Dear Dr. Hugo Najberg,

Thank you for submitting your revised Stage 2 Registered Report to PCI RR. The same two reviewers have now reviewed your revised manuscript. Most of the previous comments have been addressed satisfactorily. However, a few issues remain. I would therefore like to invite you to submit a revised manuscript, to address these remaining comments.

Thank you for your and the reviewers' thorough insights into our work. We have improved this manuscript further to answer all the points raised. In this document you will find our answers in blue, with quotes from the main manuscript in orange. Changes are highlighted in the main manuscript in blue. We also wrote a suggestion for a second iterative study at the end of our response to the recommender.

1. One main issue, as both reviewers pointed out (and I agree), concerns what the main research question was and what conclusions we can draw based on the data. In the section "The Choice of the Comparator Group Prevents Interpreting the Primary Results" on page 19, you wrote that "our contrast cannot distinguish if the intervention resulted in an absolute increase in participants' capacity to adhere to a diet". This is true, however, this was not the research question raised at Stage 1. Instead, as mentioned in the Introduction, the question was to test the assumed active "ingredient" of the training (100% versus 50% contingency) by keeping other aspects as close as possible. What you framed as a potential risk of "inducing a non-negligible effect of training into the control condition", I would actually argue is a strength, because the research question was to test the difference between 100% and 50% contingency while keeping the other aspects the same, and this control condition allows you to do exactly that. For me, it is thus inaccurate to say that the primary results cannot be interpreted because of the control condition used. Instead, I suggest making it clear that the primary research question was to test the difference between 100% and 50% contingency, and the data give a clear answer to this question: there is no difference between the two groups. I think this is a valid and informative result in itself.

You may discuss why both groups may lead to the same absolute changes (which may explain the absence of any between-group difference), but it should be clear that the question on absolute changes was not the main research question in the first place.

We believe the misunderstanding arises from a difference between the main research hypothesis ("Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training.") and the main research question ("Can food response training modify real-world consumption behavior?"), as written in Table 1 (reported below). While we fully agree that Hypothesis 1 compares the 50% and 100% conditions, our approach was based on the assumption that the 50% condition would have no effect. This would have allowed us to isolate the absolute effect of the intervention, and thus could answer the main research question. Although the main hypothesis focuses on the 100% vs. 50% comparison, its underlying question was to test a real-world effect. Our point is that the main hypothesis and its underlying design does not answer our applied research question, which we believe is crucial for readers to understand for its practical implications.

That said, we understand the importance for RR to remain close to the hypothesis and thus reworded the Discussion according to the recommender's and reviewers' points, making it clear that the primary research hypothesis tested the difference between 100% and 50% contingency, and that the data indicate there is no difference between the two groups. This now reads p.19: "The Choice of the Comparator Group Prevents Observing a Real-World Effect" and "To answer our main research question (i.e., "Can food response training modify real-world consumption behavior?", see Table 1), our hypothesis' design relied on the control group having no or a lower effect on devaluation than the experimental group", and p.20: "While our primary hypothesis (i.e., "H1: Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training", see Table 1) is clearly null, our primary question remains unresolved because of the equivalent, non-null effect of the control intervention".

Table 1: Design

Question	**Hypothesis**	**Sampling plan**	**Analysis plan**	**Interpretation given different outcomes**	**Theory that could be shown wrong by the outcomes**	**Hypothesis outcome**
Can food response training modify real-world consumption behavior?	H1: Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training.	For 90% power, alpha = .05, and n = 140 (70 per group, based on resource constraints) for a one-sided t-test, the smallest detectable effect size would be Cohen's d = 0.50	One-sided t-test between participants in the experimental vs. control training group. If homoscedasticity assumption violated, GG correction. If $p > .05$, then BF_{01} will test the null hypothesis.	If the test is significant, then we interpret food response training as improving restrictive dieting capacities. If the test is non-significant and supported by a BF01 ≥ 3, then we interpret the result as null. If the test is non-significant, and not supported by a BF01 ≥ 3, then we interpret our results as non-conclusive.	If the hypothesis is not validated, then it would give support to an independence between the already observed food response training effects on reduction on items' valuation and in-lab consumption, and real-world consumption behavior.	Not confirmed.

