
# The Efficacy of Attentional Bias Modification for Anxiety: A Registered Replication

Nathan Pond[a*], Frances Meeten[ab**], Patrick Clarke[c], Lies Notebaert[d], & Ryan Scott[a**]

[a] School of Psychology, University of Sussex, Brighton, BN1 9RH, UK

[b] Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, SE5 8AB, UK

[c] Cognition and Emotion Research Group, Faculty of Health Sciences, Curtin University

[d] Centre for the Advancement of Research on Emotion, School of Psychological Science, University of Western Australia

*Corresponding author: np286@sussex.ac.uk, **Joint Senior Author

Stage 1 Registered Report available here: https://psyarxiv.com/cf4xz

Acknowledgements

Nathan Pond is in receipt of a University of Sussex, School of Psychology funded PhD studentship.

Dr Frances Meeten receives salary support from a Medical Research Council (MRC) New Investigator Grant (MR/W005077/1).

**Abstract**

Generalised anxiety disorder is a prevalent condition linked to the presence of cognitive biases, including attention bias. Attention bias is the tendency to attend preferentially to threat-related stimuli and has been consistently observed in high anxious samples. Interventions to modify these biases have been developed with the aim of alleviating anxiety symptoms. However, while initial studies were promising, the current evidence-base for attention bias modification (ABM) procedures to alleviate symptoms is mixed. Furthermore, concerns have been raised regarding the potential for demand effects to be underlying previous significant findings. We revisited the efficacy of ABM, conducting a replication of a seminal study showing successful alleviation of anxiety symptoms following multi-session ABM training (Hazen et al., 2009). As a secondary aim, we quantified the potential influence of demand effects in the paradigm. Participants ($N = 104$) classed as high worriers attended seven lab-based sessions. This included five probe-based ABM training sessions (or sham-training control), and a pre- and post-training session in which levels of attention bias, worry, trait anxiety and depression were assessed. We adopted a Bayesian approach to analyses with a series of Bayesian t-tests evaluating the change in each outcome measure from pre- to post-training. These analyses revealed sensitive evidence for the null hypothesis of no effect for all measures except the measure of depression (which was insensitive), and no sensitive evidence for demand effects. These findings join a growing body of literature suggesting that probe-based ABM training does not lead to a robust reduction in bias or anxiety. Therefore, alternative ABM paradigms should be investigated.

*Keywords:* anxiety, worry, GAD, attention bias, CBM-A, cognitive bias modification, ABM

Supplementary material, including the raw data arising from the study, are available here: https://osf.io/mw8p6/

Generalised anxiety disorder (GAD) is a common and debilitating condition characterised by excessive and uncontrollable worry, as well as somatic symptoms including restlessness, muscle tension and fatigue (DSM-V; APA, 2013). Furthermore, GAD has been associated with impaired social and occupational functioning and increased suicide risk (Wittchen et al., 2002). GAD is estimated to have a lifetime prevalence of 3.7% according to a cross-nationally representative survey by Ruscio et al. (2017), with this figure rising to 5% in high-income countries. However, the prevalence rate has likely risen significantly since this estimate was produced, with global events such as the COVID-19 pandemic having led to a sharp increase in GAD symptoms in the general population (Jia et al., 2020). Despite the devastating impact of this condition, the leading treatment, Cognitive Behavioural Therapy (CBT), has a variable success rate, with only 50% of individuals receiving CBT for GAD achieving full remission (Loerinc et al., 2015). Given the increasing prevalence of GAD, it is essential that we identify new treatment approaches to reduce GAD symptoms.

In the model outlined by Hirsch and Mathews (2012), pathological worry (the key cognitive component of GAD) is associated with two key cognitive biases, interpretation bias and attention bias. Interpretation bias can be defined as the tendency of anxious individuals to interpret ambiguous information as threatening, whereas attention bias is the tendency to attend preferentially to threat-related stimuli (Hirsch & Mathews, 2012). Given the diverse nature of worry topics present in GAD samples, it is difficult to describe one 'archetypal' example of an attention bias present in GAD, as it is characterised by a generalised attentional bias to a variety of minor emotional/threatening stimuli that is not present in non-anxious populations (Mogg & Bradley, 2005). One such example would be an attentional bias towards

negative and/or threatening words, or to negative faces (Mogg & Bradley, 2005). However, attention biases are also cross-modal and alongside visual attentional bias, there is also evidence to support auditory attentional bias (e.g., Wang et al. (2019)). Hirsch and Mathews (2012) argue that these cognitive biases lead to the more frequent intrusion of negative threat representations into conscious perception, and thus increase the occurrence of prolonged worry periods characteristic of GAD. Though both biases play an important role in this process, whereas modification of interpretation bias and its effects on anxious symptomatology are robustly reported in the literature, the evidence is mixed regarding the malleability of attention bias compared to interpretation bias (Cristea et al., 2015; Liu et al., 2017), thus this paper focuses on attention bias.[1] In addition to the Hirsch and Mathews (2012) model, a number of theories have been proposed suggesting an important role for attention bias in both the development and maintenance of fear and anxiety (for a review, see Van Bockstaele et al., 2014), suggesting that the influence of this cognitive bias extends robustly beyond pathological worry, and into the context of anxiety. Indeed, an in-depth assessment based on Hill's (1965) criteria of causality concluded that there is evidence of a causal relationship between attention bias and anxiety (Van Bockstaele et al., 2014). Furthermore, research has shown that attention biases are prominent in GAD samples, with a meta-review by Goodwin et al. (2017) reporting

[1] It is worth noting that combined cognitive bias hypotheses are an important part of the literature, as we know that cognitive biases, such as interpretation and attention, do not necessarily operate in isolation (see Hirsch et al., 2006). For example, Hirsch and Mathews (2012) note in their model that high worriers interpret ambiguous information in a negative manner (interpretation bias), and these negative thoughts may potentially become the focus for attention bias. However, given that this report is a registered replication of a study of attention bias, and that findings regarding the malleability of attention bias have been mixed, the focus of the present report remains exclusively on attention bias.

that 69% of the 29 studies reviewed found a reliable attention bias towards threatening stimuli in GAD populations.

Given the role of these cognitive biases in GAD, it is unsurprising that they are being studied as potential targets for modification in the treatment of GAD and other anxiety disorders. Research has shown that multi-session interpretation bias modification training, in which individuals are presented or asked to generate positive or benign interpretations to emotionally ambiguous situations (Beard & Peckham, 2020), is effective at reducing GAD symptoms (Hirsch et al., 2018; 2020; 2021; Ji et al., 2021). Given this success, naturally the potential efficacy of attention bias modification (ABM) training is also a subject of interest. A common method of ABM training is threat-avoidance training (Mogg et al., 2017). Threat-avoidance training often implements visual-probe tasks to reduce attentional bias towards threat. In these tasks, two cues (one threat-related, one neutral) are presented to participants simultaneously. When the cues disappear, a target-probe is located in place of one of the cues and participants are required to respond to that target as quickly as possible. To train threat-avoidance, the target-probe is more likely to appear where the neutral cue was previously located, thus encouraging participants to orient their attention consistently away from the threat cue.

While there is a sound theoretical rationale for implementing ABM training interventions to reduce anxiety symptoms, the experimentally assessed efficacy of ABM training has produced mixed results (Martinelli et al., 2022). Given the large amount of variability across different ABM training paradigms, including variances in ABM task paradigms (dot-probe; visual probe; etc.), training settings (offline; online) and stimuli used (pictures; words), this is perhaps unsurprising (Hallion & Ruscio, 2011; Martinelli et al., 2022). Seminal studies by Amir et al. (2009) and

Hazen et al. (2009) that were the first to assess the efficacy of an ABM training intervention for GAD delivered promising results, with both studies finding a significant alleviation of reported anxiety in both a clinically diagnosed GAD sample, and a high-worry student sample, respectively. However, subsequent meta-analyses of studies into the effectiveness of ABM training for alleviating anxiety disorders have reported varying levels of efficacy, with some reviews suggesting moderate to large overall effect sizes (Hakamata et al., 2010; Linetzky et al., 2015), while others have reported very small, and in some cases non-significant, overall effect sizes (Cristea et al., 2015; Hallion & Ruscio, 2011). This has led some researchers to question the therapeutic value of ABM interventions for anxiety, with Cristea et al. (2015) arguing that the previously reported efficacy of ABM training interventions may be the result of demand effects. The moderator analysis by Cristea et al. (2015) showed that effect sizes were larger for studies conducted in a laboratory setting in which demand effects are known to be more influential. Furthermore, the authors also reported that control conditions in the studies they reviewed often led to improvements in anxiety symptoms equal to that of the ABM intervention conditions, further suggesting that the presence of expectancy effects may be influencing responding.

Despite the discouraging findings of some reviews, researchers have noted that while the increase of null findings, particularly in ABM interventions targeting social anxiety disorder (SAD) and post-traumatic stress disorder (PTSD), has led some to become wary of ABM training as an anxiety intervention, this wariness may be premature (Clarke et al., 2014; Grafton et al., 2017). Grafton et al. (2017) note that there has been a confusion in the cognitive bias modification literature, in that studies that *attempted* to modify cognitive biases and studies that *actually* managed

to modify cognitive biases are oft referred to using the same label of cognitive bias modification training. Thus, in meta-reviews the efficacy of studies attempting to alleviate anxiety symptoms that both did and did not manage to actually modify the intended cognitive bias are analysed together. The authors argue that this has led to an increase in null findings, which have artificially lowered the apparent efficacy of cognitive bias modification training. For example, in a reappraisal of the data reported by Cristea et al. (2015), Grafton et al. (2017) found that, out of nine ABM training studies that included a measure of attention bias change pre- and post-training, the six studies that failed to change attention bias had a near-zero effect size on reported anxiety (Hedges' $g$ = -0.01). However, the remainder of the studies that managed to change attention bias reported an overall moderate effect size on reported anxiety ($g$ = 0.6). Further, Clarke et al. (2014) compiled a list of 29 studies measuring both the mental health outcomes of ABM training for anxiety disorders and the change in attention bias from pre- to post-training at the group level. The authors reported that 26 of the 29 studies reviewed were consistent with the idea that when attention bias is successfully modified, an improvement in mental health outcomes is observed, and when attention bias is not successfully modified, there is no such improvement in mental health outcomes. Therefore, it was concluded that the null findings reported by some authors may likely be due to a failure of the specific methodology of the ABM training intervention applied to modify attention bias, rather than a failure of ABM training in any form to alleviate anxiety. As such, Clarke et al. (2014) conclude that though further research into the most effective methodology for ABM training interventions is warranted, interventions that do effectively reduce attention bias do lead to improvements in mental health outcomes.

