Cogn Ther Res (2014) 38:200–216
DOI 10.1007/s10608-014-9606-z

ORIGINAL ARTICLE

Attention Bias Modification Training Via Smartphone to Reduce Social Anxiety: A Randomized, Controlled Multi-Session Experiment

Philip M. Enock · Stefan G. Hofmann · Richard J. McNally

Published online: 4 March 2014

Abstract Testing feasibility and efficacy of psychological treatment via mobile devices is important, given its potential benefits for high-dosage treatment delivery, widespread and inexpensive dissemination, and efficient research methods. We conducted the first randomized controlled trial of attention bias modification training delivered via smartphones, comparing this training to control training in a double-blind design, also including a waitlist condition. All participants performed a variant of dot-probe training involving faces with neutral and disgust (representative of social threat) expressions in brief sessions three times daily over 4 weeks on their own smartphones, at home or anywhere they chose. Attention bias modification, also known as cognitive bias modification of attention, training included a contingency to induce attentional deployment away from disgust faces, whereas the control training included no contingency. Participants completed weekly Internet-based self-report symptom assessments as well as smartphone-delivered dot-probe attention bias assessments, whose reliability findings supported the viability of using smartphones for reaction-time based assessments. The between-groups training effect on attention bias scores was small, showing statistical significance in some analyses and not in others. On measures of social anxiety, intention-to-treat analyses ($n = 326$) revealed significant pre–post treatment declines with medium to large effect sizes in both training groups, whereas small declines in a waitlist group were nonsignificant. Both training groups showed greater reductions in social anxiety than did waitlist; however, the benefits under these two training conditions were statistically indistinguishable. Improvements in the two training conditions beyond those of waitlist could be attributable to any factors common to them, but not to the contingency training specific to active attention bias modification training.

Keywords Cognitive bias modification · Attentional bias · Attention training · Social anxiety · Mobile app treatment

Introduction

Anxious people, especially those suffering from anxiety disorders, often exhibit an attention bias for threat (Bar-Haim et al. 2007). Although detection of threat is essential for survival, a proclivity for attending to minor threats may needlessly heighten one's anxiety in everyday life. Laboratory studies documenting this bias usually involve variants of the dot-probe paradigm (MacLeod et al. 1986) to measure attentional deployment within pairs of neutral and threatening stimuli.

If attention bias plays a causal role in their development and maintenance of anxiety disorders, then reducing it should diminish a person's vulnerability to experience episodes of anxiety (MacLeod et al. 2002). MacLeod (1995) proposed modifying the dot-probe task by having visual probes consistently appear in the location of non-threatening words (or faces), thereby directing subjects'

Electronic supplementary material The online version of this article (doi:10.1007/s10608-014-9606-z) contains supplementary material, which is available to authorized users.

P. M. Enock (✉) · R. J. McNally
Department of Psychology, Harvard University, Cambridge, MA, USA
e-mail: p.enock@gmail.com

S. G. Hofmann
Department of Psychology, Boston University, Boston, MA, USA

attention away from threatening stimuli. He reasoned that such an attention bias modification (CBM-A, also known as ABM) training procedure could alter anxious individuals' attention bias for threat, potentially reducing anxiety proneness. Inspired by MacLeod, researchers have modified experimental paradigms, creating other cognitive bias modification (CBM) interventions, such that CBM-A represents the cognitive bias modification of attention, whereas other CBM interventions attempt to modify other forms of cognitive bias, such as interpretation bias.

Researchers have conducted randomized controlled trials (RCTs) to test the efficacy of multisession CBM-A training for reducing anxiety, reporting mixed results (for meta-analytic findings, see Beard et al. 2012; Hakamata et al. 2010, 2012; Hallion and Ruscio 2011). CBM-A treatment has also reduced symptoms in people with generalized anxiety disorder (Amir et al. 2009a), pediatric anxiety (Eldar et al. 2012), and depressive symptoms (Wells and Beevers 2010). RCTs appearing prior to the onset of our study showed strong support for CBM-A's superiority to control training (CON) for treating generalized social anxiety disorder (Amir et al. 2009b; Schmidt et al. 2009). One study also showed large effects of CBM-A versus CON for reducing social anxiety (Li et al. 2008) on the Social Interaction Anxiety Scale (SIAS; Mattick and Clarke 1998), but not on the Social Phobia Scale (Mattick and Clarke 1998) or the Fear of Negative Evaluation Scale (Watson and Friend 1969). More recently, although one study indicated superior efficacy of CBM-A versus CON (Heeren et al. 2012) for social anxiety disorder, six others have not (Boettcher et al. 2012, 2013; Bunnell et al. 2013; Carlbring et al. 2012; Neubauer et al. 2013; Sawyer et al. 2012). Although researchers have usually administered training in research laboratories, delivery through a web-based platform could increase its accessibility, as MacLeod et al. (2007) demonstrated. Additionally, four RCTs of CBM-A delivered training via Internet to participants' home PCs (Boettcher et al. 2012, 2013; Carlbring et al. 2012; Neubauer et al. 2013).

For several reasons, we investigated the feasibility of delivering CBM-A on smartphone mobile devices rather than PCs. First, the approach could augment CBM-A's effects by facilitating a higher dosage via ease of frequent training, perhaps enabling enduring changes in attentional habits more likely. Increasing the frequency of sessions is easy with smartphones, as participants can perform training anywhere throughout the day. Second, training in diverse locations could foster generalization of clinical benefits. Third, higher frequency enables brevity of sessions, perhaps increasing tolerability of this repetitive task. Fourth, mobile devices have great dissemination potential, and the popularity and convenience of mobile apps suggest that people will welcome psychological help via this venue. Fifth, establishing CBM-A's efficacy on the small screens of smartphones would confirm its robustness across varying sizes of stimulus presentation. Sixth, establishing smartphones' viability for reaction-time based tasks would open the platform to this category of psychological assessment methods.