2. Personally, I find the arguments provided in the section "The Role of the Number of Trained Items on the Effect of Response Training" not very convincing. For instance, it is unclear why repeating unhealthy-NoGo pairings for 150 times in the control condition is sufficient to reach the ceiling, while increasing it further to 300 times in the experimental condition makes no difference. Also, it is unclear why repeating unhealthy-Go pairings for 150 times in the control condition has no effect at all. As noted in the previous comments by one reviewer (Pieter Van Dessel), it is also unclear what underlying cognitive mechanisms these explanations exactly entail.

This section aimed to isolate the relevant differences between the two studies to help explain their differing results. Since the main distinction is the number of motoric repetitions, we concluded

that the devaluation effect may reach a ceiling after a certain number of repetitions, which would prevent any further differences between the control and experimental groups. Additionally, the similar devaluation of unhealthy items observed in both the Control and Experimental conditions suggests that unhealthy-Go pairings do not interfere with unhealthy-NoGo pairings. If they did, we would expect to see a difference in devaluation between the groups independently of any reached ceiling. This explanation has been clarified in the revised manuscript p.20 (changes highlighted in blue).

Regarding the underlying cognitive mechanisms, elaborating on theories about stimulus-response associations and the impact of repetition would stray too far from the manuscript's core purpose. Since this point was not a primary focus of our manuscript (and is not mentioned in the abstract), but rather an interesting comparison we wished to highlight, we are open to removing it if the recommender prefers.

The two groups had matched expectations in the current experiment. Do you think this can be a more parsimonious explanation for the current results (i.e., whether participants believe the training to be effective or not, see e.g. 10.1016/j.appet.2022.106041)? Did you also measure expectations in your previous work, and did you see a between-group difference there? Do you see correlations between participants' expectations and the effect of training in the current study?

In our previous work, we placed balanced expectations as positive control, and participants indeed had similar expectations (phi < 0.2) as in this work (d < 0.4). When correlating the devaluation effect to both measures of expectations, we observe in the experimental group correlations of 0.09 and 0.01, and in the control group correlations of -0.19 and -0.27 (i.e., a negative result means that the bigger the expectation, the bigger the devaluation). These are not enough to explain the null Exp vs. Ctrl devaluation effect according to us.

We now include this point p.20: "Additionally, both studies had balanced participants' expectation between both groups as positive control (phi < 0.2 in 2023 vs. Cohen's d < 0.4 in present study)".

3. For the exploratory analysis on "Diet success rate at each day" on page 17, effect sizes and the 95% confidence intervals are now reported, which is informative. However, the p value of .046 has been omitted, which I think is problematic. By looking at confidence intervals alone, it is often difficult to judge whether the difference is statistically significant or not. Since you claim that there is a difference in failure rate between these two groups, reporting the p value is necessary. I would also recommend explicitly acknowledging here that the analysis is post hoc, the p value is close to the threshold of .05 (thus, the evidence seems weak), and thus this exploratory result should be treated with great caution.

We removed the p-value to address a point by the reviewer Dr Van Dessel (cf. comment 9 of last response to reviews: "this result is given too much attention [...] despite a p-value that is not robust (p=.046) [...], which resembles data dredging or p-hacking"). Yet, because of the current rephrasing proposed by the same reviewer in their current point 7, we decided to put it back. This now reads p.20-21: "Although caution is needed because of the nature of exploratory results and because the p-value barely reaches significance (p = 0.046), one possible interpretation is that the intervention is being effective only for participants with a lower capacity to adhere to diets",

and p.17: "(ctrl vs. exp [95% CI]: 0.282[0.185; 0.37] vs. 0.18[0.1; 0.25], $t_{[179]}$ = -1.69, p = .046; CIs computed with bootstrapping; figure 9)".

4. The abstract and the conclusion should focus on the confirmatory results. If you mention the results from the exploratory analysis, I would suggest always adding the caveat that it is post-hoc, the evidence is weak, and it should be treated with great caution. And this should also not distract readers away from the main results. For the issue of "the lack of zero-effect comparator", first of all, I do not think it is an issue (see above). Second, it is not relevant for the primary research question, which is about the difference between the two groups, not about the absolute changes brought about by training. As such, I do not think the issue of the control condition belongs in the abstract and the conclusion.

Please refer to our reply to point 1 for details on this question. According to the recommender suggestion, we have now removed the mentioned sentence out of the abstract and rephrased the last sentence of the abstract by: "We propose conducting another study that includes a control training focused on non-food items. This would provide a clearer answer to our main research question: "Can food response training modify real-world consumption behavior?"".