It is worth noting that the ideas expressed in the above studies regarding the finding that when ABM training actually modifies bias, then anxiety is generally reduced and when ABM training does not modify bias anxiety is not reduced, have been criticised by some researchers (Cristea, 2018; Kruijt & Carlbring, 2018). Specifically, Cristea (2018) states that "identifying the trials in which both bias and outcomes were successfully changed is only possible post hoc, as these are both outcomes measured after randomisation; reverse engineering the connection between the two is subject to confounding. Bias and symptom outcomes are usually measured at the same time points in the trial, thus making it impossible to establish temporal precedence" (p. 247). Cristea (2018) goes on to argue that this increases the risk of reverse causality in our arguments (as symptom change may have caused bias change), as well as the risk of conflating demand effects as evidence for ABM procedure effectiveness (the trials in which bias and/or symptom change occurred successfully may simply be due to bias, including experimenter effects). On balance, the overall argument made by Grafton et al. (2017) and Clarke et al. (2014) that an ABM training that does not modify bias has failed at being an ABM training, and therefore should not be considered an ABM procedure, still makes logical sense. Furthermore, in response to the criticisms raised by Cristea (2018), Parsons (2018) argues that what these studies show is that ABM procedures currently do not robustly modify bias (unsurprising, given the earlier discussed variability among ABM paradigms), and therefore we must develop more effective procedures. However, it is important to acknowledge that the current pattern of results reported does not provide conclusive evidence for this particular interpretation over other possibilities. Overall, this debate perfectly highlights the importance of performing a

registered replication of a seminal ABM training study, in order to ascertain its effectiveness using a rigorous design.

A more encouraging picture arises from a recent meta-analysis by Martinelli et al. (2022), whose study represents the largest review of the efficacy of ABM training interventions for modifying bias in mental health disorders to date. Across 13 studies looking into ABM training in populations of participants with non-specific anxiety, Martinelli et al. (2022) reported an overall moderate effect size ($g$ = 0.53), suggestive that ABM interventions may yet prove to be effective for alleviating attention bias. Importantly, the authors make the distinction of separating out non-specific anxiety from other types of anxiety disorders including SAD and PTSD. Interestingly, Martinelli et al. (2022) reported that while ABM interventions are moderately effective in alleviating attention bias in those with non-specific anxiety, these interventions are far less effective for reducing attention bias in SAD populations ($g$ = 0.25, non-significant), and in the case of PTSD larger decreases in attention bias were observed in control than experimental conditions ($g$ = -0.7). This variation in efficacy will clearly be important in the future targeting of ABM.

A review of the efficacy of ABM interventions in relieving anxiety symptomology by Mogoaşe et al. (2014) found results consistent with Martinelli et al. (2022). They similarly distinguished between sub-groups of anxiety disorders and found that, across 22 studies of ABM training for alleviating the symptoms of anxiety disorders, the overall postintervention effect size was small ($g$ = 0.26). However, the effect size was found to be moderated by the type of anxiety disorder, with an overall moderate effect size reported by studies looking into generalised anxiety ($g$ = 0.61), a relatively small effect size detected in studies of SAD ($g$ = 0.24), and a non-significant effect in other types of anxiety disorders, a category

which included phobias and PTSD ($g$ = 0.16). Therefore, it may be that reviews failing to make the distinction between different types of anxiety disorder may fail to detect stronger effect sizes observed in studies solely targeting general anxiety. Given the findings of Clarke et al. (2014) and Grafton et al. (2017), one may conclude that the lack of experimental evidence of attention bias modifiability, and in turn symptom alleviation, in the context of SAD and PTSD may be due to a failure of the ABM paradigms to modify attention bias, rather than SAD and PTSD populations simply being less responsive to the induced change. Regardless, there is still evidence to suggest that in the case of general anxiety, ABM training may be an effective therapeutic method.

There is a clear need to assess the replicability of previous findings suggesting that ABM training is an effective method of alleviating GAD symptoms, including high worry. Since the publication of the Reproducibility Report: Psychology (Open Science Collaboration, 2015), which estimated a low replication rate of around 36% in psychological literature, it has become clear that it is exceedingly important to establish the replicability of key psychological findings. This replication crisis has been driven largely by the prevalence of false positive results arising from poor research practices, as well as publication bias, among other things (Hu et al., 2016). As introduced by Chambers (2013), registered reports are a new publication format that aim to reduce the likelihood of these issues occurring, and therefore performing a registered replication is a rigorous way of testing the replicability of a seminal ABM training study.

In line with this ethos, the aim of the present research was to conduct a pre-registered replication of a study assessing the influence of an ABM training intervention on symptoms of anxiety. A seminal study by Hazen et al. (2009)

exploring the effect of ABM training on symptoms of anxiety was identified as an ideal target. This study both found a substantial effect of ABM and, as it utilised a high worry student sample, offered greater feasibility for data collection. The original researchers recruited a sample of psychology undergraduates exhibiting high levels of trait worry (as assessed via the Penn State Worry Questionnaire; Meyer et al., 1990) and randomly assigned them to either an experimental ABM training group, or a sham training control group for a 5-session training course. The researchers then assessed the change in indicators of anxiety and depression, as well as the change in attention bias, from pre- to post-training assessment. The researchers found that, based on a composite measure of mental health symptoms including indicators of general anxiety and depression, ABM training led to a significant decrease in reported symptomology. Furthermore, assessment of attention bias found that the ABM intervention did indeed reduce the observed attention bias as predicted. In both cases, no such improvements were observed in the sham training control group. Consistent with the original study, the present replication predicted that the active versus sham ABM training would lead to a reduction in symptoms of general anxiety and depression as measured by a composite symptom index (CSI)[2], and a reduction in attention bias as assessed via a dot-probe based task.

In addition to the originally assessed hypotheses, in response to the concerns raised by Cristea et al. (2015) and Cristea (2018), an additional analysis was conducted to assess the hypothesis that demand effects may be influencing the

[2] A composite mental health measure was used in the interests of fully replicating Hazen et al.'s (2009) methodology. While this is not a 'standard' choice in the literature, the original authors used the CSI in the interests of reducing the familywise error rate (by analysing one CSI measure, as opposed to three individual measures). However, though the CSI is still included, we also report Bayesian analyses on the three individual mental health measures that comprise the CSI, to allow for the separation of depression and anxiety as constructs.

apparent reduction in symptomology. Demand effects occur when participants behave differently, consciously or unconsciously, in response to any aspect of the design or delivery of an experiment that indicates how they are expected to respond. First discovered by Orne (1962), demand effects have proven exceedingly difficult to mitigate (Nichols & Maner, 2008), and are particularly problematic in a training experiment such as this, in which the aim (in this case to reduce anxiety) is inherently clear. Therefore, quantifying the presence of demand effects in the present study is essential, as their presence would suggest that expectancy effects are driving the clinical outcomes of the ABM intervention. This would have implications for the underlying theory (if ABM treatment outcomes are driven by expectancy, then there may in truth not be a link between attention bias and worry/anxiety).

The relationship between a measure of Phenomenological Control (PC) and the CSI was assessed. PC describes the ability to create phenomenological experiences that match the expectancies of a given situation. One such example of a phenomenological experience would be the rubber hand illusion, in which the sight of a rubber hand being brushed at the same time brushing is felt on one's real hand leads participants to report subjective ownership of the rubber hand; this is an example of a visual hallucination leading to a subjectively reported experience that is driven by demand effects (Dienes et al., 2022; Lush et al., 2020). The application of PC to create such experiences can occur in the absence of any conscious intention. As such, where there are implicit or explicit expectations put on an individual's response or experience, those with higher PC will be more likely to exhibit the corresponding outcome; they will have increased susceptibility to demand effects (Lush et al., 2020). This difference in susceptibility can be applied to detect the influence of demand effects in any experimental paradigm with outcome variables

related to experienced phenomenology; where demand effects are present there should be a corresponding correlation between PC and the outcome measures. Note, while the relationship between responses and PC has been shown to provide an indication of demand effects in a variety of experimental paradigms (Lush et al., 2020), a null result for this experimental hypothesis does not rule out the presence of demand effects entirely. The absence of a relationship between outcome measures and PC is also consistent with the presence of demand effects that are unrelated to PC. In contrast, the presence of a relationship can be taken as a strong indication of the influence of demand effects. In the present context, if demand effects contribute to the apparent efficacy of the ABM training, then it would be predicted that participants with higher PC will exhibit greater reductions in mental health symptomology.

*Final Hypotheses*

1) ABM training will lead to a reduction in attention bias.
2) ABM training will lead to a reduction in symptoms of general anxiety and depression.
3) Demand effects may influence ABM training efficacy, contributing to a reduction in mental health symptomology.

## Method

The experiment aimed to directly replicate the procedure reported by Hazen et al. (2009) to examine whether the same ABM training paradigm can be shown to

both reduce attention bias and successfully alleviate symptoms of anxiety and depression. Therefore, all methods were in keeping with those used in the original study. The only exception to this is that Hazen et al. (2009) also conducted the Structured Clinical Interview for Axis I DSM-IV Disorders on a subsection of participants. However, this was not part of the inclusion/exclusion criteria for the study, nor was it used for any of the analyses. Therefore, we did not conduct the interview in this replication. The original frequentist analyses were replaced by Bayesian analyses, to allow for more nuanced interpretation of potentially non-significant results as denoting either evidence for the null hypothesis or indicating insensitive data. These Bayesian analyses were specified in line with the method outlined by Dienes (2021, p. 5-8), in which one first specifies a rough scale of effect, then compares the evidence for a model predicting this effect to a model predicting no effect using a Bayes factor. The final analytical decision as to whether the data supports or refutes the experimental hypotheses was made based on these Bayesian analyses.

*Statistical Hypotheses*

1) $H_0$: ABM training will not lead to a reduction in attention bias from pre- to post-training, as assessed via the change in Probe Discrimination Task score.

   $H_1$: ABM training will lead to a reduction in attention bias from pre- to post-training, as assessed via the change in Probe Discrimination Task score.

2) $H_0$: ABM training will not lead to a reduction in symptoms of general anxiety and depression from pre- to post-training, as measured via change

in both individual measures (PSWQ, STAI-T and BDI) and via change in the CSI.

$H_1$: ABM training will lead to a reduction in symptoms of general anxiety and depression from pre- to post-training, as measured via change in both individual measures (PSWQ, STAI-T and BDI) and via change in the CSI.

3) $H_0$: PC score will have no influence on the change in CSI score from pre- to post-training.

   $H_1$: The higher the PC score, the greater the reduction in CSI score from pre- to post-training.

*Participants*

Participants were students[3] from the University of Sussex gathered via flyers, in-person recruitment and online recruitment adverts. These students were invited to complete a screening questionnaire. Furthermore, recruitment was also conducted via a worry database maintained at the University of Sussex, which contains the PSWQ scores of psychology undergraduates who have indicated they are willing to be contacted for future research. Students with a PSWQ score of 60 or above were contacted via email and invited to complete the screening questionnaire.[4] Of those

[3] Originally, we intended to solely target Psychology undergraduates. However, while conducting the study it became apparent that recruiting solely from this pool would be insufficient to meet our recruitment needs. Therefore, with the approval of our PCI-RR recommender, it was added to our Stage 1 IPA that our sample would be broadened to any University of Sussex student, including undergraduates and postgraduates. This change also necessitated the addition of a cash payment option as reimbursement for the study.

