statistical Analysis section), new participants will be recruited".

Pre-post explicit liking reduction

Please add a justification for this in a narrative format - e.g. for H2 you only want to run the correlations if a devaluation effect is observed (defined by your chosen threshold) - this should be clear in your hypotheses as well. It may be good to present results without all the effect-related exclusions in the supplementary material as well for comparison.

The Analysis Plan section has been restructured for coherence. Following comments from the other reviewers, this positive control has been removed from the analysis plan.

Download the review

Review by Matthias Aulbach, 15 Aug 2023 08:20

Review for "Sugary drinks devaluation with executive control training helps to resist to their consumption"

In the Stage 1 report "Sugary drinks devaluation with executive control training helps to resist to their consumption", the authors lay out a randomized controlled trial that tests an app-based cognitive control training against an active, "sham-training" control group in its effectiveness to reduce evaluations and intake of sugar-sweetened beverages. Further, the authors aim to investigate the relation between devaluation and consumption effects as well as a dose-response effect between the amount of conducted training and consumption. The planned study is generally well-designed and tackles worthwhile questions in the field. I have some thoughts and suggestions that I hope would improve the study and its possible interpretations. Please note that this is the first time I am reviewing a Stage 1 report (which I'm quite excited about!), so please bear with me for any possible "breach of protocol".

I was somewhat unclear about one of the proposed intervention tasks. The description of the ABM task in the introduction did, in my view, not fit an attentional bias modification. Looking at the video of the task in the OSF folder, the ABM task looks more like a mix between Approach-Avoidance Task, GNG, and Cue approach task, which might actually work quite well, see van Alebeek, Veling, and Blechert (2023) (https://www.sciencedirect.com/science/article/pii/S0950329323000150). In past papers, the same authors have also termed this "Cue approach training". Why do the authors frame this as an attention modification paradigm here? I fail to see how it modifies attentional biases as it requires attention allocation to both approach and withhold items.

The task was indeed originally named CAT as we designed the task based on Shonberg et al., 2014. However, during the assessment of our RR Najberg et al., 2023: *Sci Rep,* a reviewer argued that our task included external reinforcements contrary to Shonberg's CAT, and thus that we should change the name of our task. Yet, based on the present reviewer's comment and our initial choice, we have decided to change the name back to CAT.

The authors include an active "sham-training" control condition (which I think is great) to control for confounding factors of expectations and cue exposure. This is surely a good idea but in addition, the authors could also include a measure of expectations (for example as we have done in an ongoing study of our own group: https://bmjopen.bmj.com/content/13/4/e070443.abstract). If expectations play a crucial role (and there is evidence that suggests that they might), we should then see an effect of expectations across both groups in the study. I see that this is a tangent to the current study's goals but might still be interesting to consider.

We agree that the impact of expectations on food valuation is an important topic, as demonstrated by the effort we devoted to this aspect in our two previous RRs on the topic (doi.org/10.1098/rsos.191288 & doi.org/10.1038/s41598-023-36859-x). In this study, however, the focus is more on the applied question of the translation of training-induced devaluation to the survival of a restrictive diet. However, according to the comment of the reviewer, we have now added two Likert scales on the participants' expectations on the impact of training on their diet, p9: "Expectation on the study's hypothesis will also be rated using two 5-items Likert scales at the same time, asking 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?" with 1 (Not at all) and 5 (Absolutely) as the anchors".

I must admit that I never fully grasped what exactly the dependent variable for H1 is. Is it the number of days until the first consumption of trained drinks? Or the total number of trained drinks consumed during the follow-up period? I was thinking that the best outcome measure would be the amount of drink consumed: arguably, this is easier to report than foods because of fixed common serving sizes. Such a consumption measure would be much more sensitive to change than cruder measures.

The first proposition of the reviewer is correct: the outcome for H1 is how many days participants report not drinking the trained sugary drinks after training. Our last published RR (doi.org/10.1038/s41598-023-36859-x) investigated the amount of sugary drinks consumed pre- and post-training, but here we opted for the more radical approach of asking participant to completely avoid the target items and see how much time they can maintain the restriction. Our reasoning was that if the devaluation indeed influenced actual consumption, it would manifest as an easier capacity to avoid the devalued drink consumption. We think that while less sensitive than consumption measures, our current primary outcome has the critical advantage of being less influenced by biases inherent to consumption self-reporting (social desirability, memory, etc) than consumption level since breaking the diet is a one-time event. It also fits better with the more applied perspective of the present study and the real-life situation of participants having to follow strict restrictive diets (not only to reduce their caloric intake, but also for allergic or intolerance purposes).