We also have changed the following in the conclusion: "However, exploratory results hint that it could still benefit at-risks population, although the evidence is to be treated with caution because of the nature of exploratory analyses and of the weak p-value", and "The choice of a control group including unhealthy-NoGo associations and the absence of baseline measures hinder the conclusion of the absence of real-world effect which was our main research question (see Table 1, "Can food response training modify real-world consumption behavior?")".

Kind regards,

Zhang Chen

*by **Zhang Chen**, 08 Oct 2024 10:35*
Manuscript: **https://osf.io/7cepq?view_only=4934c0215f2943cfb42e019792a30b53**
version: 2

Further note to Recommender:

We realize that the nature of our control condition prevents the paper to provide, from our point of view, the information it intended to provide on the absolute effect of response training on real-world diet maintenance (cf. Question column of Table 1 of the main manuscript). Hence, also in line with the suggestions of the reviewers, and capitalizing on the possibility of Point 2.13 of the PCI RR regulations reported below, we would like to propose conducting a second study on top / as a follow-up of the present project Stage 2.

*"**2.13 Incremental registration:** Authors may add studies to approved submissions. In such cases the approved Stage 2 manuscript will be recommended, and authors can propose additional studies for Stage 1 consideration. Where these studies extend the approved submission (as opposed to being part of new submissions), the recommender will seek to fast-track the review process. This option may be particularly appropriate where an initial study reveals a major serendipitous finding that warrants follow-up within the same article. In cases where an incremented submission is rejected (at either Stage 1 or 2), authors will retain the option of having the recently approved version of the manuscript recommended. For further advice on specific*

scenarios for incremental registration, authors are invited to contact the PCI RR Managing Board (contact@rr.peercommunityin.org).”

We would like to conduct the following study to add to this paper:

To assess whether food Go/No-Go (GNG) training can improve dietary behavior, participants will first attempt maintaining a diet that restricts highly-caloric highly-palatable foods, such as chocolate. Then, as a second phase, participants who failed to adhere to the diet (e.g., for fewer than 12 days, though this criterion may be refined) will be invited to complete a GNG training focused on the target items, while a control group will undergo training on neutral objects. Following the training, participants will attempt the same diet again. The primary outcome will be the frequency of consumption of the target foods, which helps reduce data loss due to participant dropouts compared to the number of successful days of diet.

This study would differ from the previous research in several ways: (i) by specifically recruiting participants who have the potential to improve their dieting behavior, (ii) by not limiting the target items to sugary drinks alone, (iii) by using a more flexible outcome measure based on the frequency of eating occurrences, and (iv) by including a control group that aligns better with applied research objectives (i.e. without any SR mapping as was a concern in the present study and as also recommended by Reviewer 1).

Review by Matthias Aulbach, 25 Sep 2024 08:09

First off, I think the authors addressed many of the comments very well.

However, I respectfully disagree with the authors regarding their interpretation of the null results between the two conditions and the issue around the choice of control condition, even after their revisions following Pieter van Dessel's comments. In their response to Pieter van Dessel's comment, they write "Our main question was whether a 100% association training (experimental group) led to longer diet maintenance than a 50% association." – this research question has a clear answer: it did not.

The authors give explanations as to why that might be the case and their main answer is that participants in the control condition devalued stimuli as much as those in the treatment condition. They then draw the conclusion that "The Choice of the Comparator Group Prevents Interpreting the Primary Results" (page 19). The way I see it, this assessment is only true if we assume that devaluation is the (only) mechanism of action that would drive behavioral differences between groups. However, the analyses on hypothesis 2 revealed that changes in liking did not relate to successful days of dieting, indicating that this is not the mechanism by which training would change behavior. This indicates that processes other than devaluation would be driving behavioral effects in both groups. We can only speculate as to what those mechanisms are (maybe expectations of training effects? Maybe the food exposure? "demand compliant inferences" as Pieter van Dessel suggested?) and I think the authors do a good job at that. However, these alternative explanations do not change the fact that the answer to the main research question is "no".

It is, of course, true that "our contrast cannot distinguish if the intervention resulted in an absolute increase in participants' capacity to adhere to a diet" (page 19) but that was not the question asked in the first place – the research question did not refer to changes but to differences between two specific tasks. These differences did not emerge, and the Bayes Factor implies equality between groups.