[4] A PSWQ score of 60 or above was used as the cut-off for high worry, in line with Hazen et al. (2009), who reported that scores of 60 or above represent the 90$^{th}$ percentile of normative population values. Furthermore, recent cognitive bias

who expressed an interest, respondents who were currently receiving treatment for an anxiety disorder, not fluent in English or did not have normal or corrected-to-normal vision were excluded. These exclusions were self-declared, as the consent form presented in the screening questionnaire asked participants to confirm that they met the above stated eligibility criteria. Following this, any students who had a PSWQ score of 60 or above on the screening questionnaire were invited to take part in the main study, resulting in an initial sample of 165 participants. Following the removal of data from participants who either met the above stated exclusion criteria, had missing data, whose PSWQ score at the pre-training session had dropped below 60, or who had dropped more than 10% of their pre- or post-training PDT trials (see the 'Scoring' section), the final sample consisted of $N = 104$ participants (control = 52; experimental = 52). Participants were aged between 18 and 41 years ($M = 20.12$, $SD = 3.87$). Of the sample, 86 were female, seven were male, nine were non-binary and two preferred not to specify. Furthermore, 64% were White British or White Irish, 7.7% were Asian or Asian British Indian, 2.9% were Black African or Black Caribbean, 1.9% were Chinese, 1% were Asian or Asian British Pakistani and 25% identified as either mixed race or 'other'. The sample size was determined based on a Bayesian stopping rule, adopting a sequential design with a maximal $n$ of 150[5] (as described by Schönbrodt and Wagenmakers, 2018). Specifically, data collection would continue until we had obtained a sensitive Bayes factor for the change in CSI ($B > 30$, suggesting that ABM training had led to a reduction in mental health

modification research has used a PSWQ score of 56 or above as the cut-off for high worry (see Feng et al., 2019), suggesting that the cut-off chosen here remains valid.

[5] Originally, we were targeting a maximal $n$ of 200 participants. However, given a reduction in student enrolment it became clear that this target was unfeasible. Therefore, with the approval of our PCI-RR recommender, a reduction of the maximal $n$ to 150 participants was added to our Stage 1 IPA. However, we were allowed the flexibility to exceed this number if recruitment allowed.

symptomology, or $B < 1/6$, suggesting that ABM training had not led to a reduction in mental health symptomology), or cease collection at 150 participants if the result remained insensitive. The only exception to this rule was that we would collect a minimum of 40 participants. In using the procedure detailed by Palfi & Dienes (2019, Version 3, p. 15), it was determined that given a long-term relative frequency of good enough evidence of 50%, this sample size would allow for a discriminating Bayes factor ($B > 30$ if H1 was true, and a $B < 1/6$ if H0 was true).[6] Participants were reimbursed in either course credit or £15 cash. In line with the original authors' procedure, any participants who failed to attend all seven sessions were excluded from analysis. This study was granted ethical approval by the Sciences & Technology Cross-Schools Research Ethics Committee (C-REC), University of Sussex (Application Number: ER/NP286/17).

*Materials*

**Penn State Worry Questionnaire (PSWQ).** Designed by Meyer et al. (1990) and with established internal consistency, test-retest reliability and convergent validity (Behar et al., 2003; Brown et al., 1992; Stöber, 1998), the PSWQ is a 16-item scale that gives an indication of trait worry, with higher scores indicating higher trait worry. Each item uses a 5-point Likert scale, ranging from 1 (*not at all typical of me*) to 5 (*very typical of me*), in response to statements regarding worry behaviours (for example, "Many situations make me worry"). The PSWQ is widely used across a number of clinical and research settings and is a robust measure of pathological

[6] This analysis was performed on the sensitivity of the main Bayesian analysis (the change in the CSI score). The analysis revealed that the likely *N* needed for a sensitive Bayes factor was 16 for H1, and 56 for H0. Given a maximal *N* of 150, it was apparent that a sensitive Bayes factor should be achievable.

worry that can also reliably indicate whether or not an individual has GAD (Startup & Erickson, 2006).

**State-Trait Anxiety Inventory (STAI).** Developed by Spielberger et al. (1983) and with established internal consistency, test-retest reliability, convergent validity, and internal validity (Guillen-Riquelme & Buela-Casal, 2011; Oei et al., 1990; Ortuno-Sierra et al., 2016), the STAI is a 40-item measure comprised of two subscales that can be used to measure state anxiety (the temporary feeling of anxiety triggered by a situation perceived as threatening) and trait anxiety (the general disposition to become anxious in situations perceived as threatening), respectively. Higher scores indicate greater levels of anxiety. Each item asks participants how frequently they feel certain things, responding on a 4-point Likert scale ranging from 1 (*almost never*) to 4 (*almost always*), in the case of the trait subscale, and from 1 (*not at all*) to 4 (*very much so*) in the case of the state subscale. The STAI is a psychometrically adequate measure of anxiety, with evidence suggesting that the scale can reliably distinguish between clinical and non-clinical anxiety (Ortuno-Sierra et al., 2016).

**Beck Depression Inventory (BDI-II).** Developed by Beck, Steer and Brown (1996), the BDI-II has established concurrent, content, and structural validity, as well as strong internal consistency and test-retest reliability across a number of settings, as determined by a review of 118 studies by Wang and Gorenstein (2013). The BDI-II is a 21-item measure that assesses key symptoms of depression, with higher scores indicating more severe depression. Each item asks participants to select a statement pertaining to the severity of a given depressive symptom, responding on a 4-point Likert scale ranging from 0 to 3. The BDI-II is a psychometrically strong measure

that can reliably distinguish depressed and non-depressed individuals (Wang & Gorenstein, 2013).

**Probe Discrimination Task (PDT).** Two slightly variant versions of this task were presented, one at the pre-training session and the other at the post-training session. The pre-training session version consisted of 72 trials, of which 50% featured neutral-neutral word pairs, and the other 50% featured threat-neutral word pairs.[7] All word pairs were matched for length and familiarity with emotional valence having been validated[8], and were presented supraliminally and in a randomised order, see supplementary material. Each trial commenced with a fixation cross (specifically, a plus sign) presented in the centre of the screen for 500ms. Subsequent to its disappearance, the word pair appeared on screen for 500ms, presented vertically with one word just above and the other just below the position of the fixation cross. The word pairs were presented on a black screen and written in white letters that were .5cm tall, with the words being separated by 3cm and subtending approximately 2° of visual angle. On threat-neutral trials, the threat word appeared in either position with equal probability. Subsequent to the disappearance

[7] We contacted the Hazen et al. (2009) authors to request the original word pairs used but were unable to obtain them. Therefore, we created the word pairs from the subset of words from MacLeod et al. (2002) as used by the original authors and paired them based on the pairing rules described in the paper.

[8] The threat-neutral word pairs had already been paired based on length and familiarity, and validated regarding their emotional valence, by MacLeod et al. (2002). However, the neutral-neutral word pairs needed to be created from the pool of neutral words present in MacLeod et al. (2002). Familiarity of the neutral words was determined by gathering frequency of use data from the Corpus of Contemporary American English (COCA; https://www.english-corpora.org/coca/), a database that has been used to determine word familiarity in previous cognitive bias research (see Feng et al., 2019). The words were paired based on their closeness in frequency of use. Then, the frequency scores were compiled into two lists (Frequency of Word A; Frequency of Word B) for the neutral-neutral word pairs and the mean frequency of the word lists compared with a bootstrapped t-test. This confirmed that the frequency did not significantly differ between the word lists ($p$ = .875).

of the word pairs, a probe appeared in the location of either the upper or lower word. The probe was either one dot or two dots. Participants were instructed to press one of two buttons to indicate which probe they saw. Following a response, there was a 1000ms pause before the next trial began (see Figure 1 for a visual example of a trial). At the post-training session, the task was almost identical. However, this time the task consisted of 96 trials, of which 36 featured neutral-neutral words, and the other 60 featured threat-neutral words. This difference was implemented in the interests of fully replicating the original paper.[9]

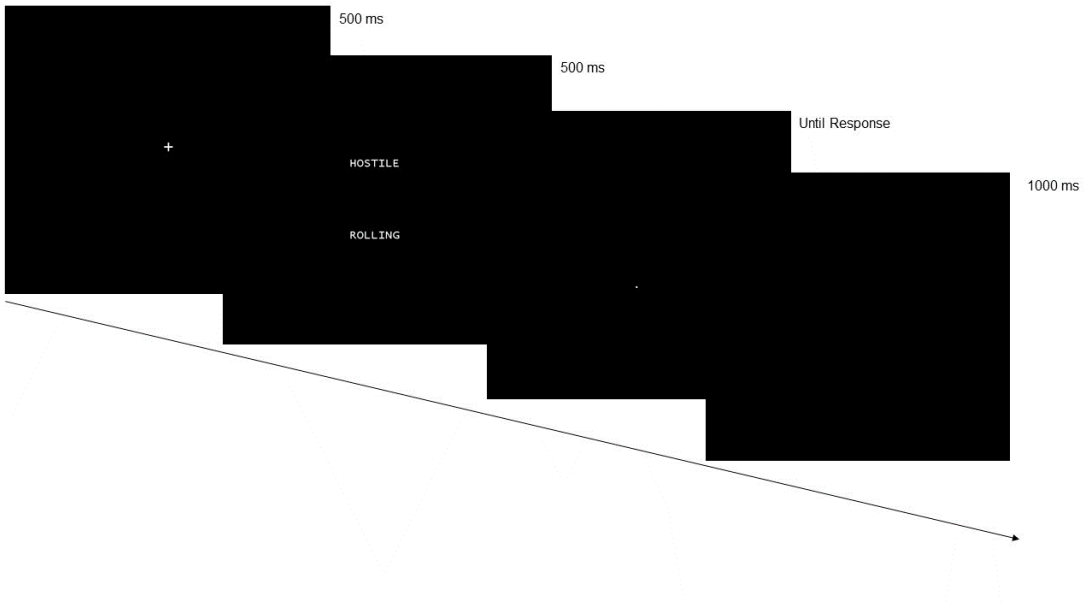


**Figure 1.** A typical PDT trial.

**Attentional Retraining for Threat Stimuli (ARTS).** This is an adapted version of the PDT designed to reduce attention to threatening stimuli that was given to the experimental group during each training session. Each trial was identical to the PDT, except in this version the word pairs were always threat-neutral, and the probe

[9] All code for the PDT tasks and the ARTS/Sham-ARTS training is included in the supplementary material.

almost always appeared behind the neutral word. Each ARTS featured 216 trials, in which the probe appeared behind the neutral word in 204.

**Sham Attentional Retraining for Threat Stimuli (Sham-ARTS).** This is an adapted, placebo version of the ARTS that was administered to the control group during each training session. Sham-ARTS was identical to the ARTS procedure, except in this version the probe appeared behind the threat word and the neutral word with equal frequency, thereby not training a bias to either stimulus type.

**Phenomenological Control (PC) Scale.** Developed by Lush et al. (2021), the PC scale is a measure of one's ability to exercise PC, with higher scores indicating a greater ability. The measure consists of 10-items that capture experiences elicited by different imaginative suggestions and the extent to which they are felt as real, measured on a scale ranging from 0 to 5. This scale has been found to have good internal consistency ($\alpha$ = .68; Lush et al., 2021).

*Procedure*

Potential participants who expressed an interest in the study were emailed a brief screening survey including an information and consent form, and the PSWQ. If they met all eligibility requirements, which included their newly assessed PSWQ score again being 60 or above, participants were invited to the first session. The study involved a total of seven sessions, all conducted by experimenters who were blind to the participants' experimental condition. The first session was a pre-training session, in which the PSWQ, STAI, BDI, and PDT were administered, to assess anxiety symptoms, depressive symptoms and level of attention bias at baseline. Data from any participants who's PSWQ score dropped below 60 since initial screening

were excluded from the study.[10] Following this, sessions two through six were training sessions, lasting approximately 15 minutes each, with participants being randomly assigned to the control or the experimental group. In line with Hazen et al. (2009), we aimed to run two training sessions a week for each participant, schedule dependent. Each training session had participants completing either the ARTS (experimental group) or the Sham-ARTS (control group), with both groups simply being instructed to become as fast as possible at discriminating between the two types of probes without making any mistakes. Finally, the post-training session was conducted one week after completion of the final training session. This session was identical to the pre-training session, except in this session the PDT used different materials, as previously described. Then, they were debriefed. PC scores were not collected during the experimental procedure, as most of the participants already had their PC scores in a PC database maintained by researchers at the University. Therefore, PC scores were taken from the database and linked to the main dataset prior to analysis being conducted. Any participants who were not already on this database were excluded from this particular analysis. The PC scale is an adapted version of the SWASH measure of hypnotisability (Lush et al., 2021). Hypnosis is a stable trait (Piccione et al., 1989), and therefore PC is equally stable.