Regarding H2: will this relation be calculated across groups? One might argue that we would only expect this effect in the intervention group because changes in the control group should be rather random. If you only use data from the intervention group, the sample size calculation does not add up because you would need 50 participants in the intervention group alone.

H2 only focuses on the experimental group. The planned sample size was a mistake, now corrected with the new estimation for H1.

Regarding H3: to really speak of a dose-response effect, one would need to assign dosage experimentally, as done in Moore, White, Finalyson, and King (2022) (https://www.sciencedirect.com/science/article/pii/S0195666322002720?casa_token=N-C479gEscQAAAAA:wAz8DeDD_lzReFYBdfACTmcNnfIXdlc_7q_x5PunfGXmcB-j7slBgnh0QfF40afP0RTKdUuCgQ). With the current design, we would probably mainly see an effect of motivation on both training use and consumption. We discussed this in some detail in our paper after having been told off by a reviewer for using the term "dose-response relationship" (see Aulbach et al., 2021; https://www.sciencedirect.com/science/article/pii/S0195666321002221). Maybe this could be taken into account by adding a measure of motivation and controlling for it in the analysis but it still does not solve the problem of self-selection. Relatedly: will the analysis for H3 be conducted across groups? If yes, adding group as a moderator in a regression could show whether this "descriptive dose-response effect" differs between groups. That (especially when combined with expectancies) would be very interesting: if more training produces larger effects but only in the intervention group while expectations are similar between groups that would be quite strong evidence for the intervention.

We thank the reviewer for their insight and agree with the comment. We have now removed all mentions of dose-response.

To maintain the focus of the study on applied outcomes, we are reluctant to add a contrast of H3 across groups to isolate the effect of expectations. Complexifying the model would also require a sample size above our available resources for this study. However, to address this comment, we have added a positive control on the balancing of participants' expectation, p11: "For H1, the expectation on the impact of training on the maintenance of the diet should be balanced between groups to interpret the results without this bias. In case of a Cohen's d above 0.4 on the average score between the two Likert scales (see Questionnaire section) between the experimental and control groups, participants further away from their group's average will be excluded until this criterion is met.".

Another aspect that limits this analysis is that the distribution is truncated at 7 and 14: there will be no participant with less than 7 or more than 14 sessions which severely limits interpretability of any found (or non-significant) effect.

The reviewer is correct on the limitation of the results interpretation induced by our procedure, and we have thus increased the maximum range up to 20 days of training. We will be very cautious on not over-interpreting null results in the light of potential ceiling effects that may appear before the 7th day of training. However, we will keep this minimum training length as it is important to enforce a minimal training length to induce detectable real-world effects.

Specific methods points:

- What is the rationale for the age range of included participants?

We wanted to target adults, and to avoid grouping different age groups (elderly or children) as they might have different neurocognitive responses to the training which can bias the results (e.g. visual impairment or fatigue, digital literacy, experience with videogames, dieting, etc).

- 20 minutes of intervention per day seems like a lot. What kind of dropout rates do you expect, especially given that you want to exclude anyone with less than seven sessions?

From our last RR in which we required a similar amount in a similar online procedure, we expect roughly 25% of participants to complete the study, which is compensated by a large-scale recruitment and automatized online screening.

- Will excluded participants be replaced or will the analysis be conducted with what you have after exclusions? If it is the latter: how will that affect statistical power at different exclusion rates?

Participants excluded to comply with the positive controls will be replaced if their exclusion results in the reduction of the sample size below the planned thresholds. It now reads in the Sampling plan section p4: "If exclusions to comply with the positive controls reduce the sample size below these thresholds (see Statistical Analysis section), new participants will be recruited." and in the analysis plan section p9: "Participants excluded this way will be replaced only if their exclusions result in a sample size below the planned threshold".

With regards to the replacement of dropouts and participants with missing data, they will not be accounted for in their corresponding analyses and thus replaced. Finally, distribution outliers will not be replaced, because 1) it would create a circularity where the thresholds for exclusion would change each time we replace participants, eventually creating a new threshold that could have included past outliers, and 2) this procedure improves the reliability of the mean and its variance, thus increasing the quality of the results and our power.

- Participants receive 8 sugary drinks for reduction. That seems like a lot of different drinks! I would consider lowering this number, as otherwise drinks will need to be included that are not drunk very often in the first place.

