Dear Dr. Andrew Jones,

We would like to thank you for the opportunity to revise our manuscript, and the reviewers for their helpful feedback. Below, we list the reviewers' comments **in bold**, and explain how we have addressed each comment in the revision. Text from the revised manuscript is in green. We have also highlighted these changes in the revised manuscript.

On behalf of my co-authors,
Kind regards,
Zhang Chen

Before addressing the comments by the reviewers below, we would like to mention a minor change that we have made. That is, for the food rating task with the 200-point slider, we initially wrote that the cursor would always start at the 0 point, using the default setting of jsPsych. However, since participants need to move the cursor in order to advance to the next trial, in our previous experiments we have observed that participants tended to not use ratings around 0 (because they tended to move the cursor away from 0), causing the overall distribution of rating scores to be bimodal. To reduce the potential influence of this bimodal distribution on the item selection procedure, the program now does not show the cursor on the slider initially. Instead, participants can click anywhere on the slider, and only after this first click, the cursor appears. They can still further adjust the position of the cursor until they confirm their rating. We have revised the manuscript to reflect this (Page 14).

> Participants could click anywhere on the slider and a cursor would then appear on the clicked location. They could further adjust the position of the cursor until it accurately reflected their rating. They could then click on a 'Continue' button beneath the slider to advance to the next trial.

Comments by Reviewer 1 (Alexander MacLellan)

This paper aims to investigate whether Go/NoGo cues vs Approach/Avoidance cues influence behaviour when selecting appetitive food. The question is an interesting and valid question, and the authors have a well detailed, and considered study planned. There is a clear logic and rationale, with clear implications for cognitive training programmes for unhealthy eating. The manuscript in its current state is well written, providing sufficient detail to allow for replication, and meets the majority of the issues to consider at Stage 1 of a PCI RR. My comments are as follows:

Response: We thank the reviewer for their positive evaluation and helpful feedback.

Main comments:

- The sampling plan is very well described with an extended justification of their first sample size estimate, however the decision to recruit 100 participants per group when they state 58 in each is sufficient. Might it not be better to continue recruit until that sample size is reached (after exclusions etc.) rather than over-recruit?

Response: Thank you for the comment which made us realize that our explanations were unclear. Given the inherent uncertainty in effect size estimates, we had decided to take a conservative approach and use 80% of the estimated effect size in our power analysis. We have now revised the manuscript to make the reasoning clearer.

> Page 12: We therefore expected the effect size for both the effects of go/no-go and approach/avoidance actions on choices to be around Cohen's d of 0.533. Given the inherent uncertainty in effect size estimates, we used Cohen's d of 0.426 (i.e., 0.533 * 80%) as the expected effect size in a power analysis in G*Power (version 3.1.9.6; Faul et al., 2007), which showed that 60 participants were needed (with a two-sided one sample t test) for 90% power with an alpha level of .05.
>
> Page 13: Again, to be conservative, we used Cohen's d of 0.488 (i.e., 0.611 * 80%) in a power analysis, which showed that 90 participants per condition (180 participants in total) are needed (with a two-sided independent samples t test) for 90% power with an alpha level of .05.

To leave room for potential exclusions, we thus decided to recruit 200 participants in total. This number is also close to the maximum number of participants we can recruit given our resources for this project. Our previous experiences with the paradigm have shown that the number of participants that we will need to exclude will be relatively low. However, to avoid the situation where the final sample size in any group would be lower than 90, we have further adopted the reviewer's suggestion of continuing recruiting participants if necessary, until we have at least 90 participants in each group.

> Page 13: To be more conservative in our sample size planning, we therefore used 80% of the expected effect size and 90% power in the power analyses above. To allow for potential exclusions,we decided to recruit 100 participants per condition (N = 200 in total). In case the sample size in an instruction group is below 90 after exclusions (see below), we will continue recruiting participants for that specific instruction group, until the final sample sizes in both groups are 90 after exclusions.

- **The authors then state they will recruit 160 participants in the 'Participants' section (page 13), which is inconsistent with their sample size estimation.**

Response: Thank you very much for catching this and we apologize for this mistake. We have corrected it in the manuscript.