[10] In the Stage 1 report it was stated that participants would be explicitly told that they would be receiving either an experimental or a placebo condition, and that they would be asked which they thought they had received at the end of the study. Due to an accidental omission, these steps were not conducted in practice. This was reported to the PCI-RR prior to Stage 2 submission. The former deviation is discussed in more detail in the Discussion section, but we briefly expand on the latter here regarding not asking participants which condition they believed they were in. Research suggests that 'suspicion probes' such as this (wherein participants are asked to reveal whether they figured out an element of the experiment) are ineffective, as participants will usually say what they believe the researcher wants to hear (Nichols & Maner, 2008). The PC analysis we report is a far more powerful method of investigating demand effects.

# Results

Data cleaning was conducted in Microsoft Excel (Microsoft Corporation, 2025), and statistical analyses were performed using IBM SPSS Statistics (Version 30; IBM Corp., 2024).[11]

*Scoring*

The PSWQ, STAI-T and BDI from the pre- and post-training sessions were scored for each of the experimental conditions (ARTS; sham-ARTS), and the means and standard deviations for each of the measures were calculated at pre- and post-training for each experimental condition (for a full breakdown, see Table 1). Additionally, a CSI was calculated. This measure, as calculated by Hazen et al. (2009), represents an overall, standardised indicator of general mental health symptomology, inclusive of measures of anxiety and depression. Firstly, the mean PSWQ, STAI-T and BDI scores for each time point were standardized according to normative estimates of general population means and standard deviations. These estimates of population means and standard deviations are those used in Hazen et al.'s (2009) original paper (PSWQ: $M = 45.7$, $SD = 13.5$; STAI-T: $M = 37.96$, $SD = 9.42$; BDI: $M = 7.65$, $SD = 5.9$). Then, the mean of these three standardised scores was calculated for each time point in each of the experimental conditions. Therefore, each CSI score represents the average level of depression and anxiety symptomology in each group relative to levels of symptomology in the general population.

[11] Added at Stage 2 for clarification.

The PDT from the pre- and post-training sessions was scored for each of the experimental conditions. Firstly, in line with Hazen et al.'s (2009) exclusion criteria, RTs from neutral-neutral word pair trials, trials in which participants failed to accurately detect the probe or trials with extreme RT values (RT < 150ms, or RT > 1500ms) were excluded from analysis. If more than 10% of a participant's pre- or post-training PDT trials were excluded due to these criteria, then that participant's data was excluded from analysis ($n$ = 14 participants were excluded). In the PDT task, the position of probes was crossed with threat word position to create four within-subject conditions: threat upper\probe upper (TU\PU); threat upper\probe lower (TU\PL); threat lower\probe upper (TL\PU); threat lower\probe lower (TL\PL). Harmonic means were calculated for each of these four conditions, in line with Hazen et al.'s (2009) method. Then, the four conditions were combined using the following formula to create an attention bias score: [(TU\PL - TU\PU) + (TL\PU - TL\PL)] / 2. This resulted in a final attention bias score in which a positive value represents faster discrimination of probes following threat words as opposed to neutral words (biased attention towards threatening words), and a negative value represents faster discrimination of probes following neutral words as opposed to threat words (reduced attention towards threatening words). Finally, the mean and standard deviation for the PDT were calculated at pre- and post-training for each experimental condition (see Table 1).

**Table 1.**

*Means and SDs for clinical symptomology and attention bias measures by Group and Time*

| Measure | ARTS Group | | Sham-ARTS Group | |
|---|---|---|---|---|
| | Pre-training | Post-training | Pre-training | Post-training |
| | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* |
| CSI | 2.08 (0.75) | 1.80 (0.81) | 2.07 (0.86) | 1.86 (0.93) |
| PSWQ | 67.88 (4.38) | 66.35 (4.99) | 69.02 (5.01) | 66.54 (6.45) |
| z-score | 1.64 (0.32) | 1.53 (0.37) | 1.73 (0.37) | 1.54 (0.48) |
| STAI-T | 57.38 (7.45) | 55.15 (8.57) | 58.23 (8.07) | 56.10 (9.28) |
| z-score | 2.06 (0.79) | 1.83 (0.91) | 2.15 (0.86) | 1.93 (0.98) |
| BDI-II | 22.62 (9.04) | 19.69 (9.64) | 21.37 (9.41) | 20.08 (9.79) |
| z-score | 2.54 (1.53) | 2.04 (1.63) | 2.32 (1.59) | 2.11 (1.66) |
| PDT | 0.14 (27.36) | 0.03 (22.45) | 9.32 (32.96) | -3.91 (21.32) |

*Note.* CSI = Composite Symptom Index, PSWQ = Penn State Worry Questionnaire, STAI-T = State-Trait Anxiety Inventory – Trait Subscale, BDI-II = Beck Depression Inventory II, PDT = Probe Discrimination Task. Z-scores refer to the variant of the relevant outcome measure standardized according to normative estimates of general population means and standard deviations, as derived from Hazen et al. (2009, pp. 630).

*Preliminary Analyses*

Bayesian Credibility Intervals, assuming a uniform prior and normal approximation (and thus equivalent to a 95% confidence interval), were calculated for the difference in pre-training scores between conditions (ARTS vs. Sham-ARTS) for each of the four measures (PSWQ, STAI-T, BDI, PDT; see supplementary material (Table S1) for a full breakdown of the credibility intervals. Doing so

revealed that across all pre-training outcome measures assessed there was an overlap in the credibility intervals for the two conditions, suggesting that pre-training scores did not differ substantially between them. Further, for each experimental condition, Bayesian Credibility Intervals, again assuming a uniform prior and normal approximation, were calculated for the difference between the pre-training scores for the PSWQ and STAI-T and their corresponding general population means (as previously derived by Hazen et al., 2009). This analysis revealed that the pre-training credibility intervals in each condition for the PSWQ and STAI-T did not overlap with the general population estimate of 45.7 and 37.96, respectively. This suggests that at pre-training the present sample was above the normative levels of worry and trait anxiety observed in the general population.

*Main Analyses*

A series of Bayesian t-tests were conducted to assess each hypothesis. For all Bayes Factors we adopted the conventional thresholds of values greater than 3 indicating evidence for the alternate hypothesis and values less than 1/3rd indicating evidence for the null. Robustness regions are reported as: RRconclusion [x1, x2], where x1 is the smallest and x2 is the largest SD that gives the same conclusion: $B < 1/3$; $1/3 < B < 3$; $B > 3$. All Bayes factors were calculated using an online Bayes factor calculator (URL: https://harry-tattan-birch.shinyapps.io/bayes-factor-calculator/). For every Bayes factor, we also report the corresponding t and p-values.

Firstly, the change in attention bias scores was assessed (see Figure 2). A Bayes factor was computed on the difference between groups in their respective change in attention bias scores from pre- to post-training (ARTS pre-post attention bias minus Sham-ARTS pre-post attention bias). $H_1$ was modelled as a half-normal

distribution with a mode of 0 and *SD* of 17.49. This *SD* is the raw effect size originally observed by Hazen et al. (2009; see Table 2, *'Mean Threat Bias Scores by Group'*).[12] The results indicated that there was no evidence of a meaningful difference between conditions in the change in PDT score from pre- to post-training. Though a moderate effect size was observed, this was driven by a change in the control group as opposed to the experimental group, and so this remains sensitive evidence for the null hypothesis, $MDiff$ = -13.12, $SE$ = 7.55, $t(102) = 1.74$, $p = .085$, $B_{HN(0,\ 17.49)} = 0.16$, $RR_{B<1/3}[7.4, \infty]$, $d = 0.34$.[13]

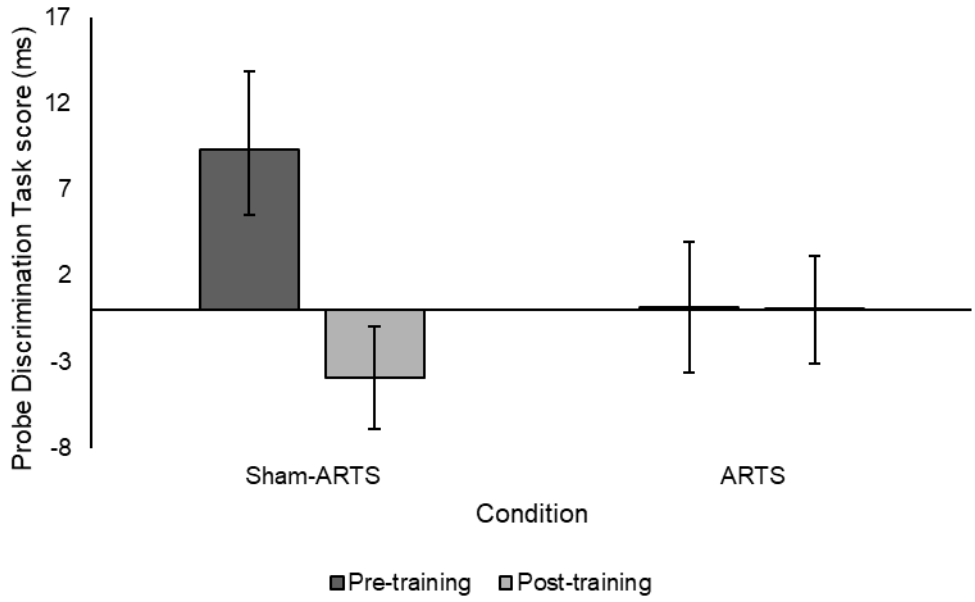


**Figure 2.** Mean Probe Discrimination Task scores (+/- 1 *SE*) at pre- and post-training for each experimental condition (Sham-ARTS; ARTS).

[12] Implementing a half-normal prior for $H_1$ with a *SD* equal to the raw effect size observed by Hazen et al. (2009) reflects our expectation that if $H_1$ is true, we would observe an effect approximately equivalent to what the original authors observed, and an effect larger than this is less likely.

[13] Our Stage 1 report stated directional hypotheses but did not specify alpha or test sidedness. Consistent with standard practice, we therefore conducted two-sided tests with $\alpha = .05$.

Next, the change in CSI score was assessed (see Figure 3). Bayes factors were computed on the difference between groups in their respective change in CSI scores from pre- to post-training (ARTS pre-post CSI minus Sham-ARTS pre-post CSI). $H_1$ was modelled as a half-normal distribution with a mode of 0 and *SD* of 0.86. This *SD* is the raw effect size originally observed by Hazen et al. (2009; see Table 1, *'Means and SDs for composite symptom index and individual measures by Group and Time'*). The results indicated that there was no evidence of a meaningful difference between conditions in the change in CSI score from pre- to post-training, and therefore sensitive evidence for the null hypothesis, $MDiff = 0.07$, $SE = 0.12$, $t(102) = -0.62$, $p = .535$, $B_{HN(0,\ 0.86)} = 0.24$, $RR_{B<1/3}[0.61, \infty]$, $d = 0.12$.