As our sugary drink pictures dataset (cf our OSF project page) includes 9 different types of sugary drinks across 57 pictures, with some brands displayed multiple times (e.g., the Emmy iced coffees, or Redbulls "organics" set), we believe that a sample of 8 drinks is small enough to avoid 'diluting' the training effect, but large enough to allow for the whole range of sugary drinks to be represented. It is true that for one category of sugary drinks, participants may be used to drink one or two of the same bottles, but when we consider the whole range of what sugary drinks entails (iced coffees, iced teas, juices, sport drinks, etc), we believe that 8 drinks may represent the real variety of the consumption, in turn reducing the probability for compensatory strategy to take place during the dieting phase.

- Relatedly, the consumption frequency question has a very subjective scale. Why not ask e.g. "On how many days during the last month have you drank this?" Otherwise, participants will scale the responses to their own standards: For person A, "often" might mean once a week, for person B it might mean several portions a day.

The subjectivity of the scale should not impact the study, as we are not comparing participants with each other in any confirmatory analysis on this scale. For the related positive control on this variable, it is unlikely that both groups will have a large enough average difference of this scale's perception to substantially impact the results.

- I'm not sure if the liking measure is ideal to measure what is of key interest. With the current wording, it seems very context-dependent: I might like cola on a hot afternoon but imagining drinking it in the morning, I would not like it. From the introduction I read that the main interest is relatively stable preferences. If you measure a very variable momentary preference, you will get a lot of noise in your measure and might not measure so much what you care about.

From our two last RR using this scale, we observed a robust pre-post-training effect in the experimental group (Cohen's d = 1.7 and 2.1 for Najberg et al., 2021 and 2023 respectively), showing small intra-subject variance. This shows that factors outside of the main effect of intervention, such as the time of day, had little impact on this variable. Additionally, using the same question as in our previous RR would facilitate comparison with this literature and contribute to a replication effort.

In the analysis plan section, there is mention of “deltas” – I assume those are pre-post changes? Please clarify.

This is now clarified in the manuscript p10: “pre-post-training differences are computed”.

Regarding the compensatory strategies: I would find it interesting to see if the intervention group engaged more in those because it would suggest that the training works for the specific trained foods but does not generalize to similar items.

This data is certainly interesting, and we will surely explore them at stage 2.

The authors write that a pre-post reduction of explicit liking is necessary for investigating H2. I do not think this is necessary: you would just correlate one distribution with another (liking – consumption) and it should be irrelevant where on the scale this correlation occurs/where these distributions are on the scale.

Indeed, following the comments from the recommender and reviewers, this positive control has now been removed from the analysis plan.

In summary, I think that this study will produce very interesting evidence on the effects of mobile cognitive bias modification. I hope that the authors find my comments and suggestions helpful to further improve the study. I’m aware I referred to some of my own and my colleagues’ papers which is somewhat frowned upon in peer review. By no means do I insist they be cited in any publications, I only included those because I know those best and found them helpful pointers for the issues at hand.

Thank you for giving me the opportunity for this stage 1 review and possibly contributing to making this study even better.

We thank the reviewer for the positive assessment and the very constructive comments of our work.

Review by Pieter Van Dessel, 14 Jul 2023 08:42

In this registered report, the authors aim to investigate the impact of a combination of Go/No Go training and approach bias training on participants' consumption of their favorite sugary drinks. The study addresses an important and relevant topic, as the exploration of online training methods to improve unhealthy consumption patterns holds practical and theoretical significance. The authors are commendable for their commitment to good scientific practices by conducting this study as a registered report.

We thank the reviewer for their thorough revision and the constructive comments.

However, while reading the paper, I identified several areas where the authors could enhance the quality of their work. Primarily, the introduction falls short of the expected standards. It contains inaccuracies and lacks clarity regarding the constructs under investigation. It is crucial for the authors to be more precise when describing the behavior of interest, clearly distinguishing between observed behavior and the explanation of behavior in terms of mental processes. For instance, in the first sentence of the abstract there are already several inaccuracies: "Food executive control training has been shown to reduce the perceived value of palatable food items". The authors mention a reduction in the perceived value of palatable food items, but since the participants' perception is not directly probed, it would be more accurate to refer to it as self-reported value.

This point has been corrected.

