Biological Psychology 152 (2020) 107866

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Commentary

# Does attention bias modification induce structural brain changes? A commentary on Abend et al. (2019)

Sam Parsons

*Department of Experimental Psychology, University of Oxford, United Kingdom*

ARTICLE INFO

*Keywords:*
Attention bias modification
ABM
Structural brain changes
Commentary
Statcheck
Statistical power

ABSTRACT

The strength of research conclusions must follow from the strength of evidence. In this commentary I raise several issues with a recent paper "Brain structure changes induced by attention bias modification training" (Abend et al., 2019). I follow the paper's five highlights to discuss; the absence of discussion of negative ABM results, the low power of the study itself. Centrally, I raise the concern that the conclusions rest all their weight on a single statistically significant group x time interaction form brain-wide analysis. Most importantly, this test is not internally consistent following statcheck. The reported result requires checking, and potentially a correction that would shift the results entirely.

Can only five sessions of attention bias training (ABM) induce structural brain changes? In a recent Biological Psychology paper, Abend et al. (2019) argue yes. With merely five Dot-Probe based ABM sessions, we can expect changes in the inferior temporal cortex - brain changes that may underlie ABM's therapeutic effects on anxiety. The key (and only) significant interaction finds that Fractional Anisotropy (FA) significantly changes over the training period in the ABM group (n = 15), but not the attention control group (n = 14). This was the single group x time interaction effect to survive FDR correction from a whole-brain analysis. This result is used to support the conclusion (and paper title), that ABM induces brain structure changes.

I argue that the conclusions of this paper are overstated and, more importantly, the core result may require a correction. I have structured this commentary following the study's highlights and raise several methodological, statistical, and reporting issues.

## 1. Attention bias modification (ABM) reduces anxiety via attention training tasks

In their first highlight, Abend et al. portray an optimistic representation of the ABM literature. To be clear, the authors did not test the hypothesis that ABM reduces anxiety. Throughout the paper, the authors maintain that ABM has "robust clinical effects" (p. 5) and only refer to studies supporting ABM, such as an early meta-analysis (Hakamata et al., 2010). However, many papers have emphasised the severe limitations of the dot-probe task, ABM, and call into question the quality of ABM research. For brevity, I direct the reader to a meta-analysis of cognitive bias modification (including ABM) effects on anxiety and depression that finds few positive conclusions (Cristea, Kok, & Cuijpers, 2015). Outliers drove significant positive effects of ABM and removing a single one reduced beneficial effects to non-significance. The authors conclude that, low-quality under-powered studies and significant publication bias plague the ABM literature. By selectively citing only positive results the authors leave the reader with an unrealistic impression of the literature.

## 2. We mapped ABM-induced (vs control) short- and longer-term structural changes

Critical readers might question this study's capability to map brain structure changes with precision, given the small sample size (n = 29 included in the final analyses). Low power reduces the likelihood of detecting true effects. The knock-on effect is that statistically significant effects are less likely to reflect true effects, i.e. are more likely to be false positives. Button et al. (2013) argue that small sample sizes undermine the reproducibility of results, as significant effect sizes are likely overestimates of the true effect. This is more likely to be the case when only significant results are reported. A more recent paper reports that sample sizes of at least 80 are required to recover medium effect sizes (Geuter, Qi, Welsh, Wager, & Lindquist, 2018). With this in mind, far larger samples would be needed to confirm Abend et al. core result. To the authors credit, the small sample size is the first limitation discussed. However, the authors use this limitation to argue that larger samples would help find more interactions and offer greater precision to their results. These interpretations suggest that the authors already consider the result to be robust. Yet, the low power of the study

*E-mail address:* sam.parsons@psy.ox.ac.uk.

https://doi.org/10.1016/j.biopsycho.2020.107866
Received 29 September 2019; Accepted 10 February 2020
Available online 19 February 2020

suggests a different conclusion; the result is likely a false positive.

ABM researchers, indeed all researchers, must avoid this error in statistical reasoning. We must be vigilant to the, incorrect, assumption that adequately powered studies will conform to - or find 'better' results than - their underpowered counterparts.

## 3. ABM led to specific longer-term structural changes in inferior temporal cortex

The authors report three significant tests ($p < .05$ following FDR correction, the number of tests corrected for was not reported) for the whole brain analysis (each was found in a different location). Two were main effects of time. I will focus on the significant group x time interaction, $F(2,54) = 12.76$, $p = .00001$, as this single test bears the weight of the authors conclusions.

This statistical result is not internally consistent. The *p*-value, as reported, does not match the *F* statistic and degrees of freedom. I used statcheck to examine the reported statistic (Epskamp & Nuijten, 2016). Statcheck is a spell-check for statistics; it compares the test statistic and degrees of freedom to the *p*-value. The *p*-value matching $F(2,54) = 12.76$ is $p = .00002895806$. I checked the smallest *F* value for the test to become consistent. $F = 13.70$ (to two decimal places) was the smallest value consistent with the reported *p*-value. Such inconsistencies are not uncommon (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016), and the cause is likely innocuous. However, the authors should check these reported statistics and, if necessary, issue a corrigenda. Readers can find the code at https://osf.io/cws56/.