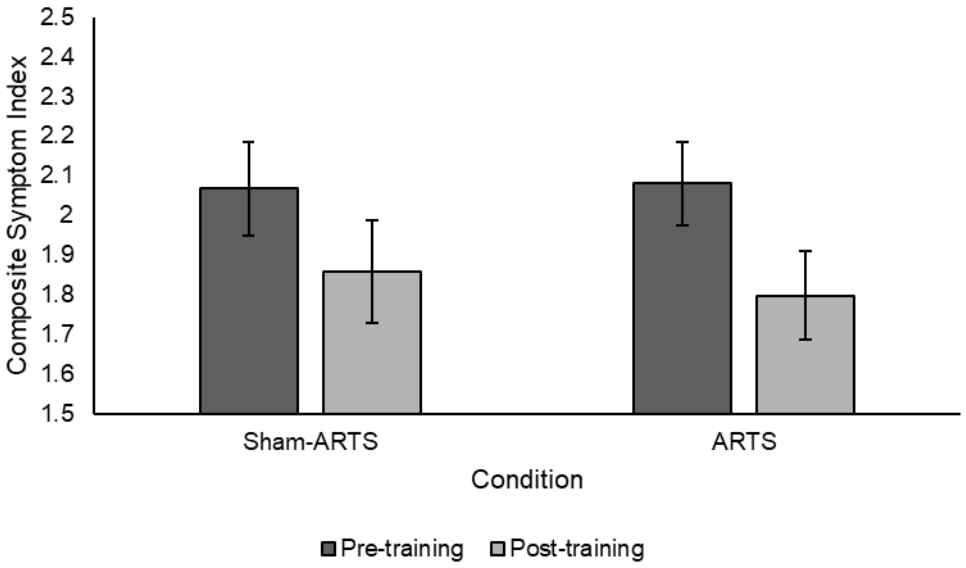


**Figure 3.** Mean Composite Symptom Index scores (+/- 1 *SE*) at pre- and post-training for each experimental condition (Sham-ARTS; ARTS).

In addition to the comparisons conducted by the original authors, Bayes factors were computed on the difference between groups in their respective changes in each of the individual mental health indicators (PSWQ, STAI-T and BDI), pre minus post training as above. All SDs are the raw effect sizes originally observed by Hazen et al. (2009; see Table 1, *'Means and SDs for composite symptom index and individual measures by Group and Time'*). For the mean difference in PSWQ scores, $H_1$ was modelled as a half-normal distribution with a mode of 0 and *SD* of 8.14. The results indicated that there was no evidence of a meaningful difference between conditions in the change in PSWQ score from pre- to post-training, and thus provides sensitive evidence for the null hypothesis, *MDiff* = -0.94, *SE* = 0.98, $t(102) = 0.96$, $p = .338$, $B_{HN(0,\ 8.14)} = 0.07$, $RR_{B<1/3}[1.41, \infty]$, $d = 0.19$. For the mean difference in STAI-T scores, $H_1$ was modelled as a half-normal distribution with a mode of 0 and *SD* of 6.72. The results indicated that there was no evidence of a meaningful difference between conditions in the change in STAI-T score from pre- to post-training, and therefore sensitive evidence for the null hypothesis, *MDiff* = 0.1, *SE* = 1.25, $t(102) = -0.08$, $p = .939$, $B_{HN(0,\ 6.72)} = 0.19$, $RR_{B<1/3}[3.78, \infty]$, $d = 0.02$. For the mean difference in BDI scores, $H_1$ was modelled as a half-normal distribution with mode of 0 and *SD* of 7.25. The results were insensitive, and therefore no meaningful conclusion can be drawn regarding the difference between conditions in the change in BDI score from pre- to post-training, *MDiff* = 1.63, *SE* = 1.29, $t(102) = -1.27$, $p = .208$, $B_{HN(0,\ 7.25)} = 0.68$, $RR_{1/3<B<3}[0, 15.2]$, $d = 0.25$.

Further to these, an additional analysis was conducted to assess the potential impact of demand characteristics on the outcome of the intervention on symptomology. Specifically, we tested the prediction that higher PC scores would lead to greater reductions in the CSI in the ARTS vs. the Sham-ARTS condition. This

prediction was derived from the hypothesis that individuals with higher PC scores may (consciously or unconsciously) utilise PC to experience the expected change in symptomology, thus showing a greater response to demand characteristics (Lush et al., 2020). As such, the extent to which this hypothesised relationship was observed would be an indication of the degree to which demand characteristics were influencing the results; as would occur if the expectations were greater in the experimental versus the sham condition. This was assessed by examining the interaction between Condition (ARTS vs. Sham-ARTS) and PC in a multiple regression predicting change in CSI. If the interaction term was greater than zero, this would indicate that PC was influencing the apparent efficacy of ABM. The difference from zero was evaluated by applying a Bayes Factor to the interaction term. To the authors' knowledge, no research has yet investigated the relationship between PC and the efficacy of ABM. Therefore, a theory of the relationship based on the ratio-of-means heuristic (Dienes, 2019) was used to model $H_1$ for the Bayes Factor. If we theorise that PC is required for a change in CSI, then there can be no change in CSI without using PC and thus both scales will approach zero together. This is an idealised theory in which change in CSI cannot occur independently of PC. However, if we theorise that PC is one important contributory factor leading to a change in CSI, then the ratio represents the maximum interaction effect one may expect to observe. Therefore, the ratio of the mean CSI change score and the mean PC score was used as an estimate of the maximum slope. Specifically, $H_1$ was modelled as a half-normal distribution with a mode of 0 and *SD* = (Mean CSI Change score / Mean PC score)/2. Of the sample, 77.9% (*n* =81) had their PC scores available on the PC database and were thus able to be included in this analysis. The results indicated that the interaction term was insensitive, meaning we cannot make a

meaningful conclusion regarding whether or not higher PC scores lead to greater reductions in the CSI in the ARTS vs. the Sham-ARTS condition, $b$ = -0.02, $SE$ = 0.19, $t(77)$ = -0.08, $p$ = .934, $B_{HN(0,\ 0.05)}$ = 0.95, $RR_{1/3<B<3}[0, 0.49]$, $\beta$ = -0.03.

*Exploratory Analyses*

A series of simple linear regressions were run to assess the relationship between PC scores and each of the outcome measures, to assess whether PC can predict the change in mental health symptoms in response to an intervention regardless of condition assignment. This is to evaluate whether expectations may influence response to a therapeutic intervention. Firstly, a simple regression predicting the influence of PC on the change in CSI scores. The effect of PC was significant, $b$ = 0.22, $SE$ = 0.09, $t(78)$ = 2.35, $p$ = .021, $\beta$ = 0.26, suggesting that higher PC scores resulted in a greater reduction in CSI scores.

Secondly, a simple regression predicting the influence of PC on the change in the PDT measure. The effect of PC was non-significant, $b$ = -5.46, $SE$ = 6.8, $t(78)$ = -0.8, $p$ = .425, $\beta$ = -0.09, suggesting that higher PC scores may not result in a reduction in the PDT measure.

Finally, a series of simple regressions predicting the influence of PC on the change in each of the individual measures of mental health symptomology (PSWQ; BDI-II; STAI-T) were run. The effect of PC on the change in STAI-T scores was significant, $b$ = 2.39, $SE$ = 1.08, $t(78)$ = 2.22, $p$ = .029, $\beta$ = 0.24, suggesting that higher PC scores resulted in a greater reduction in STAI-T scores. However the effects of PC on the change in PSWQ, $b$ = 1.25, $SE$ = 0.85, $t(78)$ = 1.48, $p$ = .144, $\beta$ = 0.16, and BDI-II, $b$ = 1.86, $SE$ = 1.09, $t(78)$ = 1.7, $p$ = .092, $\beta$ = 0.19, scores were

non-significant, suggesting that higher PC scores may not result in a greater reduction in PSWQ or BDI-II scores.

## Discussion

The present findings revealed that there was sensitive evidence for the null hypothesis regarding the change in the PDT measure from pre- to post-training, which would suggest that multi-session ABM training did not lead to a reduction in attention bias. Furthermore, the present findings revealed sensitive evidence for the null hypothesis regarding the change in the CSI measure from pre- to post-training, meaning that multi-session ABM training did not lead to an overall reduction in symptoms of general anxiety and depression. When individual analyses were run on the measures that comprised the CSI, it was revealed that there was sensitive evidence for the null hypothesis for the change in the PSWQ and the STAI-T from pre- to post-training. This suggests that ABM training did not lead to a reduction in levels of trait worry or trait anxiety. However, the evidence remained insensitive for the change in the BDI measure, meaning that it remains unclear whether ABM training led to a change in depression. In addition to this, to assess the impact of potential demand effects, a multiple regression evaluating the influence of the PC X Condition interaction effect on the change in CSI scores from pre- to post-training was conducted. This revealed an insensitive interaction term, meaning that it remains unclear whether higher PC scores lead to greater reductions in the CSI in the ARTS condition compared to the Sham-ARTS control condition. Therefore, we cannot make a robust conclusion regarding the potential influence of demand effects on the efficacy of the ABM procedure. Finally, a series of exploratory regression analyses

on the effect of PC scores on each of the outcome measures (independent of experimental condition) revealed that higher PC scores significantly predicted a greater reduction in CSI and STAI-T scores, suggesting that individual differences in the ability to exercise PC may predict the efficacy of the procedure in reducing anxiety symptoms. This effect remained non-significant for the remaining outcome measures (PSWQ; BDI; PDT), suggesting that PC may not predict change in levels of worry, depression or attention bias.

In the present study, multi-session, probe-based ABM training was not shown to be effective in alleviating attention bias, or anxiety symptomology. Our findings did not replicate those of Hazen et al. (2009), but are in line with previous literature suggesting that ABM procedures may not lead to a change in anxious symptomology, or bias (Cristea et al., 2015; Hallion & Ruscio, 2011). While it is worth noting that a non-significant shift in attention bias was observed in the control condition that was not present in the experimental condition, we believe this is due to experimental noise. Research has shown that sham-training control conditions may have some efficacy beyond placebo training and thus act as 'low dose' ABM training (Blackwell et al., 2017; Gladwin et al., 2019; Tiggemann & Kemps, 2020). For example, presenting a probe with a 50-50 contingency in the location of the threat or neutral stimulus may facilitate greater attentional flexibility. However, were this a true effect, it seems implausible that a 'high dose' experimental condition, in which participants were encouraged to attend maximally to the neutral stimulus, would have no effect. Given that the current procedure did not modify attention bias as intended in the experimental group, these results align with previous assertions from Grafton et al. (2017) and Clarke et al. (2014), suggesting that when an ABM procedure fails to modify bias, naturally one would not expect a change in symptoms

to follow. In a state of the science review, Vrijsen et al. (2024) note that one of the main explanations for the inconsistent efficacy of ABM procedures is the unreliability with which current paradigms result in the intended alteration of attention bias. Therefore, it may simply be the case that probe-based ABM procedures are not effective in alleviating attention bias and thus alternative paradigms for bias modification need to be explored.

