Dear Dr Gaillard and peer reviewers,

Thank you for the thoughtful responses on our manuscript. We appreciate its difficult finding reviewers now, so we valued the largely positive comments. Below we have separated the reviewer comment, our response, and the location of any relevant change in the revised manuscript.

Reviewer 1

|  |  |  |
| --- | --- | --- |
| Reviewer comment | Location | Our response |
| First, some of the details of the experimental procedure are a bit fuzzy to me. Were all pairings between images presented (i.e., fully factorial design: smoking-neutral; neutral-neutral; smoking-smoking) or did all presented pairs contains one smoking and one neutral cue? Obviously, the latter is the case of interest, but I was curious if these other, less informative, trials were present and, if so, how they processed (simply removed, used for exclusion but nothing else, etc.).  Also, which IAPS pictures were used as the neutral cues? I assume IAPS pictures labeled as neutral, but this would be important to confirm given the well-documented attentional bias effects towards threatening or negative imagery (e.g., Kappenman et al., 2015). Finally, it was not indicated how long the inter-trial fixation cross remained on the screen for. Many of these points could be addressed with the addition of a task figure to the manuscript, which I would recommend overall to readers who may be less familiar with the typical visual probe task. | Page 9, line 152 (visual probe task section)  Page 11, line 187 (task diagram)  Page 27, line 461 (disclosures) | Thank you for the prompt to make the materials clearer to readers.  We realised we relied on the open code to provide some of the task and processing details. We have expanded the visual probe task section in the method and added a new data processing section to the results to explain how the neutral trials were excluded prior to analysis and the confirmatory analyses here focus on smoking/non-smoking trials.  In the original version of the manuscript, we included a disclosures section before the reference list which summarised the code/data links and provided a link to the Gorilla open materials page. We could create an anonymous version of the OSF link, but not to Gorilla, so we omitted the link from the previous version of the manuscript to comply with blinding for peer review. However, we have added the Gorilla open materials link back into the manuscript to demonstrate our response to this point.  All the images are available on the Gorilla open materials, but for ease of access, we have added a spreadsheet with the IAPS ID numbers, and valence and arousal means from the original validation documents. The mean valence was 6.26 and mean arousal was 3.94. The original scale was 1 (low) to 9 (high), so our neutral images can be interpreted as on average moderately pleasant and weakly arousing.  We have also added a task diagram into the manuscript as Figure 1 to make the procedure easier to follow for readers who do not want to follow a link to Gorilla to preview the experiment |
| Second, the authors test (rather well might I add) for the presence of a between-group effect: i.e., the existence of a difference in attentional bias between daily and non-daily smokers. At a more fundamental level however, the results presented in Figure 2 suggest to me that a basic attentional bias was not found within either group, beyond a difference in the group. While I am not an expert in attention bias and smoking research, this struck me as important, given that the authors cite many papers purportedly demonstrating this effect (albeit in a less-powered design, collapsing across daily and non-daily smokers). I suggest the authors interrogate the basic attentional bias effect, across groups and SOAs, which could be done in their current statistical framework by analyzing the significance of the intercept term in their ANOVA. Eyeballing the descriptive results reported in the paper, and Figure 2, it does not seem to me that any significant bias effect was found. If this analysis is done correctly—and with the degree of scrutiny the authors have committed to the rest of their analysis—I think this could be a strong challenge to what is apparently a quite widespread finding, given that, by the author’s own admission, their design is better powered than those in the past. | Page 20, line 320 | Thank you for this suggestion. For our second set of exploratory analyses, we included trial type as an additional IV and explored whether participants showed the predicted attentional bias effect towards smoking images.  The prompt to question whether participants showed the attentional bias effect reminded us of Grice et al. (2020) and the idea of persons as effect sizes. Therefore, we added the percentage of participants who provided faster responses to smoking images. These rarely deviated from 50%, showing almost chance level responses to whether each group responded faster to smoking images. We also added an additional figure to visualise these results, colour-coding whether participants showed a negative slope for faster responses to smoking images.  Given the non-significant exploratory 2x2x2 ANOVA and the percentage of participants showing the attentional bias effect, there was not strong enough evidence to conclude participants produced an attentional bias effect in the first place, in addition to the lack of meaningful difference between smoking groups. |
| Lastly, I would suggest the authors briefly describe Laken’s (2018) small telescopes method for the readers. This is not yet a common statistical tool and I think it would bolster the impact of Figure 3 for readers to understand what is at stake with this analysis—namely that we can conclude, with high confidence, that the difference between groups is basically zero in both SOA conditions. In line with my previous comment, I would recommend running the same analysis on the basic attentional bias effect as well, using the same rationale. | Page 17, line 285 | We have expanded the exploratory analyses section for equivalence testing. We have added a paragraph explaining the logic behind equivalence testing and commented on our use of the small telescopes method of choosing an effect size boundary. We comment on the robustness of these choices given different options and explain which method we focus on for our conclusions given the analysis was exploratory.  Given the non-significant 2x2x2 mixed ANOVA for exploratory analyses and only half the participants showed an attentional bias effect, we do not think it is necessary to report an additional set of equivalence tests for each combination of smoking group and SOA condition, noting whether the difference between smoking and non-smoking trials is statistically equivalent. We provide full open data and code, so interested readers are welcome to explore this further. |
| On p. 4 the authors state that “This [the lack of reliability in many cognitive tasks] is not necessarily a problem for experimental research as the tasks are designed to emphasise differences between groups or conditions”. I am not totally convinced this is right, because, of course, for a test to have any resolution to predict group difference it must have some ability to discriminate between individuals. To take the extreme case, it is not possible for a test to be entirely unreliable at the individual level, but predict a group difference in the aggregate. I am nitpicking here to be sure, but perhaps consider rephrasing this section—after all, the importance of reliability in assessing group differences in the visual probe task is a key contribution of the present paper. | Page 6, line 98 | We have made it clearer this is the message from Hedge et al. (2018), not a claim from us, but toned down the comment to explain the competing aims of experimental and correlational measures. On pages 2-3 of Hedge et al., they explain different meanings of reliability. For correlational / individual differences research, reliability decreases for smaller between-subject variance. However, in experimental research, large within-subject variance makes it difficult to produce robust and replicate effects. This is behind their explanation that tasks like the Stroop replicate almost 100% of the time, since you get a Stroop effect in almost any group of participants. However, large within-subject variation means the reliability across time is poor. We have reworded it to make this distinction clear, since our conclusion is the traditional methods in attentional bias research using response times are not suited to correlational / individual differences research, such as cognitive bias modification. |
| On the bottom of p. 6, the authors state “This procedure was repeated twice and presented in two blocks, creating 384 trials overall with 64 trials in each SOA and trial type condition.” What is the trial type condition? This is the only time this term is used. | Page 9, line 152 (visual probe task section) | Our response to the first comment covers this concern. In hindsight, the details of our task and design were not clear when we omitted the Gorilla open materials link. We extended the visual probe task description in the method and explained the role of the trial type condition in the processing section of the results. |