In the first study of CBM-A on mobile devices, we tested the feasibility of reducing social anxiety and worry via iPhone, iPod Touch, and Android-based smartphones (Enock and McNally 2010, summarized in Enock and McNally 2013). Using a multiple baseline across subjects single-case design (Barlow et al. 2009), we administered 1 or 2 weeks of CON followed by 3 weeks of frequent CBM-A training to 16 anxious individuals. Participants, mostly Harvard University students, generally found treatment to be acceptable. They reported training on their mobile devices primarily at home (67 % of sessions), but also in libraries (9 %), bathrooms (6 %), and other sites. Although symptoms declined significantly, improvements did not differ between the CON and CBM-A phases.

In the present RCT, we tested the efficacy of smartphone-delivered CBM-A. We randomly assigned participants to one of three conditions: CBM-A, CON, or waitlist. Our study tested the following hypotheses: (1) that social anxiety would decrease more in the CBM-A group than in the CON group; (2) that social anxiety would decrease more in both CBM-A and CON groups than in the waitlist group; and (3) that attention bias scores would diverge at the start of training and grow further apart during training, with lower scores in the CBM-A group than in the CON training group. If CBM-A causes greater decreases in social anxiety than does CON, then this would imply that the putative active ingredient of CBM-A, the contingency linking probes to locations opposite to threat stimuli, produces these differential decreases. If, however, CBM-A and CON yield similar symptom declines, both greater than those of waitlist, this would imply that performing either training has anxiolytic effects that are not specific to the theoretically active ingredient. We included measures of worry and depression as secondary measures to characterize the breadth of clinical benefit.

Method

Participants

On the study website, 826 individuals signed up, indicating their desire to participate in the study. There were 429 participants (52.2 % male) randomized in the study, ranging in age from 18 to 68 years old ($M = 34.8$, SD = 11.4). The racial/ethnic composition of the sample was 81.2 % white, 11.7 % Asian, 4.9 % Hispanic or Latino, 1.4 %

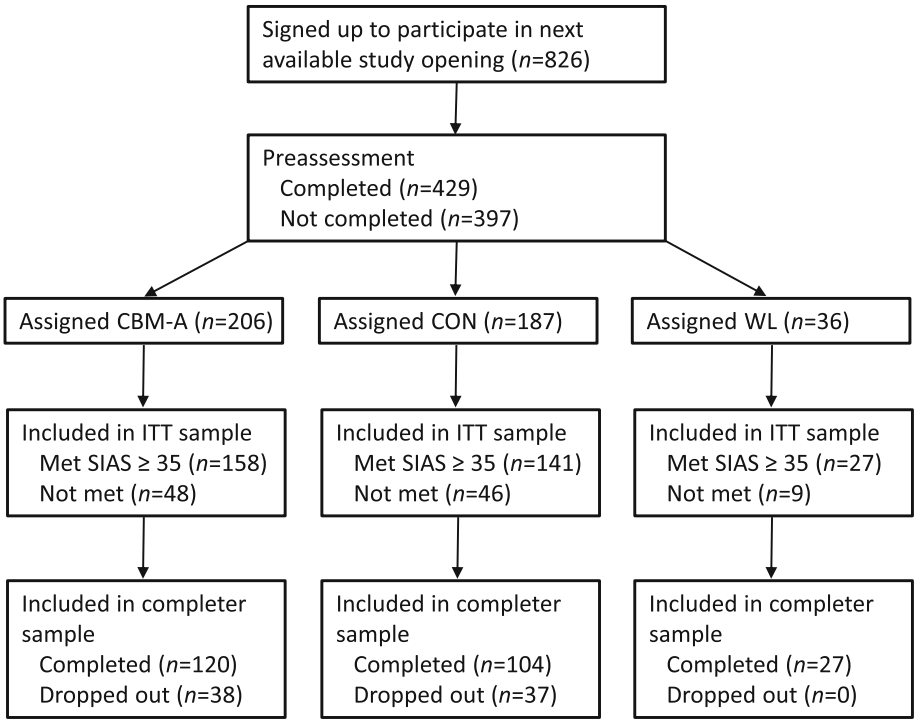

Fig. 1 Participant randomization, inclusion in ITT analyses, and dropouts

Black or African American, 0.9 % American Indian or Alaskan Native, 0.2 % Native Hawaiian or Pacific Islander, and 2.3 % other. The mean years of education was 17.0 (SD = 3.0). Participants used these handheld devices during the study: 52.1 % iPhone (Apple Computer, Inc., Cupertino, CA, USA), 22.4 % iPod Touch (Apple Computer), and 24.1 % any Android-based (Google, Mountain View, CA, USA) smartphone.[1]

Data collection occurred between September 12, 2010 and January 6, 2012. Prior to June 18, 2011, we randomly assigned participants to either the active (CBM-A) or control (CON) conditions with a .5 probability of assignment to each condition. From June 18, 2011 to January 6, 2012, we randomly assigned participants to CBM-A, CON, or waitlist (WL) conditions with a .33 probability for each condition.