In support of this assertion, a recent meta-analysis of the efficacy of differing ABM paradigms by Rooney et al. (2024) found that the only task type resulting in a significant change in clinical symptom outcomes were gaze-contingent tasks, with a medium effect size ($g = 0.45$). Gaze-contingent tasks involve the use of eye trackers and include any task where participants use their gaze to selectively engage with positive or non-threatening stimuli (Rooney et al., 2024). For example, participants may be presented with a positive-negative stimuli pair (or a 4x4 stimulus array), and active engagement with the positive stimulus would either be necessary to progress to the next trial or be positively reinforced with music. For all other ABM paradigms, the resultant change in clinical symptoms were small and non-significant. Additionally, the largest effect sizes in terms of bias modification were for tasks which trained participants to engage with non-threatening stimuli, which includes gaze-contingent paradigms. Taken together, these results suggests that ABM tasks promoting an active shift in gaze to engage with non-threatening stimuli result in the highest treatment efficacy.

The nuance regarding the effectiveness of targeting attentional engagement is important in understanding the lack of efficacy of the probe-based ABM paradigm. Clarke et al. (2013) note that attention bias can take the form of engagement, wherein threatening stimuli are quickly selected and preferentially processed, or

delayed disengagement, wherein the ability to avert attention from a threatening stimulus is inhibited. Interestingly, Rooney et al. (2024) found that ABM paradigms targeting engagement bias for modification resulted in much larger bias change effect sizes than tasks targeting delayed disengagement, suggesting that ABM paradigms should focus on the modification of engagement bias. This engagement bias occurs very early during cognitive processing, with a review of event-related potential (ERP) research by Clauss et al. (2022) reporting that the N2pc component (thought to be associated with attentional engagement of a threatening stimulus) can be detected ~170ms after the onset of a threatening stimulus. If bias is occurring at this early stage of processing, it would seem unlikely that a probe-based task, in which the target typically appears 500ms following stimulus onset, will be able to effectively modify this engagement bias. It is known that covert attentional shifts can occur in as little as 50-100ms (Muller & Rabbit, 1989), meaning that the initial engagement with threat may likely have passed and thus the training contingency is being applied too late into cognitive processing. Additionally, the most effective ABM procedures are those that encourage active engagement with positive stimuli (Rooney et al., 2024), and while gaze-contingent paradigms achieve this by rewarding engagement with positive stimuli, probe-based tasks offer no such reward. In these tasks, it is simply assumed that participants will implicitly learn that reassigning attention to the non-threatening stimulus equates to success on the task, even though they are given no feedback on their performance. Although there is some reinforcement in that reallocating attention to non-threatening stimuli results in one's attention being in the right location to identify the target probe, this may not be a strong enough contingency to reliably change patterns of attention. While this contingency may offer greater control over demand effects relative to other

paradigms that make use of more explicit reinforcement, this also results in a less tangible reward for engaging with neutral stimuli than a gaze-contingent paradigm can offer; the next trial will occur regardless of whether engagement with the neutral stimulus has occurred. Therefore, probe-based tasks may not provide a strong enough reinforcement contingency to reliably change patterns of attention, perhaps explaining why probe-based tasks are not an effective ABM method.

Given the lack of efficacy displayed by the probe-based ABM procedure to modify bias or anxiety symptomology, this joins a growing body of evidence to suggest that probe-based paradigms are not appropriate for clinical implementation as a treatment for anxiety. It is known that currently available ABM training procedures do not reliably modify attention bias as intended (Vrijsen et al. 2024). There is however some evidence to suggest that when bias changes, symptoms do change (Clarke et al., 2014). In the format used in the present study, there is no evidence to suggest ABM training has clinical utility. However, there may be merit in pursuing work to develop and assess the effectiveness of alternative ABM training paradigms. As aforementioned, gaze-contingent paradigms may be a promising alternative with robust treatment outcomes reported so far (Rooney et al., 2024) and could potentially be translated to a clinical setting. However, these methods cannot be translated into a remote task for service users to participate in at home, so further work may be needed to identify tasks that successfully result in active engagement with non-threatening stimuli without the requirement of an eye tracker.

With regards to concerns about the potential for demand effects to be influencing the reported efficacy of ABM procedures (e.g. Cristea et al., 2015), the present study was the first to empirically assess their potential impact. It emerges that further research is needed. The lack of sensitive evidence to indicate a greater

influence of expectancies on response to training in the experimental group compared to the control group, alongside a negligible effect size of the interaction term, is promising. This would suggest that demand effects were not likely to be having a substantially differential effect in the experimental group compared to the control group, meaning that demand effects were likely controlled for well in the paradigm. However, given the insensitive nature of this analysis, this conclusion is not robust. While probe-based ABM procedures that make use of an active sham-control condition, as well as an implicit learning contingency as opposed to explicit instructions, may control for demand effects better than other ABM paradigms (cf. Clarke et al., 2014), the potential influence of demand effects can still not be ruled out. Given that participants elect to attend an experimental training regime for worry, the goal of the training (i.e. to produce a reduction in worry) is inherently clear. Any potential weaknesses in the sham-training control condition are therefore extremely consequential; it is quite feasible that participants could notice the proportion of times the probe aligns with the threatening or neutral stimulus, thus unblinding them to their condition and influencing expectations. The potential influence of these demand effects on the results of ABM procedures cannot be evaluated without using a measure such as PC or an alternate measure of expectations. Therefore, further research is needed before the potential for demand effects to be underlying ABM treatment effects can more certainly be dismissed.

Briefly, it bears mentioning that the accidental omission wherein participants were not explicitly informed that they would be assigned to an experimental or control condition (see Footnote 10) may have slightly decreased demand effects in this replication. When participants are made aware that there is an experimental and a control condition, some may reasonably attempt to determine which condition they

are in. This could be achieved by attending to the frequency with which the probe appears in the location of the neutral vs. threatening stimulus during the ARTS training. Having made a determination, the participant might reasonably be expected to be influenced by it, for example if they come to believe they are in the experimental condition they may exhibit a greater effect on the subsequent measures. We hesitate to report this is a 'limitation', as in terms of detecting a real experimental effect for the change in bias and mental health symptomology this slight deviation is beneficial. Nonetheless, it bears acknowledging that the demand effects may be lower in this replication than in Hazen et al.'s (2009) original study.

Our exploratory findings showed that individuals with higher PC scores may experience a greater shift in levels of trait anxiety in response to an ABM training intervention. This is in line with our assumption that when expectations may influence participant responses, we observe a positive relationship with PC. Research (cf. Dienes & Lush, 2023) has shown that individual differences in trait PC may predict the extent to which participants respond to experimental paradigms where the expectation of what they will experience is clear (Lush et al., 2020; 2021; 2022). Given that participants signed up to an experimental training intervention for worry, the expectation (i.e. that the training will lead to a reduction in anxiety symptomology) will inevitably be clear and thus, as we have demonstrated, participants with higher trait PC may experience increased efficacy of the procedure (regardless of condition assignment) in alleviating symptoms.[14] Given that expectations arose from attending an experimental computerised training for worry,

[14] The relationship between PC and expectations is well established (cf. Dienes & Lush, 2023) and hence provides the most likely basis for the observed change. However, it would be necessary to directly measure expectations to definitively establish this explanation versus one based on an unknown relationship between PC and, for example, spontaneous recovery.

it is likely that expectations present in other clinical settings (for example, when attending talking therapy) are far greater. This could mean that individual differences in trait PC may predict service user responses to psychological treatments. Therefore, further research is warranted to investigate the potential influence of trait PC on therapeutic outcomes from a variety of treatment approaches, not just those in the realm of ABM.

In light of the present findings, future research should move away from probe-based ABM paradigms to focus on the efficacy of alternative tasks, which may address the issue of the inconsistent ability of current ABM paradigms to modify attention bias (Vrijsen et al., 2024). For example, while gaze-contingent paradigms show much promise, the research to date has focussed on a range of clinical populations, with only two of the studies reviewed by Rooney et al. (2024) focussing on an anxious sample (Lazarov et al., 2017; Price et al., 2016). Therefore, more research into the efficacy of gaze-contingent ABM paradigms in anxious populations is warranted. In addition to this, it is notable that there has been great heterogeneity in the parameters of previously implemented ABM tasks, including variations in stimulus presentation time, stimulus type (e.g. words, pictures), as well as the number of trials and sessions implemented. Therefore, future research systematically assessing the influence of varying different task parameters on ABM training effectiveness would be greatly beneficial. Furthermore, further research into whether PC scores predict the efficacy of the treatment contingency of different ABM paradigms should be conducted, to assess the potential influence of demand effects on symptomology outcomes. Finally, as aforementioned, future research should assess whether trait PC may predict responses to psychological treatment interventions beyond the realm of ABM training.

The present study was subject to some limitations. Firstly, given that the present study was a replication of a seminal ABM study, the word pair stimuli used may now be considered low in ecological validity (MacLeod et al., 2019), as they are less informationally rich than emotional information encountered in real life. It is possible that an ABM procedure making use of more ecologically valid stimuli may have produced different results. For example, Mazidi et al. (2025) have recently developed the 'Talking Heads', a pair of AI generated heads that discuss a topic to the participant, one in a negative and the other in a neutral manner. These stimuli are much closer visual match to the way in which potentially threatening information may be transmitted in a more normative social context. Future research would benefit from making use of these more realistic and informationally rich stimuli. Secondly, due to reductions in the available participant pool, the original maximal *n* planned for data collection had to be slightly reduced. While sensitivity was achieved for the crucial analysis conducted in the original study, namely the effect on the composite measure, and for two of the three individual measures, analyses of the change in depressive symptoms and demand effects remained insensitive. Finally, while the absence of a sensitive relationship between PC and the change in CSI scores is promising regarding the adequate control of demand effects in the present ABM paradigm, this still cannot rule out the presence of demand effects that are unrelated to PC.

In conclusion, the present study failed to replicate the findings of Hazen et al. (2009), with results showing that multi-session, probe-based ABM training did not lead to a reduction in levels of attention bias, or anxiety symptomology. The present findings join others in providing no support for the use of probe-based ABM

procedures in the treatment of anxiety disorders. Therefore, our attention should turn to alternative ABM paradigms, such as gaze-contingent tasks.

## References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Amir, N., Beard, C., Burns, M., Bomyea, J. (2009). Attention modification program in individuals with generalized anxiety disorder. *Journal of Abnormal Psychology, 118*(1), 28-33. https://doi.org/10.1037/a0012589

Beard, C., & Peckham, A. D. (2020). Interpretation bias modification. In J. S. Abramowitz & S. M. Blakey (Eds.), *Clinical handbook of fear and anxiety: Maintenance processes and treatment mechanisms* (pp. 359–377). American Psychological Association. https://doi.org/10.1037/0000150-020

Beck, A. T., Steer, R. A,, & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation

Behar, E., Alcaine, O., Zuellig, A. R., & Borkovec, T. D. (2003). Screening for generalized anxiety disorder using the Penn State Worry Questionnaire: A receiver operating characteristic analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 34*(1), 25–43. https://doi.org/10.1016/ S0005-7916(03)00004-1

Blackwell, S. E., Woud, M. L., & MacLeod, C. (2017). A question of control? Examining the role of control conditions in experimental psychopathology using the example of cognitive bias modification research. *The Spanish Journal of Psychology, 20,* E54. https://doi.org/10.1017/sjp.2017.41

Brown, T. A., Antony, M. M., & Barlow, D. H. (1992). Psychometric properties of the Penn State Worry Questionnaire in a clinical anxiety disorders sample.