I think an interesting implication of this study's results then is that the contingency of the pairing does not matter for devaluation to occur. This begs the question: if we were to run the study with a different control condition (say, waitlist control) as the authors propose, what should the intervention look like? Based on the current results, the contingency between stimuli and reaction does not seem to matter (all else being equal).

As also reported in the response to the recommender's point, we believe the misunderstanding arises from a difference between the main research hypothesis ("Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training.") and the main research question ("Can food response training modify real-world consumption behavior?"), as written in Table 1 (reported below). While we fully agree that Hypothesis 1 compares the 50% and 100% conditions, our approach was based on the assumption that the 50% condition would have no effect. This would have allowed us to isolate the absolute effect of the intervention, and thus could answer the main research question. Although the main hypothesis focuses on the 100% vs. 50% comparison, its underlying question was to test a real-world effect. Our point is that the main hypothesis and its underlying design does not answer our applied research question, which we believe is crucial for readers to understand for its practical implications.

That said, we understand the importance for RR to remain close to the hypothesis and thus reworded the Discussion according to the recommender's and reviewers' points, making it clear that the primary research hypothesis tested the difference between 100% and 50% contingency, and that the data indicate there is no difference between the two groups. This now reads p.19: "The Choice of the Comparator Group Prevents Observing a Real-World Effect" and "To answer our main research question (i.e., "Can food response training modify real-world consumption behavior?", see Table 1), our hypothesis' design relied on the control group having no or a lower effect on devaluation than the experimental group", and p.20: "While our primary hypothesis (i.e., "H1: Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training", see Table 1) is clearly null, our primary question remains unresolved because of the equivalent, non-null effect of the control intervention".

We also agree with the necessity to run another study with a different control group and propose to do so as an incremental registration to the present study. Please refer to our reply to the recommender for details.

Regarding the issues around the choice of control condition, I further refer the authors to (Kakoschke et al., 2018). While that article is about Approach-Avoidance Training, the same logic applies here and has been spelled out by the authors, but it might be good to cite this paper, too.

We thank the reviewer for pointing out this relevant literature which is now mentioned p.19 and 20.

One another note, I think there was a misunderstanding regarding my comment 11: I did not mean to say that the dieting phase impacted the devaluation but merely the instruction to avoid those and/or the participants' decision to try to avoid them.

Thank you for this clarification.

Review by Pieter Van Dessel, 23 Sep 2024 11:53

The authors have done a commendable job revising their manuscript. I have only a few additional suggestions:

1. In the Abstract, the authors write: “We interpret this result as the effect on diet maintenance reaching ceiling in both groups.” This phrasing does not sound very objective or scientific, as it suggests there is only one correct interpretation. I recommend revising this to: “One possible interpretation of this result is that…”

Agreed, this has now been edited accordingly.

2. In the same sentence, the authors state that this hypothesis (which I would call an interpretation) “is supported by the finding of equivalent target item devaluation in both groups.” I’m missing the logic here. If you are arguing that the effect reached a ceiling, why then support this with a statement that devaluation was equivalent? I believe the authors are not referring to ceiling effects here (and I did not see evidence for ceiling effects, in the sense that nearly everyone maintained their diet). Instead, I think they aim to interpret the null result as suggesting that training in both groups may be equally effective. This interpretation is indeed (to some extent) supported by the equivalent target item devaluation.

We agree with the reviewer’s interpretation and thank them for pointing out this mistake. This mention of ceiling might have been a relic from re-writing the abstract multiple times. This now reads: “One possible interpretation of this result is that the training created an equivalent effect in both groups, a hypothesis supported by the finding for equivalent target item devaluation in both groups”.

3. The following sentence in the Abstract does not logically follow from the previous one, and it is also slightly unclear: “Food response training may also have not improved restrictive dieting adherence in a resourceful, healthy population, as supported by a difference in dieting adherence found only in participants with early failures (18% failure in the experimental group vs. 28.2% in the control group at first quartile).” I suggest rephrasing this to indicate that an alternative interpretation for the null finding (compared to the idea that both trainings impact responses to the same extent) is that there is a difference between the two types of training only in groups that struggle with diet adherence, but this difference is not observed overall due to the small number of participants with such issues in this study. You could then refer to the initial evidence for this by noting the small difference in dieting adherence when considering only participants with early failures.

Agreed, we have now clarified this section. It now reads: “Another possible interpretation is that the training only induced an effect on the few participants prone to fail the diet early, while we recruited mostly resourceful healthy population, as supported by a difference in dieting adherence found only in participants with early failures (18% failure in the experimental group vs. 28.2% in the control group at first quartile)”.