We recruited from a range of sources, mostly word of mouth stemming from a news article in the *Economist* magazine (Gee 2011), but also online messageboards, Craigslist, the Harvard Study Pool, flyers posted locally, and Google search and AdWords. Harvard students who sought course credit for their participation received it, whereas others received no compensation other than the opportunity to receive CBM-A after the 2-month follow-up. We encouraged participation from individuals with high levels of social anxiety, generalized anxiety or worry, or both, but there was no exclusion criterion during signup based on symptoms.

To participate, individuals were required only to: (1) have access to an iPhone, iPod Touch, or Android-based phone with Wi-Fi or other Internet access during the 4 weeks of training, (2) be aged 18 or older, and (3) be sufficiently proficient in English to read and understand the instructions and consent form. For data analyses, we applied cutoff scores to examine socially anxious individuals. A flow chart of participant assignment and dropout is included (see Fig. 1).

Materials

Face Stimuli

We used the same face stimuli (Matsumoto and Ekman 1988) as other studies (Amir et al. 2008, 2009b; Schmidt et al. 2009), balanced across the factors of sex (male vs. female) and ethnicity (Caucasian vs. Japanese). Each model portrayed disgust and neutral expressions.

Smartphone Attention Bias Modification Training Task

Participants trained on their own handheld devices, which had a screen width of approximately 5 cm (horizontal) × 7.5 cm

[1] Data from study signup was missing for one participant, and the answer on handheld device type was missing for 1.4 % of participants.

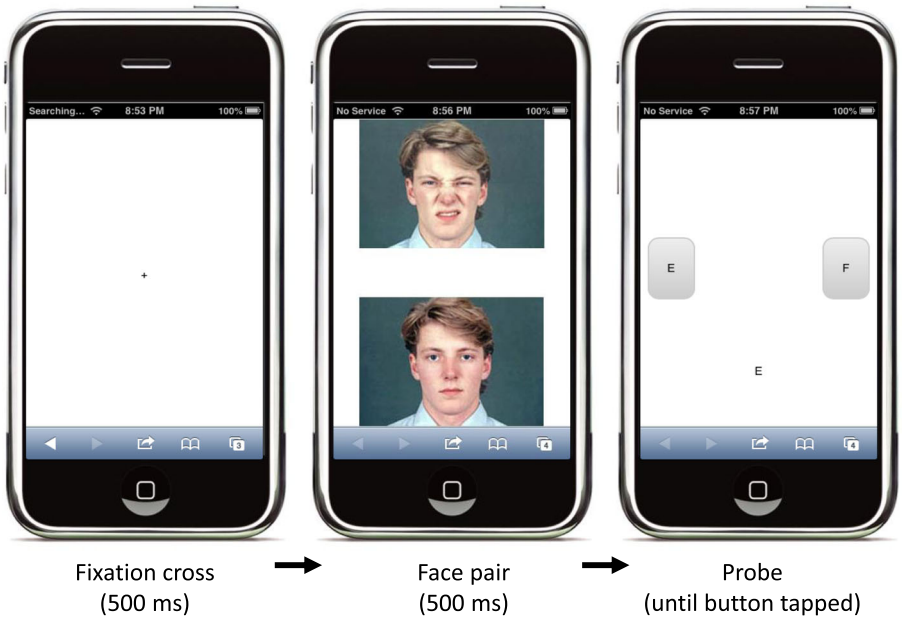

Fig. 2 Trial screen sequence for dot-probe attention bias assessment and training. Images are to scale, as they appeared on an iPhone model 3GS

(vertical). Screenshots from an iPhone 3GS illustrate, to scale, the sequence of screens and timings for one trial of dot-probe training/assessment (Fig. 2). We implemented CBM-A on a website by using JavaScript and HTML code for task presentation and response collection with MySQL and PHP code for server-side programming. Instructions walked participants through creating a home screen shortcut icon. Hence, the user experience of accessing the training page resembled that of a mobile app, but it lacked an app's typical user-friendly interface. In selecting a browser web app instead of a native app, we considered timing precision issues. Large differences in precision have little impact on reaction time task scores (Damian 2010; Ulrich and Giray 1989), minimizing concerns about using a web app.

In this paradigm, a fixation cross appeared for 500 ms, followed by the simultaneous presentation of two faces for 500 ms. One face appeared on top, and the other appeared on the bottom of the screen. On any given trial, both faces were of the same model.[2] In 20 % of the trials, the model displayed the identical neutral expression, whereas in the remaining 80 % of the trials the model displayed a neutral expression and a threatening expression conveying disgust. Immediately thereafter, a probe replaced one of the two faces. The probe was either an E or an F. The participants' task was to push one of two buttons on the screen as quickly as possible indicating whether an E or F was present. In the control version of the task (given to the CON group), the probe replaced each type of face equally often. In the active training version (given to the CBM-A group), the probe replaced the neutral (nonthreatening) face every time a neutral-threat face pair appeared. Thus, the goal was for participants to learn implicitly to attend to neutral rather than to threat faces, and to disengage quickly if their attention was captured first by the threat faces. Each training session comprised 80 trials. The median session duration was 2.24 min, and most sessions (86.6 %) took 2–2.5 min to complete.

Smartphone Attention Bias Assessment Task

Participants used the same device for assessment as for training. Also, the stimuli were the same. In using the training stimuli in assessment, we intended to maximize the chance of detecting a change in attention bias during the experiment. Indeed, our previous study's data failed to show attention bias reduction for new stimuli (Enock and McNally 2010, summarized in Enock and McNally 2013). The assessment task differed from training in that each session comprised 160 trials, and that all trials involved threat-neutral facial pairs.