Minor comments:

- **As a minor comment, but in the interest of transparency, it would be a positive to see a statement added at this stage as to the level of bias control achieved in this study (e.g. Level 6), as well as statements on data and code availability.**

Response: We have now added a statement on the level of bias control achieved, and provided a link to the OSF repository (page 11).

> The current manuscript achieved level 6 of bias control according to the policies of Peer Community in Registered Reports: *No part of the data or evidence that will be used to answer the research question yet exists and no part will be generated until after IPA*. All experimental materials, raw data and analysis code are available at https://osf.io/24apk/.

Note that the repository currently only contains the experimental materials and experimental code. We will add raw data and analysis files after data collection at Stage 2.

- The first specified analysis, ratings before the training, appears to be frequentist, but no details of frequentist assessment criteria are given.

Response: We initially indeed planned to conduct frequentist ANOVAs on the ratings before the training. However, for the sake of consistency with the main analysis (in which we will use Bayesian mixed models), we have decided to also use Bayesian ANOVA to analyze the rating data (page 22).

> The average ratings of the items in the four training conditions were then computed, and submitted to a 2 (response, go vs. no-go; within-subjects) by 2 (consequence, approach vs. avoidance; within-subjects) by 2 (instruction group, go/no-go vs. approach/avoidance; between-subjects) Bayesian repeated-measures ANOVA in JASP. We used the default prior settings in JASP, and computed the Bayes factor for each effect across matched models. Bayes factors (BF_{01}) quantified the relative likelihood of the data under the null hypothesis against that under the alternative hypothesis. We expected the BF_{01} for the main effect of consequence, and that for the interaction effect between consequence and instruction group to be larger than 3, which would provide support for the null hypothesis (Wagenmakers et al., 2018). This would suggest that before the training, the average ratings for the approach and avoidance items were matched.

And similarly, on page 23:

> The average ratings of the items in the four training conditions were computed, and submitted to a 2 (response, go vs. no-go; within-subjects) by 2 (consequence, approach vs. avoidance; within-subjects) by 2 (instruction group, go/no-go vs. approach/avoidance; between-subjects) Bayesian repeated-measures ANOVA. We expected the BF_{01} for the main effect of go/no-go response, and that for the interaction between response and instruction group to be larger than 3. This would suggest that before the training, the average ratings for the go and no-go items were matched.

Comments by Reviewer 2 (Katrijn Houben)

The manuscript addresses an important and relevant question regarding GNG and AAT procedures as implemented in food-related research (and other applied domains). The introduction offers a comprehensive theoretical and empirical background, effectively leading to the study's aims and hypotheses. The methods are

clearly described, ensuring reproducibility, and the analyses are detailed and well-explained. I particularly also want to emphasize that I appreciate the authors' dedication to conducting rigorous and reproducible research.

Given the manuscript's high quality, I recommend its publication with only minor revisions, as explained below.

- The authors report power analyses and ultimately decide to include 100 participants per condition (N = 200 in total). In the next paragraph detailing the sample, the authors however state that they planned to recruit 160 participants (who were at least 18 years old). So there appears to be an inconsistency here that needs clarification.

Response: Thank you for spotting this mistake. We have corrected this in the manuscript. The planned initial sample size is 200, rather than 160.

- The method also describes a memory task which is not previously mentioned in the study aims/hypotheses nor in the statistical analyses section. Similarly measures of dietary restraint and hunger are collected but not further elaborated. I presume that these would be part of the exploratory analyses? Perhaps it would be good if the authors could also provide a brief rationale for inclusion of these measures.

Response: We have now added a brief rationale for the inclusion of the memory tasks (page 20):

> Previous work has shown that memory of stimulus-response contingencies correlated with training effects in both GNG and AAT (e.g., Chen & Veling, 2022; Van Dessel, De Houwer, & Gast, 2016). We therefore included memory tasks to explore the role of memory in our novel training paradigm.

On page 21, we added a rationale for including measures of dietary restraint and hunger:

> These measures were included to provide descriptive information about the sample regarding their eating behavior. They also offer additional context that may be relevant in exploratory analyses to understand how these factors might influence the effectiveness of the training paradigm.

I am confident that the research findings will contribute to current developments in this research area both from a methodological and theoretical perspective. I look forward to reading the manuscript again once the data have been collected and analyzed.

Response: We thank the reviewer for their positive evaluation and helpful feedback.