My aim is not to act as a *p*-value pedant. Whole brain analyses can include many thousands of tests. The number of tests performed will determine whether correcting the *p*-value would be the difference between a significant and non-significant (FDR-corrected) result. To illustrate, an FDR corrected $p = .00001$ becomes $p = .05$ with 5000 comparisons. Whereas, with 5000 corrections the corrected *p*-value becomes $p = .1447903$. On the other hand, more than 1725 comparisons would lead to the corrected *p*-value of $p = .00002895806$ being greater than .05, and thus, non-significant following correction. In addition to checking the reported statistics, the authors should make the reader aware of the number of tests being corrected for.

Assuming that the interaction remains significant following a potential correction, we should lend attention to the direction of change. Readers may miss that fiure 2, panel A, plots normalised changes in FA (compared to the first scan). As the authors report in table 1, both group's FA values converge in the third scan - unlike the divergence implied by the figure. At baseline, the ABM and ACT (control group) FA values were .136 and .125, respectively. Two weeks later, at the third scan, FA values were .128 and .127. Thus, the baseline differences in FA between groups lessens as a result of training. The obvious question is: Were ABM-related training-related effects observed, or was the effect due to regression to the mean? Incorporating regression to the mean into the interpretations of results might lead us to question the strength of inferences we can draw from this single result.

The single significant interaction leads the authors to conclude that these structural brain changes are region specific. This interpretation requires a higher standard of evidence than reported. First, no evidence is provided that other areas did not change - i.e. absence of evidence is not evidence of absence. Likewise, the small sample size renders distinguishing between significant and non-significant results impossible. A well-powered, pre-registered, replication with the aim of testing region specific brain changes is needed to support these conclusions.

## 4. Temporal changes in prefrontal, occipital cortices were noted across groups

Significant main effects of time on FA were reported; one in the ventromedial prefrontal cortex, and one in the middle occipital gyrus. In both cases FA fluctuates; decreasing from scan 1 to scan 2 then increasing from scan 2 to scan 3, and vice versa. It is not clear if participant's repeated engagement in an attentional task, natural fluctuations due to measurement error or regression to the mean, caused FA changes. For instance, changes from scan 1–2 occurred during the same testing session. Is it plausible to conclude that structural brain changes would occur in the course of a single testing session? The authors suggest that weakening FA in one region relates to strengthening FA in another, r(29) = -.53 95 % CI [-.75, -.20] (I computed confidence intervals to highlight the precision in the estimate, see https://osf.io/cws56/ for code). However, correlations at this sample size are not stable. Schönbrodt and Perugini (2013) demonstrated that sample sizes need to be in the hundreds to achieve adequate precision. Well-powered - ideally pre-registered - replication attempts are necessary to test the accuracy of these conclusions.

## 5. ABM training induces structural changes; these may relate to clinical effects

Throughout this commentary, I have argued that Abend et al.' conclusions require greater evidence. The conclusion rests on a single significant interaction, for which the p-value may require correcting, which may render the result non-significant, and that could be explained in large part by regression to the mean. Low statistical power increases the likelihood that reported effects are false positives. The literature on ABM related clinical effects is much less certain than suggested. Nor were clinical effects examined in this study. Thus, suggesting that ABM induces structural changes in the brain, which may underlie clinical effects on anxiety, oversells the evidence reported. More substantial evidence would be required to support the claim that ABM training induces structural brain changes and to infer any relationships to clinical effects.

I raise these comments as part of a wider commentary on ABM research. Mixed reports of ABM's clinical efficacy must be incorporated into any discussion of the literature. Similarly, the continued use of the dot-probe task as the go-to task of choice, despite well documented reliability issues, renders questions of individual differences near-unanswerable (for a summary and further reading, see Parsons, Kruijt, & Fox, 2019). Larger sample sizes than typically used are necessary for ABM research to provide robust evidence of ABM-induced behavioural, structural, and clinical effects. The strength of conclusions must follow from the strength of the evidence for ABM research to function as a worthy scientific endeavour.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

Sam Parsons is currently supported by a grant from the Economic & Social Research Council (ESRC), UK. Grant code: ES/R004285/1.

## References

Abend, R., Rosenfelder, A., Shamai, D., Pine, D. S., Tavor, I., Assaf, Y., et al. (2019). Brain structure changes induced by attention bias modification training. *Biological Psychology, 146*, 107736.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365.

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: Meta-analysis. *The British Journal of Psychiatry, 206*(1), 7–16.

Epskamp, S., & Nuijten, M. B. (2016). *Statcheck: Extract statistics from articles and recompute p values.* Retrieved from http://CRAN.R-project.org/package=statcheck. (R

package version 1.2.1).
Geuter, S., Qi, G., Welsh, R. C., Wager, T. D., & Lindquist, M. A. (2018). Effect size and power in fMRI group analysis. *Biorxiv*295048.
Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J., C., Fox, N., A., Leibenluft, E., … Pine, D., S. (2010). Attention Bias Modification Treatment: A Meta-Analysis Toward the Establishment of Novel Treatment for Anxiety. *Biological Psychiatry, 68*, 982–990. https://doi.org/10.1016/j.biopsych.2010.07.021.
Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205–1226.
Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395. https://doi.org/10.1177/2515245919879695.
Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612.