Modified Posner Cueing Task

Local participants able to visit our laboratory ($n = 62$) performed a modified Posner cueing task on a desktop computer with the Google Chrome browser, a 19" monitor with a 5:4 aspect ratio. The task was programmed via

[2] Due to a programming error, in two of the eight pairs, the face of one model (ID YY2) was paired with the face of another model (ID ES2) who shared similar features. Because many participants had already completed training when we discovered this problem, we maintained the same stimuli throughout the study without mentioning it to participants. Two participants reported noticing the mismatch. However, the error did not violate the requirement for attention bias modification that each pair consist of a neutral and a threat stimulus.

JavaScript and HTML. The distance from the participants' eyes to the screen was 50–70 cm. We used the method of Amir et al. (2003), including their positive, neutral, and social threat words. In this paradigm, a black screen appeared with two white box outlines on either side (left and right) of a fixation cross ("+") for 500 ms. A word then appeared in one of the boxes for 250 ms, after which an asterisk appeared in either the same box (a valid cue) or the other box (an invalid cue). On some trials, no word appeared (no cue). Holding the mouse in both hands, with their thumbs on the mouse buttons, participants pressed the left or right button to indicate on what side the asterisk appeared. Response latencies typically show a cue dependency effect; that is, mean latencies are lower for validly cued trials than for invalidly cued trials, with uncued trials' mean latency falling midway between that of validly and invalidly cued trials. Participants performed 24 practice trials and 288 training trials per session, including three rest breaks. Participants were orally instructed that the asterisk would usually, but not always, appear in the same location as the word and told to perform the task as quickly and as accurately as possible. The session took about 10 min.

Social Interaction Anxiety Scale (SIAS)

The SIAS (Mattick and Clarke 1998) is a self-report measure of social anxiety, psychometrically strong in online administration (Hedman et al. 2010). The SIAS-17 excludes the three reverse-scored items, reducing its overlap with extraversion (Rodebaugh et al. 2007). We used the full version for screening and the SIAS-17 as an outcome measure. In our data, internal consistency was acceptable (SIAS at preassessment, $\alpha = .85$; SIAS-17 across assessment points $\alpha = .85$–$.95$).

Liebowitz Social Anxiety Scale, Self-Report Version (LSAS-SR)

The LSAS-SR (Baker et al. 2002; Cox et al. 1998) assesses fear and avoidance in social and performance situations. The self-report version has similar properties (Fresco et al. 2001) to the clinician-administered version of the scale (Liebowitz 1987), as does the Internet-administered LSAS-SR (Hedman et al. 2010). In our data, internal consistency was acceptable (Fear subscale, $\alpha = .91$–$.96$; Avoidance subscale, $\alpha = .91$–$.95$).

Penn State Worry Questionnaire (PSWQ)

The PSWQ (Meyer et al. 1990) measures worry. The Internet-administered version shares the strong internal consistency and concurrent validity of the paper-and-pencil version (Zlomke 2009). In our data, internal consistency was acceptable ($\alpha = .93$–$.95$).

Table 1 Reliability coefficients for internal consistency and test–retest reliability

	r, CBM-A	*r*, CON
Smartphone-delivered dot-probe		
Internal consistency		
Preassessment/practice	−.05	.00
Day 2	.07	−.23
Week 1	.18	.14
Week 2	.35	.04
Week 3	.50	.39
Week 4 (postassessment)	.53	.31
Test–retest reliability		
Preassessment-day 2	−.25	.05
Day 2–week 1	.08	−.13
Week 1–week 2	.45**	.26*
Week 2–week 3	.70*	.49**
Week 3–week 4	.61**	.63**
PC-delivered modified Posner task		
Internal consistency		
Preassessment	−.22	
Week 4 (postassessment)	−.04	
Test–retest reliability		
Pre–postassessment	.07	

Values have been augmented by the Spearman–Brown formula for estimating full-length reliability from split halves. The Monte Carlo reliability estimation method used precludes significance testing; hence, *p* values were not calculated. Modified Posner cueing task reliability estimates were performed with groups combined

Significance key: * $p < .01$; ** $p < .001$

Depression Anxiety Stress Scales, 21-Item Version (DASS)

The DASS (Lovibond and Lovibond 1995) measures depression, anxiety, and stress. We included it primarily to assess depression. We employed the 21-item version (Antony et al. 1998) and report results from the DASS depression subscale. In our data, internal consistency was acceptable (DASS-Depression, $\alpha = .90$–$.95$).

Reliability of Smartphone-Delivered Attention Bias Assessments

To estimate the internal consistency of the smartphone-delivered attention bias assessments, we employed a Monte Carlo simulation process that repeats the steps (2,000 times) of randomly reselecting halves and calculating split-half reliability of bias scores of the halves (Enock et al. 2012). We used this method because other ones yield

unstable estimates of bias score reliability (e.g., Cronbach's alpha or standard split-half reliability).

For test–retest reliability, we calculated a Pearson-*r* correlation of scores at each adjacent time point. In all reliability analyses, we used data from protocol completers, rather than the full intention-to-treat (ITT) sample, and we excluded missing data, rather than employing last observation carried forward (LOCF). Reliabilities were low or nonsignificant for the first two time points (preassessment and Day 2) and significant in the later weeks. The results appear in Table 1.

Reliability of PC-Delivered Modified Posner Cueing Task Assessments

Employing the Monte Carlo method described above, we analyzed CBM-A and CON together, due to the small number of observations (41 preassessment and 37 postassessment sessions). Results showed no reliability in the attention bias scores (see Table 1).