Reviewer 2

|  |  |  |
| --- | --- | --- |
| Reviewer comment | Location | Our response |
| Where is the logical connection between the data shown and the arguments made about these two groups (their differneces, why those exist, why readers might care).  Maybe the authors set out to resolve some mixed findings in the literature, but they ended up with another mixed finding.  Was there a specific hypothesis and rationale? Why were there seperate groups based on daily or non daily smoking?  What is gained by finding or not finding a group difference between daily and non daily smokers?  Too many times I was left with the impresssion that the authors wanted the collected data to speak for itself, without regaurd for the motivation to the research question. | Page 4, line 51 | Thank you for highlighting this gap. In a previous submission with stringent word limits, we omitted this content to focus on the key components of the study. We agree the connection between smoking behaviour and attentional bias was not clear for the implications, so we have added a new paragraph to focus on theoretical models of attentional bias, then commented on competing predictions for which smoking group should produce greater attentional bias. |
| Investigating attention bias is misguided with the typical approach: RT differences. It should only (and is rarely) be approached with methods that could gaurentee a higher chance of obtaining an effect with practical significnace (something with a SMD greater than .4, see Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. <https://doi.org/10.1037/a0015808>).  RT differences are simply unable to produce a reliabile enough estimate to correlate with, well, much of anything (see the simons task and others here: Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. <https://doi.org/10.3758/s13428-017-0935-1>).  Given that the methods used here have such high limitations (although commonly used by researchers), their data cannot speak much to a research question. | Page 30, line 320 | We agree that RT differences are potentially problematic. That is the argument we are making in the manuscript as there are no meaningful differences in attentional bias between daily and non-smokers, using methods commonly used in the field. In response to reviewer one’s comments, we also provide exploratory analyses where we include trial type as an IV and RT as an outcome, and our conclusions did not change. Studies focus on RT differences using the visual probe task and compare their groups of interest or investigate how attentional bias changes over time. Our argument - and the reason we submitted to the Journal of Trial and Error - is we believe this approach does not provide meaningful results. Given the large sample and analyses supporting statistical equivalence, we would encourage researchers not to focus on this approach in future.  We also tried more advanced methods like a generalised mixed effects model to avoid treating response time as strictly normal and included participants and the stimulus pairs as random effects. Our conclusions still did not change with the only significant effect of SOA, meaning participant’s responses were different depending on the presentation time of stimuli. We have not included this in the manuscript as the exploratory analyses keep increasing, but since we include open data, interested readers are welcome to try other methods.  For reliability, this is also a key message from our manuscript. Researchers rarely report reliability measures for their cognitive tasks and we already cited Hedge et al. in our rationale for reporting the reliability estimates for our visual probe task. Combined with the lack of meaningful group differences, the poor reliability of the visual probe task means our take home message is researchers should not rely on RT differences in attentional bias research. |