Furthermore, instead of using the term "executive control training," which implies training of the mental construct 'cognitive control,' a construct that cannot be directly observed, it would be better to consistently refer to the specific tasks employed, such as the go/no go training task and the attention bias modification task (and please make a note there as well that this task does not directly modify attentional bias, it merely targets this bias). I recommend that the authors critically review all the constructs discussed in the paper, ensuring accurate definitions and clear differentiation between behavioral effects and mental constructs.

To address this comment, "executive control training" has been replaced by "food Go/NoGo and cue-approach training" or simply "food response training" as a general term (also in the title). Similar corrections have been made to improve the quality of the introduction.

Similarly, it is essential to avoid making inaccurate claims. For example in the second paragraph of the introduction, the authors state that conventional reflective approaches to reduce overconsumption behaviors usually fail because they target conscious processes, while (palatable) food consumption is largely driven by environmental cues. However, this represents only one explanation of findings and it should not be presented as if it is a truism.

We agree and have now removed this sentence.

Additionally, in the next sentence, the authors claim that recent evidence indicates automatic motivational processes driving unhealthy overconsumption can be modulated by executive control training (ECT), with ECT robustly reducing the perceived value of targeted cues in the eating domain. These claims are not accurate and the sentences would better be discussed in reversed order. First, highlight evidence that specific types of training can reduce self-reported value. Then indicate that this reduction has been explained as potentially targeting automatic motivational processes driving unhealthy consumption (while noting other explanations as well). It is crucial to avoid oversimplifying the literature. In this sense, it is also worth noting that training effects are often limited, and there is little evidence supporting real-life effects of cognitive bias modification, except perhaps in the context of alcohol approach bias modification for alcoholic patients. It is essential to address this omission and discuss relevant work in the field.

We agree and have adjusted the first five paragraphs of the introduction accordingly. Yet, we would note that the focus of the present study is not on the devaluation per se, notably because i) we have demonstrated in two previous RR that the devaluation effect is robust with our intervention and why we think it is the case (doi.org/10.1098/rsos.191288 & doi.org/10.1038/s41598-023-36859-x), and ii) several papers already discuss the mechanisms underlying response training (e.g. Veling et al,. 2017, doi.org/10.1007/s40429-017-0131-5) and the robustness of the related literature (e.g. Carbine & Larson, 2019, doi.org/10.1080/17437199.2019.1622144)

In the present paper, we wanted to focus on the translation of this effect to an applied outcome, namely the capacity to resist the consumption of the devalued items in the context of a restrictive diet. Hence, we prefered not to reorient and develop the introduction too strongly on the devaluation aspects (and mechanisms thereof). That said, according to the suggestion of the reviewer, we have now modified this section to first introduce the tasks individually, then their putative underlying mechanism and limitation, and to then approach the conscious/automatic processes argument.

Moving on to more minor issues within the paper:

- The explanation of the sample size rationale lacks clarity, as the authors consistently fail to state the effect of interest for the t-test as Cohen's d. Additionally, it is unclear why only an effect of 7 days would be of interest, considering that finding such an effect appears highly unlikely. Moreover, with 36 participants per group, the sample size seems relatively small, which raises concerns about the informativeness of the results. Further explanation is needed to address these concerns adequately.

For sample size calculation, our goal is to aim for an absolute effect size of interest that can be considered relevant in an applied daily life or clinical setting (i.e., “how much more days should a patient maintain his diet for the solution to be worth considering?”), instead of looking for any significant effect that can be found. This is why the effect size of interest is presented as its estimated population parameters (mean and sd) instead of the resulting Cohen’s d whose interpretation is based on community-driven benchmarks. To facilitate the interpretation of this aspect, the resulting Cohen’s d for H1 is now reported in the manuscript, and the rationale for presenting population parameters is clarified. It now reads p4: “Our rationale for the sampling plan is to try and detect at minimum the smallest effect that would be relevant to a daily living or clinical setting, instead of searching for the minimal effect expected with the current literature. As such, when possible, the population parameters (e.g., differences in means) are used as effect sizes instead of absolute effect sizes (e.g., Cohen’s d).

For H1, the estimated smallest effect size of interest that would be relevant to an applied setting is a difference in means of 5 days more of restrictive dieting in the experimental than control training groups, with an estimated standard-deviation of 10 (Cohen’s d = 0.5).”

- It would be beneficial to report Bayesian factors for significant results in addition to other statistical measures.