To ensure that our implementation of the task was effectively measuring reaction times and that unreliability was not due to factors such as participant distraction or lack of effort during the task, we created scores for the cue validity effect, collapsing across all word types. These scores do not measure attention bias and are not expected to relate to other variables of interest in this study. Rather, they index the relative cost of a participant responding to an invalid cue compared with a valid cue. Reliability estimates for these cue validity scores were $r = .83$ at preassessment and $r = .89$ at postassessment, with a test–retest reliability of $r(35) = .62$, $p < .001$, suggesting that the task did measure reaction times suitably for reliable assessment.

Procedure

Participants began by completing the consent form and answering questions online as they signed up for the study via KeySurvey, a web survey platform. They answered demographic questions before completing the SIAS and PSWQ. We asked participants with a high score on either measure (35 or higher on the SIAS or 56 or higher on the PSWQ) whether they were able to attend two laboratory sessions (pre-training and post-training). We did not screen for other psychopathology, and nor did we inquire about current treatment at sign-up. Those who did not meet the symptom cutoff scores or who declined to attend the laboratory sessions completed the procedure with online and email instructions only.

Eligible individuals visited a website that outlined the training protocol and provided instructions for accessing the training web page on their handheld devices. Participants created an icon for easy access, alongside their home screen apps, and scheduled daily calendar reminders for their training. Those who did not attend the laboratory sessions commenced training on their own, after an initial email from us.

For those attending the laboratory sessions, in the initial meeting, the experimenter explained to the participant how to use the dot-probe task on his or her handheld device and administered the modified Posner cueing task on a laboratory computer. This took approximately 25 min.

We instructed participants to perform three training sessions per day during the 4-week training period. We asked them to perform one session in the morning (4 a.m.–12 p.m.), one session in the afternoon (12 p.m.–8 p.m.), and one session in the evening (8 p.m.–4 a.m.). Participants could make up for missed sessions that evening, and they were limited to three sessions per day. Daily emails sent to participants reminded them to continue training. At the end of any session, participants could view the number and percentage of sessions they had completed. Participants completing fewer than 80 % of sessions in the previous 5 days or since the start saw the notice, "A good target is to complete 80 %."

Participants received emails containing links to the following online assessment battery of measures on Day 1, Day 8, Day 15, Day 22, Day 29 (the day after the end of training), 1-month follow-up (Day 58), and 2-month follow-up (Day 88): LSAS-SR, SIAS, PSWQ, and DASS. During the training period, instructions asked participants to consider only the past week for their responses. The timing of smartphone-delivered attention bias assessments was as follows: On Day 1, participants completed a practice session with 80 trials (including 16 trials where both faces showed neutral expressions, in addition to 64 critical trials), which also served as preassessment. Our primary attention bias assessments all occurred after training had begun and participants had practiced on the dot-probe task, with sessions consisting of 160 trials on Day 2, Day 8 (henceforth Week 1), Day 15 (Week 2), Day 22 (Week 3), and Day 28 (Week 4 or postassessment). They completed these assessment sessions upon visiting the training web page, before doing any training on that day. Participants who did not visit the training page on the scheduled day completed the session the next time they visited the training web page.

For participants who attended laboratory sessions, the second visit took place at approximately the end of the training period. They completed the final assessment of the Posner task.

All participants in training conditions were told their condition and debriefed only after they had completed the 2-month follow-up questionnaire. Those who withdrew from the study were debriefed and told their condition if they explicitly requested this information.

Participants in the WL condition completed the same online assessment battery of measures at the same intervals as participants in the training conditions, but they did not use their handheld device for the study during the main period of 4 weeks. After this period, they began 4 weeks of the active training either immediately or at a convenient time for them.

Data Analyses

We used SPSS software (SPSS Inc. 2009) for mixed ANOVA analyses and R (R Development Core Team 2012) for all other analyses.

Inclusion Criteria for Intention-to-Treat (ITT) Analyses

Our primary analyses involved an ITT approach on participants with high social anxiety, defined as an SIAS of at least 35 at preassessment. Heimberg et al. (1992) used an SIAS cutoff of 34, which was one standard deviation above the mean of their community sample. In that study, the cutoff classified 82 % of the social anxiety disorder sample correctly as cases and 18 % of the community sample incorrectly as cases.

For inclusion in analyses, participants also needed to have completed the preassessment self-report measures and to have been randomized. For those assigned to CBM-A or CON, completing at least one training session signified randomization, and for those in WL, there was no such requirement. Although 826 people completed screening measures, we were often unable to run participants immediately following their signup, and some had waited 4 months to begin. Unsurprisingly, 60.5 % of those signing up never began the study.

The ITT analysis sample comprised 326 participants. The group sizes were $n = 158$ (CBM-A), $n = 141$ (CON), and $n = 27$ (WL). The lag between signup and preassessment varied ($M = 34.1$ days, $Mdn = 11$, SD = 41.4). Given the absence of a diagnostic interview in this study, clinical cutoff scores that are both sensitive and specific to diagnostic status are useful. They provide metrics for estimating the proportion of individuals who might have received a diagnosis of social anxiety disorder (SAD) or generalized anxiety disorder (GAD) by clinical interview at preassessment if we had conducted interviews. In the ITT sample, 96.6 % of participants' LSAS-SR scores exceeded 30, signifying likely SAD (Rytwinski et al. 2009); 69.3 % of LSAS-SR scores exceeded 60, signifying likely *generalized* SAD (Rytwinski et al. 2009); and 50.3 % of PSWQ scores exceeded 65, signifying likely GAD (Fresco et al. 2003). Although diagnostic cutoffs for the DASS-Depression are unavailable, to our knowledge, the mean ($M = 19.0$) and median ($Mdn = 16.0$) fell within the moderate range (14–20 points; Lovibond and Lovibond 1995) of depression symptom severity.