4. In the Discussion, the title “The Choice of the Comparator Group Prevents Interpreting the Primary Results” seems inaccurate. The choice was actually well-considered, so it's unclear why it should be reconsidered simply because no effect was observed. The study showed clear

evidence that the difference in contingencies is not enough to induce a difference in overall diet maintenance and explicit ratings. That is a clear and valid result and it should be highlighted as such. However, as with every result, there are several possible interpretations. I suggest first stating this result clearly—it seems valid enough, with moderate evidence (looking at the Bayes factors)—and then discussing possible explanations, such as the possibility that the control training had a strong effect.

Please see our combined response to points 4 and 5 in the next section.

5. Further in that discussion, the authors state: "However, it entails the risk of inducing a non-negligible effect of training into the control condition. Our design assumed that the control group would always have a lower effect on devaluation than the experimental group and could thus be used for an unequivocal interpretation of the mechanistic effect of an intervention." I disagree with these statements. A design does not assume anything, and inducing a non-negligible effect in the control condition is not a "risk." The control group was designed to control for everything except the contingencies, and this is exactly what happened. Hence, the results are clear and they are valid. However, there is also another question: whether response training produces any effect. This question is also a valid question but it is not the focus of the current study as it would need to be answered with a different design. Hence, one should first explain and discuss the key finding (no effect of the contingency manipulation) and only then note that this of course does not mean that response training had no effect. In fact, one possible explanation is that both groups produced similar effects...

As also reported in the response to the recommender's point, we believe the misunderstanding arises from a difference between the main research hypothesis ("Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training.") and the main research question ("Can food response training modify real-world consumption behavior?"), as written in Table 1 (reported below). While we fully agree that Hypothesis 1 compares the 50% and 100% conditions, our approach was based on the assumption that the 50% condition would have no effect. This would have allowed us to isolate the absolute effect of the intervention, and thus could answer the main research question. Although the main hypothesis focuses on the 100% vs. 50% comparison, its underlying question was to test a real-world effect. Our point is that the main hypothesis and its underlying design does not answer our applied research question, which we believe is crucial for readers to understand for its practical implications.

That said, we understand the importance for RR to remain close to the hypothesis and thus reworded the Discussion according to the recommender's and reviewers' points, making it clear that the primary research hypothesis tested the difference between 100% and 50% contingency, and that the data indicate there is no difference between the two groups. This now reads p.19: "The Choice of the Comparator Group Prevents Observing a Real-World Effect" and "To answer our main research question (i.e., "Can food response training modify real-world consumption behavior?", see Table 1), our hypothesis' design relied on the control group having no or a lower effect on devaluation than the experimental group", and p.20: "While our primary hypothesis (i.e., "H1: Participants in the experimental training will report more successful days of high sugary drinks restrictive dieting than the control training", see Table 1) is clearly null, our primary question remains unresolved because of the equivalent, non-null effect of the control intervention".

Regarding the comment: “However, there is also another question: whether response training produces any effect. This question is also a valid question but it is not the focus of the current study as it would need to be answered with a different design”. We propose to correct this discrepancy between the main research question, main hypothesis, and the design, with a new, incremental registered study added to the present Stage 2. Please refer to our reply to the recommender for details.

6. On page 20, the authors write: “We explain the smaller Group x Session interaction in the current vs. the 2023 study by…” This could be interpreted as the authors being biased toward supporting only one interpretation when there are several others (such as Type I or Type II errors). It would be better to say: “One possible explanation for the smaller Group x Session interaction in the current vs. the 2023 study is that…”

This phrasing is now being used. This now reads p.20: “Given these differences in parameters, one possible explanation for the smaller Group x Session interaction in the current vs. the 2023 study is that […]”.

7. On page 21, the authors state: “We interpret this exploratory result…” Here, I would also suggest presenting this as one possible interpretation and I would also advise more caution. For instance: “Although caution is needed because this is an exploratory result with confidence intervals barely reaching significance, one possible interpretation is that…”

This phrasing is now being used. This now reads p.21: “Although caution is needed because of the nature of exploratory results and because the p-value barely reaches significance ($p = 0.046$), one possible interpretation is that the intervention is being effective only for participants with a lower capacity to adhere to diets”.

8. I noticed the word “expect” used a couple of times when the authors likely meant “except.”

This has now been corrected p19 and 20.