*Behavior Research and Therapy, 30*(1), 33–37. https://doi.org/10.1016/0005-7967(92)90093-V

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609-610. https://doi.org/10.1016/j.cortex.2012.12.016

Clarke, P. J. F., MacLeod, C., & Guastella, A. J. (2013). Assessing the role of spatial engagement and disengagement of attention in anxiety-linked attentional bias: a critique of current paradigms and suggestions for future research directions. *Anxiety, Stress, & Coping, 26*(1), 1–19. https://doi.org/10.1080/10615806.2011.638054

Clarke, P. J., Notebaert, L., & MacLeod, C. (2014). Absence of evidence or evidence of absence: reflecting on therapeutic implementations of attentional bias modification. *BMC psychiatry, 14*(1), 1-6. https://doi.org/10.1186/1471-244X-14-8

Clauss, K., Gorday, J. Y., & Bardeen, J. R. (2022). Eye tracking evidence of threat-related attentional bias in anxiety-and fear-related disorders: A systematic review and meta-analysis. *Clinical psychology review, 93,* 102142. https://doi.org/10.1016/j.cpr.2022.102142

Cristea, I. (2018). Author's reply. *The British Journal of Psychiatry, 212*(4), 247-247. doi:10.1192/bjp.2018.43

Cristea, I. A., Mogoașe, C., David, D., & Cuijpers, P. (2015). Practitioner review: Cognitive bias modification for mental health problems in children and adolescents: A meta‐analysis. *Journal of Child Psychology and Psychiatry, 56*(7), 723-734. https://doi.org/10.1111/jcpp.12383

Dienes, Z. (2019). How do I know what my theory predicts?. *Advances in Methods and Practices in Psychological Science*, *2*(4), 364-377. https://doi.org/10.1177/2515245919876960

Dienes, Z. (2021). Obtaining evidence for no effect. *Collabra: Psychology*, *7*(1), 28202. https://doi.org/10.1525/collabra.28202

Dienes, Z., & Lush, P. (2023). The role of phenomenological control in experience. *Current directions in psychological science*, *32*(2), 145-151. https://doi.org/10.1177/09637214221150521

Dienes, Z., Lush, P., Palfi, B., Roseboom, W., Scott, R., Parris, B., . . . Lovell, M. (2022). Phenomenological control as cold control. *Psychology of Consciousness: Theory, Research, and Practice, 9*(2), 101-116. doi:https://doi.org/10.1037/cns0000230

Feng, Y. C., Krahe, C., Sumich, A., Meeten, F., Lau, J. Y., & Hirsch, C. R. (2019). Using event-related potential and behavioural evidence to understand interpretation bias in relation to worry. *Biological Psychology*, *148*, 107746. https://doi.org/10.1016/j.biopsycho.2019.107746

Gladwin, T. E., Möbius, M., & Becker, E. S. (2019). Predictive attentional bias modification induces stimulus-evoked attentional bias for threat. *Europe's Journal of Psychology, 15*(3), 479-490. https://doi.org/10.5964/ejop.v15i3.1633

Goodwin, H., Yiend, J., & Hirsch, C.R. (2017). Generalized Anxiety Disorder, worry and attention to threat: A systematic review. *Clinical Psychology Review, 54,* 107-122. https://doi.org/10.1016/j.cpr.2017.03.006

Grafton, B., MacLeod, C., Rudaizky, D., Holmes, E. A., Salemink, E., Fox, E., & Notebaert, L. (2017). Confusing procedures with process when appraising the

impact of cognitive bias modification on emotional vulnerability. *The British Journal of Psychiatry, 211*(5), 266-271. doi:10.1192/bjp.bp.115.176123

Guillen-Riquelme, A., & Buela-Casal, G. (2011). Psychometric revision and differential item functioning in the State Trait Anxiety Inventory (STAI). *Psicothema*, *23*(3), 510-515.

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*(11), 982-990. https://doi.org/10.1016/j.biopsych.2010.07.021

Hallion, L. S., & Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychological bulletin, 137*(6), 940-958. https://doi.org/10.1037/a0024355

Hazen, R.A., Vasey, M.W., & Schmidt, N.B. (2009). Attentional retraining: A randomized clinical trial for pathological worry. *Journal of psychiatric research, 43*(6), 627-633. https://doi.org/10.1016/j.jpsychires.2008.07.004

Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Hirsch, C. R., & Mathews, A. (2012). A cognitive model of pathological worry. *Behaviour research and therapy, 50*(10), 636-646. https://doi.org/10.1016/j.brat.2012.06.007

Hirsch, C. R., Clark, D. M., & Mathews, A. (2006). Imagery and interpretations in social phobia: Support for the combined cognitive biases hypothesis. *Behavior therapy*, *37*(3), 223-236. https://doi.org/10.1016/j.beth.2006.02.001

Hirsch, C. R., Krahé, C., Whyte, J., Bridge, L., Loizou, S., Norton, S., & Mathews, A. (2020). Effects of modifying interpretation bias on transdiagnostic repetitive negative thinking. *Journal of consulting and clinical psychology*, *88*(3), 226. https://doi.org/10.1037/ccp0000455

Hirsch, C. R., Krahé, C., Whyte, J., Loizou, S., Bridge, L., Norton, S., & Mathews, A. (2018). Interpretation training to target repetitive negative thinking in generalized anxiety disorder and depression. *Journal of consulting and clinical psychology, 86*(12), 1017. https://doi.org/10.1037/ccp0000310

Hirsch, C.R., Krahé, C., Whyte, J., Krzyzanowski, H., Meeten, F., Norton, S., & Mathews, A. (2021). Internet-Delivered Interpretation Training Reduces Worry and Anxiety in Individuals With Generalized Anxiety Disorder. *Journal of Consulting and Clinical Psychology, 89*(7), 575-589. https://doi.org/10.1037/ccp0000660

Hu, C., Wang, F., Guo, J., Song, M., Sui, J., & Peng, K. (2016). The replication crisis in psychological research. *Advances in Psychological Science*, *24*(9), 1504. doi: 10.3724/SP.J.1042.2016.01504

IBM Corp. (2024). IBM SPSS Statistics for Windows (Version 30.0) [Computer software]. IBM Corp.

Ji, J. L., Baee, S., Zhang, D., Calicho-Mamani, C. P., Meyer, M. J., Funk, D., ... & Teachman, B. A. (2021). Multi-session online interpretation bias training for anxiety in a community sample. *Behaviour research and therapy*, *142*, 103864. https://doi.org/10.1016/j.brat.2021.103864

Jia, R., Ayling, K., Chalder, T., et al. (2020). Mental health in the UK during the COVID-19 pandemic. *BMJ Open*;10:e040620. doi:10.1136/bmjopen-2020-040620

Kruijt, A., & Carlbring, P. (2018). Processing confusing procedures in the recent re-analysis of a cognitive bias modification meta-analysis. *The British Journal of Psychiatry, 212*(4), 246-246. doi:10.1192/bjp.2018.41

Lazarov, A., Pine, D. S., & Bar-Haim, Y. (2017). Gaze-contingent music reward therapy for social anxiety disorder: A randomized controlled trial. *American Journal of Psychiatry, 174*(7), 649-656. https://doi.org/10.1176/appi.ajp.2016.16080894

Linetzky, M., Pergamin‐Hight, L., Pine, D. S., & Bar‐Haim, Y. (2015). Quantitative evaluation of the clinical efficacy of attention bias modification treatment for anxiety disorders. *Depression and anxiety, 32*(6), 383-391. https://doi.org/10.1002/da.22344

Liu, H., Li, X., Han, B., & Liu, X. (2017). Effects of cognitive bias modification on social anxiety: A meta-analysis. *PloS one*, *12*(4), e0175107. https://doi.org/10.1371/journal.pone.0175107

Loerinc, A.G. et al. (2015). Response rates for CBT for anxiety. *Clinical Psychology Review, 42,* 72–82. https://doi.org/10.1016/j.cpr.2015.08.004

Lush, P, Botan, V, Scott, R B, Seth, A K, Ward, J and Dienes, Z. (2020). Trait phenomenological control predicts experience of mirror synaesthesia and the rubber hand illusion. *Nature*, 11. a4853 1-10. ISSN 0028-0836 https://doi.org/10.1038/s41467-020-18591-6

Lush, P., Scott, R. B., Seth, A. K., & Dienes, Z. (2021). The phenomenological control scale: Measuring the capacity for creating illusory nonvolition, hallucination and delusion. *Collabra: Psychology*, *7*(1), 29542. https://doi.org/10.1525/collabra.29542

Lush, P., Seth, A., Dienes, Z., & Scott, R. B. (2022, April 22). Trait phenomenological control in top-down and bottom-up effects: ASMR, Visually Evoked Auditory Response and the Müller-Lyer illusion. https://doi.org/10.31234/osf.io/hw4y9

MacLeod, C., Grafton, B., & Notebaert, L. (2019). Anxiety-Linked Attentional Bias: Is It Reliable? Annual Review of Clinical Psychology, 15, 529–554. https://doi.org/10.1146/annurev-clinpsy-050718-095505

MacLeod, C., Rutherford, E., Campbell, L., Ebsworthy, G., & Holker, L. (2002). Selective Attention and Emotional Vulnerability: Assessing the Causal Basis of Their Association Through the Experimental Manipulation of Attentional Bias. *Journal of Abnormal Psychology (1965)*, *111*(1), 107–123. https://doi.org/10.1037/0021-843X.111.1.107

Martinelli, A., Grüll, J., & Baum, C. (2022). Attention and interpretation cognitive bias change: A systematic review and meta-analysis of bias modification paradigms. *Behaviour Research and Therapy,* 104180. https://doi.org/10.1016/j.brat.2022.104180

Mazidi, MacLeod, C., Ranjbar, S., Myles, O., & Grafton, B. (2025, April 7). The Talking Heads Attentional Bias Assessment Task: A Readily Available, Reliable, and Effective Task for Assessing Attentional Bias. https://doi.org/10.31219/osf.io/c2vup_v1

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour research and therapy, 28*(6), 487-495. https://doi.org/10.1016/0005-7967(90)90135-6

Microsoft Corporation. (2025). Microsoft Excel (Version 2507). [Computer software].

Mogg, K., & Bradley, B. P. (2005). Attentional bias in generalized anxiety disorder versus depressive disorder. *Cognitive therapy and research*, *29*, 29-45. https://doi.org/10.1007/s10608-005-1646-y

Mogg, K., Waters, A.M., & Bradley, B.P. (2017). Attention bias modification (ABM): Review of effects of multisession ABM training on anxiety. *Clinical Psychological Science, 5*(4), 698-717. https://doi.org/10.1177/2167702617696359

Mogoaşe, C., David, D., & Koster, E. H. (2014). Clinical efficacy of attentional bias modification procedures: An updated meta‐analysis. *Journal of Clinical Psychology, 70*(12), 1133-1157. https://doi.org/10.1002/jclp.22081

Müller, H. J., & Rabbitt, P. M. (1989). Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. *Journal of Experimental psychology: Human perception and performance, 15*(2), 315-330. https://doi.org/10.1037/0096-1523.15.2.315

Nichols, A. L., & Maner, J. K. (2008). The Good-Subject Effect: Investigating Participant Demand Characteristics. *The Journal of General Psychology*, *135*(2), 151–166. https://doi.org/10.3200/GENP.135.2.151-166

Oei, T. P., Evans, L., & Crook, G. M. (1990). Utility and validity of the STAI with anxiety disorder patients. *British Journal of Clinical Psychology, 29*(4), 429-432. https://doi.org/10.1111/j.2044-8260.1990.tb00906.x

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am. Psychol. 17,* 776–783. https://doi.org/10.1037/h0043424

Ortuno-Sierra, J., Garcia-Velasco, L., Inchausti, F., Debbane, M., & Fonseca-Pedrero, E. (2016). New approaches on the study of the psychometric properties of the STAI. *Actas espanolas de psiquiatria*, *44*(3), 83-92.