Clinical Outcome Analyses

We report results from four measures separately: LSAS-SR, SIAS-17, PSWQ, and DASS-Depression subscale. For each measure, we first conducted a one-way ANOVA to ensure that preassessment scores did not significantly differ among the groups. Second, we conducted a 3 Group (between-subjects factor: CBM-A, CON, WL) × 5 Time (within-subjects factor: preassessment, Week 1, Week 2, Week 3, and postassessment) mixed ANOVA. In all clinical outcome ANOVAs, Mauchly's Test of Sphericity indicated a lack of sphericity; hence, we employed the Greenhouse-Geisser correction for all main effects of time and the Group × Time interaction effects. We selected the ANOVA and LOCF methods for these analyses, rather than multilevel linear modeling, due to their customary use in treatment research including CBM-A RCTs.

To further explore the main effects of time and interaction effects, we calculated pre–post change scores from the postassessment (end of training) score minus the preassessment (prior to training) score. We then conducted two-tailed Welch's t tests on these scores. To assess whether scores declined significantly in each group, from preassessment to postassessment, we employed one-sample t tests. To assess whether the degree of declines in scores differed between individual groups, we conducted independent samples t tests for each pair. To test the hypothesis that change scores of both training groups together differed from those of the WL group, we conducted a contrast with weights coded as 1 (CBM-A), 1 (CON), and −2 (WL). Finally, we conducted two-tailed independent samples t tests comparing the groups' scores at postassessment and at 2-month follow-up. We include Cohen's d effect sizes for each t test.[3]

The size of the WL group ($n = 27$) in the ITT sample was much smaller than that of the CBM-A ($n = 158$) and CON ($n = 141$) groups. These unequal sample sizes between WL and the other groups would be problematic for t tests if unequal variances were present. Therefore, we used Welch's method in t tests, which does not assume

[3] All p values are two-tailed. T tests employ Welch's method (equal variances are not assumed). Those appearing in the second paragraph under each symptom heading (e.g., under LSAS above) apply to comparisons of change scores, whereas those in the third paragraph apply to scores at one time, either postassessment or two-month follow-up. For pre–post Cohen's d effect sizes within each group, we used the formula $(M_{post} - M_{pre})/SD_{pre}$. For between-groups Cohen's d effect sizes, we employed the standard Cohen's d formula using pooled standard deviation, $(M_{Group1} - M_{Group2})/SD_{pooled}$.

equal variances, and we also report Levene's tests for unequal variances.

Dot-Probe Attention Bias Score Analyses

We calculated a bias score for each session by subtracting the mean response time for threat-location probes from the mean response for neutral-location probes. Hence, positive bias scores represent attention bias toward threat faces, and negative bias scores represent bias away from threat, or, equivalently, bias toward neutral faces.

We analyzed bias scores in two ways: (1) a multilevel linear modeling approach tailored to the specific circumstances of the present data and missing data, and (2) a traditional approach comprising ANOVA and *t* tests with LOCF for missing data, often used in the CBM-A literature.

For the tailored analyses, we excluded preassessment/practice sessions due to poor data quality (see Procedure and Data Reduction sections). Due largely to dropout from the study, a sizable amount of dot-probe assessment data points were missing, especially at later time points (e.g., 26.8 % of sessions missing at Week 4 versus 2.7 % at Day 2, 15.7 % at Week 2). The LOCF policy presupposes that scores of any dropouts would have remained constant. In contrast, multilevel linear modeling does not make this assumption, and it uses observed data to estimate missing scores.

Multilevel linear modeling (a.k.a., linear mixed effects modeling) is a form of regression appropriate for mixed designs with between-groups conditions and repeated measures assessments (Gelman and Hill 2007). Accordingly, we used model comparison methods to test the hypothesis that CBM-A group bias scores would be lower than those of the CON group during and after training. This is a test of the *main effect* of group, collapsed across assessments, following the start of training, *not* an interaction effect. As to the Group × Time interaction effect, the hypothesis tested is that the difference between the groups' scores would change during the training. We employed the R package "nlme" (Pinheiro et al. 2010) with maximum likelihood estimation and an autoregressive correlation structure to model the presumptive pattern of closer time points having closer scores than distant ones. Bias score was the dependent variable. Participants were modeled as a random effect, with intercepts allowed to vary. Time was entered as a continuous predictor.

For our traditional analyses, we included the preassessment/practice sessions. We also calculated one score per participant, termed the Post Mean, to serve as postassessment, a mean of the five assessments that occurred after training commenced. We employed mixed ANOVA analyses, focusing on the Group × Time interaction effect, as well as *t* tests, all using LOCF for missing data.

Modified Posner Cueing Task Attention Bias Score Analyses

To focus on the most theoretically relevant tests and save space, we calculated one bias score per session. As the difference between participants' responses to social threat words and nonthreat words has the most theoretical interest, we first collapsed neutral and positive words into one category, nonthreat. For each session, we then calculated cue validity effect scores for threat and nonthreat words by subtracting the mean response time for validly cued trials from the mean response time for invalidly cued trials, within trials of each word type. In creating a bias score for a session, we followed the logic of Fox et al. (2001) such that the cue validity effect should be larger for threat trials than for nonthreat trials in socially anxious individuals. Hence, we calculated the attention bias score by subtracting the cue validity effect score for nonthreat words from the cue validity effect score for threat words. A bias score greater than zero signifies an attention bias toward threat. We conducted statistical tests similar to those of the clinical outcome pre–post analyses.