We prefer not to systematically report the BFs for all analyses (at least not in the main manuscript), to avoid complicating the interpretations without adding much value. Significant results are interpreted using the frequentists values, with the BFs to support the null results in case of non-significance. Hence, reporting BFs for all tests might be superfluous, especially since we would like to stress more on the effect sizes than on the inferential statistics for interpretation.

- The paper states that adherence to a restrictive diet constitutes a robust and valuable dependent variable to assess the real-world effect of food ECT, as it is not biased by memory or the relationship with the experimenter. It would be helpful to explain the basis for this claim. Why would there not be a bias by “memory”. Isn’t there always a bias by a broad cognitive construct like memory? Additionally, it is important to note that experimenter demand still presents a possible explanation of any effect that may be observed. Including a demand compliance question at the end could provide more information about the possibility of participants in the experimental group conforming to the hypothesis and falsely indicating that they did not consume the sugary drink.

We think that any memory bias coming from a behavior as specific as avoiding a target food item would be negligible. Individuals in a targeted restrictive diet would unlikely forget they have broken the diet by eating a forbidden item, especially if they have to detect only the first failure. Hence, we think that relative to food journals or food frequency questionnaires which focus on subtle modulation of daily eating behavior, biases on dieting behavior are negligible, even if existent. To reflect this, we have modified the manuscript to now read p3: “it is easier to report and less biased by memory”.

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.

- More background information is required to understand the rationale behind combining Go/No Go (GNG) and Attentional Bias Modification (ABM) training. The authors should provide a clearer explanation and reference their previous studies that employed this combination. It is crucial to highlight the added value of this research compared to their prior work.

The combination of both tasks is now discussed in the introduction. It reads now, p2: "Our previous work has demonstrated that the combination of these task in a response training intervention robustly reduces the self-reported explicit liking of the targeted unhealthy food cues, alongside a potential increase in the healthy items valuation and a decrease in the unhealthy items self-reported consumption[10,11]".

- The statement, "This contrast will allow us to control for the confounding factors of cue exposure and of expectations developed by the participants on the effects of the intervention," raises questions about expectations as a confound. The authors should clarify why expectations would not drive the effects as in fact, I think expectation may always drive any behavior effect.

Additionally, it is evident that the control group does not control for expectations. 100% and 50% contingencies are typically noticed very well by participants (and in fact this awareness seems crucial for effects) and lead to different inferences (including causal inferences or predictions).

Our last two RRs measured the participants' expectation on our hypotheses (reduced self-reported valuation and consumption of the trained items), both using the same experimental and control training groups. We observed phis ranging from 0.21 and 0.26 depending on the measures (see supplementary material of doi.org/10.1098/rsos.191288 & doi.org/10.1038/s41598-023-36859-x), showing a small to medium impact of the training group on the expectation: The 100% / 50% contingencies were noticed, but not enough to consider it as a break-point.

We however agree that this point was not sufficiently examined and thus added a rating of participants' expectation on the effect of the intervention on increased dieting behavior and balance them across groups in as a positive control. This can be found in the manuscript p9: "Expectation on the study's hypothesis will also be rated using two 5-items Likert scales at the same time, asking 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?" with 1 (Not at all) and 5 (Absolutely) as the anchors." and p11: "For H1, the expectation on the impact of training on the maintenance of the diet should be balanced between groups to interpret the results without this bias. In case of a Cohen's d above 0.4 on the average score between the two Likert scales (see Questionnaire section) between the experimental and control groups, participants further away from their group's average will be excluded until this criterion is met".

- H3) should be rephrased. The current phrasing suggests that the more participants train, the larger the effect of the intervention will be on their dieting behavior. However, if an effect is found, it does not necessarily imply that the intervention caused it. An alternative explanation could be that participants who train more are more motivated to stop drinking sugary drinks and thus exhibit reduced consumption.

We agree and we thank the reviewer for raising this point. H3 now reads "H3) The amount of days of training in the experimental condition will correlate positively with the number of days of successful dieting" indicating no direct cause between training duration and dieting behavior.

- Including the compensatory strategy in an ANOVA to determine its role in explaining the observed effects would be beneficial and should be considered.

While compensatory strategies are clearly interesting, the interaction between them and dieting behavior would require a sample size above those allowed by our resources to be examined with sufficient power. Yet, we will certainly investigate this aspect with exploratory analyses and possibly report at stage 2.

By addressing these concerns and incorporating the suggested improvements, the authors can significantly enhance the quality and clarity of their paper, making it a valuable contribution to the research community.

We thank their positive evaluation of our work.