Palfi, B., & Dienes, Z. (2019). The role of Bayes factors in testing interactions. https://doi.org/10.31234/osf.io/qjrg4

Parsons, S. (2018). *Moving forward with questions of process and procedure in cognitive bias modification research: Three points of consideration.* OSF preprint. https://dx.doi.org/10.31234/osf.io/k3vxc

Piccione, C., Hilgard, E. R., & Zimbardo, P. G. (1989). On the degree of stability of measured hypnotizability over a 25-year period. *Journal of Personality and Social Psychology, 56*(2), 289–295. https://doi.org/10.1037/0022-3514.56.2.289

Price, R. B., Greven, I. M., Siegle, G. J., Koster, E. H., & De Raedt, R. (2016). A novel attention training paradigm based on operant conditioning of eye gaze: Preliminary findings. *Emotion*, *16*(1), 110-116. https://doi.org/10.1037/emo0000093

Rooney, T., Sharpe, L., Todd, J., Michalski, S. C., Van Ryckeghem, D., Crombez, G., & Colagiuri, B. (2024). Beyond the modified dot-probe task: A meta-analysis of the efficacy of alternate attention bias modification tasks across domains. *Clinical Psychology Review,* 102436. https://doi.org/10.1016/j.cpr.2024.102436

Ruscio, A. M., Hallion, L. S., Lim, C. C., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., ... & Scott, K. M. (2017). Cross-sectional comparison of the epidemiology of DSM-5 generalized anxiety disorder across the globe. *JAMA psychiatry*, *74*(5), 465-475. doi:10.1001/jamapsychiatry.2017.0056

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, *25*(1), 128-142. https://doi.org/10.3758/s13423-017-1230-y

Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory STAI (Form Y).* Palo Alto, CA: Consulting Psychologists Press.

Startup, E., & Erickson, T. M. (2006). The Penn State Worry Questionnaire (PSWQ). In G. C. L. Davey & A. Wells (Eds.), *Worry and its psychological disorders* (pp. 101–119). Wiley Publishing.

Stöber, J. (1998). Reliability and validity of two widely-used worry questionnaires: Self-report and self-peer convergence. *Personality and Individual differences, 24*(6), 887-890. https://doi.org/10.1016/S0191-8869(97)00232-8

Tiggemann, M., & Kemps, E. (2020). Is sham training still training? An alternative control group for attentional bias modification. *Frontiers in Psychology, 11,* 583518. https://doi.org/10.3389/fpsyg.2020.583518

Van Bockstaele, B., Verschuere, B., Tibboel, H., De Houwer, J., Crombez, G., & Koster, E. H. W. (2014). A review of current evidence for the causal impact of attentional bias on fear and anxiety. *Psychological Bulletin, 140*(3), 682-721. doi:https://doi.org/10.1037/a0034834

Vrijsen, J. N., Grafton, B., Koster, E. H., Lau, J., Wittekind, C. E., Bar-Haim, Y., ... & Wiers, R. W. (2024). Towards implementation of cognitive bias

modification in mental health care: State of the science, best practices, and ways forward. *Behaviour Research and Therapy,* 104557. https://doi.org/10.1016/j.brat.2024.104557

Wang, Y. P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. *Brazilian Journal of Psychiatry, 35,* 416-431. https://doi.org/10.1590/1516-4446-2012-1048

Wang, Y., Xiao, R., Luo, C., & Yang, L. (2019). Attentional disengagement from negative natural sounds for high-anxious individuals. *Anxiety, Stress, & Coping*, *32*(3), 298-311. https://doi.org/10.1080/10615806.2019.1583539

Wittchen, H.U. (2002). Generalized anxiety disorder: Prevalence, burden, and cost to society. *Depression and Anxiety, 16*(4), 162–171. http://doi.org/10.1002/da.10065

| Question | Hypothesis | Sampling plan | Analysis Plan | Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes | Results |
|---|---|---|---|---|---|---|---|
| Does ABM training lead to a reduction in attention bias? | ABM training will lead to a significant reduction in attention bias, as assessed via a Probe Discrimination Task. | This study will use a sequential design with a maximal *n* of 150. Specifically, data collection will continue until either we have a sensitive Bayes Factor (*B* > 30, or < 1/6) for the composite symptom index analysis, or we have reached a sample size of 150. | A Bayes factor will be computed on the difference between groups in their respective change in attention bias scores from pre- to post-training (ARTS pre-post attention bias minus Sham-ARTS pre-post attention bias). The distribution will be modelled as a half-normal with a mode of 0 and an *SD* of 17.49. This *SD* is the raw effect size originally observed by Hazen et al. (2009; see Table 2, '*Mean Threat Bias Scores by Group*'). | For all Bayes Factors we will adopt the conventional thresholds of values greater than 3 indicating evidence for the alternate hypothesis and values less than 1/3rd indicating evidence for the null. Any Bayes Factor falling between 0.33 – 3 will be deemed insensitive. | BF > 3: ABM training results in a reduction in attention bias.<br><br>BF greater than 1/3 and less than 3: Data is insensitive; therefore, no conclusion can be drawn.<br><br>BF < 1/3: ABM training does not result in a reduction in attention bias. | N/A. This hypothesis is a manipulation check, asking whether a training designed to reduce attention bias *actually* reduces attention bias. | BF < 1/3: ABM training did not lead to a reduction in attention bias. |
| Does ABM training lead to a reduction in anxiety/mood disorder symptomology? | ABM training will lead to a significant reduction in symptoms of general anxiety | This study will use a sequential design with a maximal *n* of 150. Specifically, data collection will | A Bayes factor will be computed on the difference between groups in their respective change in | For all Bayes Factors we will adopt the conventional thresholds of values greater | BF > 3: ABM training results in a reduction in symptoms.<br><br>BF greater than | Hirsch and Mathews (2012) model of pathological worry | BF < 1/3: ABM training did not result in a reduction in anxiety symptomology |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | and depression, measured both individually (PSWQ, STAI-T and BDI) and as a composite symptom index. The composite symptom index is the mean of standardised estimates of the PSWQ, STAI-T and BDI. These estimates are standardised based on general population estimates of scores on these scales, and therefore represent how many standard deviations one's mental health symptomology lies above or below the general population average. | continue until either we have a sensitive Bayes Factor (*B* > 30, or < 1/6) for the composite symptom index analysis, or we have reached a sample size of 150. | composite symptom index scores from pre- to post-training (ARTS pre-post composite symptom index minus Sham-ARTS pre-post composite symptom index). The distribution will be modelled as a half-normal with a mode of 0 and an *SD* of 0.86.<br><br>A Bayes factor will be computed on the difference between groups in their respective change PSWQ scores from pre- to post-training (ARTS pre-post PSWQ score minus Sham-ARTS pre-post PSWQ score). The distribution will be modelled as a half-normal with a mode of 0 and an *SD* of 8.14.<br><br>A Bayes factor will be computed on the difference | than 3 indicating evidence for the alternate hypothesis and values less than 1/3rd indicating evidence for the null. Any Bayes Factor falling between 0.33 – 3 will be deemed insensitive. | 1/3 and less than 3: Data is insensitive; therefore, no conclusion can be drawn.<br><br>BF < 1/3: ABM training does not result in a reduction in symptoms. | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | between groups in their respective change STAI-T scores from pre- to post-training (ARTS pre-post STAI-T score minus Sham-ARTS pre-post STAI-T score). The distribution will be modelled as a half-normal with a mode of 0 and an *SD* of 6.72.<br><br>A Bayes factor will be computed on the difference between groups in their respective change BDI scores from pre- to post-training (ARTS pre-post BDI score minus Sham-ARTS pre-post BDI score). The distribution will be modelled as a half-normal with a mode of 0 and an *SD* of 7.25.<br><br>The *SDs* used to define $H_1$ for each B listed above, represent the | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | relevant raw effect sizes originally observed by Hazen et al. (2009; see Table 1, '*Means and SDs for composite symptom index and individual measures by Group and Time*'). | | | | |
| Do demand effects predict the efficacy of the ABM training in reducing mental health symptomology? | The higher the Phenomenological Control score, the greater the reduction in composite symptom index. | This study will use a sequential design with a maximal *n* of 150. Specifically, data collection will continue until either we have a sensitive Bayes Factor (*B* > 30, or < 1/6) for the composite symptom index analysis, or we have reached a sample size of 150. | A Bayes Factor will be computed on the interaction term of a multiple regression predicting composite symptom index from Experimental Condition (ARTS vs. Sham ARTS) and PC. If PC partially predicts a change in composite symptom index, then the interaction will be greater than zero. The H1 distribution will be modelled as a half-normal with a mode of 0 and *SD* = (Mean Composite | For all Bayes Factors we will adopt the conventional thresholds of values greater than 3 indicating evidence for the alternate hypothesis and values less than 1/3rd indicating evidence for the null. Any Bayes Factor falling between 0.33 – 3 will be deemed insensitive. | BF > 3: Phenomenological Control and the change in mental health symptomology are related, suggesting that demand effects may be influencing apparent ABM efficacy.<br><br>BF greater than 1/3 and less than 3: Data is insensitive; therefore, no conclusion can be drawn.<br><br>BF < 1/3: Phenomenological control and the change in mental | Cristea et al.'s (2015) suggestion that demand effects may drive the supposed therapeutic outcomes of ABM training interventions. | 1/3 < BF < 3: The data was insensitive, and therefore no robust conclusion can be drawn. Therefore, the potential influence of demand effects on the change in mental health symptomology cannot be robustly dismissed; further research is needed. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | Symptom Index Change score / Mean PC score)/2. This *SD* has been modelled based on the ratio-of-means heuristic (Dienes, 2019), given the absence of previous research exploring this effect. | | health symptomology are not related, suggesting that demand effects do not influence apparent ABM efficacy. | | |

**Guidance Notes**

- **Question**: articulate each research question being addressed in one sentence.
- **Hypothesis**: where applicable, a prediction arising from the research question, stated in terms of specific variables rather than concepts. Where the testability of one or more hypotheses depends on the verification of auxiliary assumptions (such as positive controls, tests of intervention fidelity, manipulation checks, or any other quality checks), any tests of such assumptions should be listed as hypotheses. Stage 1 proposals that do not seek to test hypotheses can ignore or delete this column.
- **Sampling plan**: For proposals using inferential statistics, the details of the statistical sampling plan for the specific hypothesis (e.g power analysis, Bayes Factor Design Analysis, ROPE etc). For proposals that do not use inferential statistics, include a description and justification of the sample size.
- **Analysis plan**: For hypothesis-driven studies, the specific test(s) that will confirm or disconfirm the hypothesis. For non-hypothesis-driven studies, the test(s) that will answer the research question.
- **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**: For hypothesis-driven studies that employ inferential statistics, an explanation of how the authors determined a relevant effect size for statistical power analysis, equivalence testing, Bayes factors, or other approach.
- **Interpretation given different outcomes**: A prospective interpretation of different potential outcomes, making clear which outcomes would confirm or disconfirm the hypothesis.
- **Theory that could be shown wrong by the outcomes**: Where the proposal is testing a theory, make clear what theory could be shown to be wrong, incomplete, or otherwise inadequate by the outcomes of the research.