Results

Protocol Compliance

Participant Dropout

We defined protocol completers as participants who met inclusion criteria for the ITT sample, completed postassessment self-report measures at the end of Week 4, and completed at least 20 of the 83 training sessions in the protocol for CBM-A or CON (not applicable for WL), which amounted to 1,280 critical training trials. Dropouts, defined as participants in the ITT sample who were not completers, were $n = 38$ (24.1 %, CBM-A), $n = 37$ (26.2 %, CON), and $n = 0$ (WL). In most cases, we did not obtain a reason from participants for their dropout; hence, we do not attempt to report reasons for dropout.

Training Session Completion

Out of the 83 training sessions in the 4-week protocol, participants (including dropouts) performed a mean of 53.9 sessions (64.9 %). Among protocol completers, a mean of 64.8 (78.1 %) sessions were performed, and the number of sessions was not significantly correlated (all $ps > .11$) with

Table 2 Symptom measure means and standard deviations for intention-to-treat analysis

	CBM-A ($n = 158$)[a]	CON ($n = 141$)[b]	WL ($n = 27$)[c]
LSAS			
Pre	74.2 (23.5)	72.8 (24.3)	80.1 (29.9)
Post	57.6 (30.7)	58.2 (31.3)	75.0 (34.7)
One-month follow-up	57.5 (29.3)	57.9 (30.6)	NA[d]
Two-month follow-up	58.1 (28.2)	57.3 (31.4)	NA
Pre–post change	−16.6 (22.9)	−14.6 (19.2)	−5.1 (22.4)
Pre–post Cohen's *d*[e]	−0.71	−0.60	−0.17
SIAS-17			
Pre	43.2 (8.8)	42.1 (9.7)	47.1 (11.7)
Post	33.9 (13.9)	33.4 (13.9)	44.4 (14.2)
One-month follow-up	34.1 (13.3)	32.7 (13.8)	NA
Two-month follow-up	34.7 (13.5)	32.8 (13.5)	NA
Pre–post change	−9.3 (10.5)	−8.7 (10.4)	−2.7 (8.2)
Pre–post Cohen's *d*	−1.06	−0.90	−0.23
PSWQ			
Pre	63.7 (11.7)	62.2 (13.7)	64 (12.5)
Post	57.7 (12.4)	57.6 (14.2)	62.1 (12.4)
One-month follow-up	57.5 (12)	56.8 (14.5)	NA
Two-month follow-up	57.1 (12.8)	57 (14.6)	NA
Pre–post change	−6.0 (9.6)	−4.6 (7.5)	−1.9 (8.7)
Pre–post Cohen's *d*	−0.51	−0.34	−0.15
DASS-depression			
Pre	19.2 (10.2)	18.1 (10.9)	22.4 (13)
Post	14.6 (11.6)	14.1 (12)	21.8 (12.9)
One-month follow-up	14.9 (10.8)	14.7 (11.6)	NA
Two-month follow-up	14.6 (11.5)	14.0 (11.7)	NA
Pre–post change	−4.6 (9)	−4.0 (8.3)	−0.7 (8.5)
Pre–post Cohen's *d*	−0.45	−0.37	−0.05

These data are from ITT analyses. Missing data, mainly attributable to study dropout, were filled in via LOCF. Values are mean scores on the given scale, with standard deviations in parentheses

[a] Number of observations missing and filled via LOCF in CBM-A group, cumulatively: 0 (pre), 37 (post), 42 (1-month follow-up), 46 (2-month follow-up)

[b] Number of observations missing and filled via LOCF in CON group, cumulatively: 0 (pre), 35 (post), 36 (1-month follow-up), 39 (2-month follow-up)

[c] No observations missing

[d] WL participants were offered active training immediately following the WL period; therefore, we did not collect follow-up data from them

[e] Cohen's *d* within each group is $(M_{post} - M_{pre})/SD_{pre}$

change on any clinical outcome measure, in either training condition.

Symptom Assessment Completion

Of the five weekly online self-report assessments, participants completed 83.5 % (CBM-A), 83.3 % (CON), and 98.5 % (WL).

Clinical Outcome Measures

Descriptive statistics for the clinical outcome measures are presented in Table 2.

Tests for Unequal Variance Between Groups

Levene's tests for unequal variances comparing CON and WL scores were nonsignificant for all measures at preassessment, postassessment, and for pre–post change scores (all $ps > .09$). For comparing CBM-A and WL scores, Levene's tests were nonsignificant for all measures at postassessment and for pre–post change scores (all $ps > .15$), but significant for the preassessment SIAS-17 ($p = .033$) and DASS-Depression ($p = .025$), suggesting unequal variances. On the preassessment SIAS-17, SD = 8.8 for CBM-A and SD = 11.7 for WL. On the preassessment DASS-Depression, SD = 10.2 for CBM-A and SD = 13.0 for WL. Thus, the results of testing for unequal variance show no significant differences between WL and either of the other groups except for two instances where the preassessment variance in the WL group was higher.

LSAS-SR

On the LSAS-SR, scores at preassessment did not significantly differ among the groups ($F(2, 69.8) = 0.74$, $p = .48$). The 3 Group × 5 Time mixed ANOVA yielded a main effect of time ($F(2.7, 308) = 31.12$, $p < .001$, $\eta_p^2 = .088$) and a trend of a main effect of group ($F(2, 323) = 2.59$, $p = .077$, $\eta_p^2 = .016$). The Group × Time interaction was significant ($F(5.3, 308) = 2.67$, $p = .018$, $\eta_p^2 = .016$).

Scores decreased from pre to postassessment in the two training groups (CBM-A, $t(157) = -9.13$, $p < .001$, $d = -0.71$; CON, $t(140) = -9.04$, $p < .001$, $d = -0.60$) but not significantly in WL ($t(26) = -1.18$, $p = .25$, $d = -0.17$). Declines in CBM-A and CON did not differ ($t(295.8) = -0.83$, $p = .41$, $d = -0.10$) but declines were greater in CBM-A versus WL ($t(36.0) = -2.47$, $p = .018$, $d = -0.51$) and in CON versus WL ($t(33.7) = -2.07$, $p = .046$, $d = -0.60$). A contrast on pre–post change scores with weights coded as 1 (CBM-A), 1 (CON), and

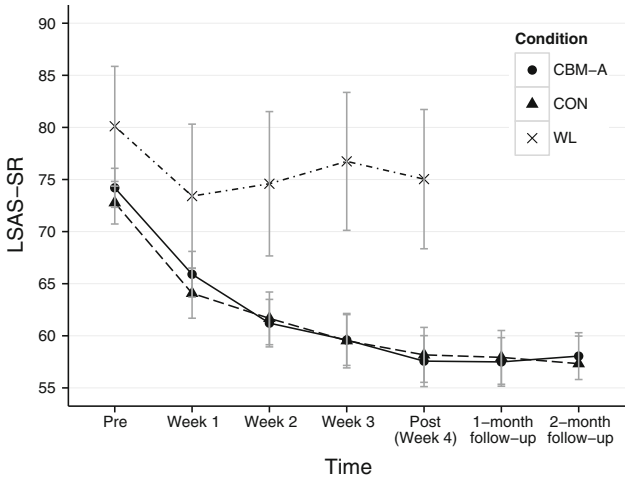

Fig. 3 LSAS-SR scores during main protocol and follow-up. This figure depicts the means and standard errors, with missing data filled via LOCF

−2 (WL) confirmed that the two training groups showed greater declines than did WL ($t(30.3) = -2.36$, $p = .025$). From postassessment to 2-month follow-up, scores did not change significantly within CBM-A or CON ($ps > .48$).

LSAS-SR scores for CBM-A and CON did not differ at postassessment ($t(291.9) = -0.17$, $p = .87$, $d = -0.02$) or at 2-month follow-up ($t(283.3) = 0.21$, $p = .83$, $d = 0.02$). However, CBM-A and WL scores differed at postassessment ($t(33.3) = -2.45$, $p = .020$, $d = -0.56$), as did CON and WL scores ($t(34.6) = -2.35$, $p = .025$, $d = -0.53$).

For a comparison of LSAS-SR scores across all time points and conditions, see Fig. 3.

SIAS-17

On the SIAS-17, scores at preassessment did not significantly differ among the groups ($F(2, 69.4) = 2.22$, $p = .12$). The 3 Group × 5 Time mixed ANOVA yielded a main effect of time ($F(2.4, 308) = 42.54$, $p < .001$, $\eta_p^2 = .12$) and a main effect of group ($F(2, 323) = 6.54$, $p = .002$, $\eta_p^2 = .039$). The Group × Time interaction was significant ($F(4.9, 308) = 5.05$, $p < .001$, $\eta_p^2 = .030$).

Scores decreased from pre to postassessment in the two training groups (CBM-A, $t(157) = -11.11$, $p < .001$, $d = -1.06$; CON, $t(140) = -10.02$, $p < .001$, $d = -0.90$) but not significantly in the WL group ($t(26) = -1.75$, $p = .092$, $d = -0.23$). Declines in CBM-A and CON did not differ ($t(294.2) = -0.47$, $p = .64$, $d = -0.06$), but declines were greater in CBM-A versus WL ($t(42.4) = -3.70$, $p < .001$, $d = -0.64$) and in CON versus WL ($t(43.8) = -3.34$, $p = .0017$, $d = -0.60$). A contrast on pre–post change scores with weights coded as 1 (CBM-A), 1 (CON), and −2 (WL), confirmed that the two training groups showed greater declines than the WL group ($t(34.3) = -3.74$, $p < .001$). From postassessment to 2-month follow-up, scores did not change significantly within CBM-A or CON ($ps > .16$).

SIAS-17 scores for CBM-A and CON did not differ at postassessment ($t(292.8) = 0.28$, $p = .78$, $d = 0.03$) or at 2-month follow-up ($t(292.8) = 1.20$, $p = .23$, $d = 0.14$). However, CBM-A and WL scores differed at postassessment ($t(35.0) = -3.57$, $p = .001$, $d = -0.76$), as did CON and WL scores ($t(36.2) = -3.69$, $p < .001$, $d = -.79$).

PSWQ

On the PSWQ, scores at preassessment did not significantly differ among the groups ($F(2, 73) = 0.55$, $p = .58$). The 3 Group × 5 Time mixed ANOVA yielded a main effect of time ($F(2.9, 308) = 19.75$, $p < .001$, $\eta_p^2 = .058$) but no main effect of group ($F(2, 323) = 1.15$, $p = .32$, $\eta_p^2 = .0071$). The Group × Time interaction was significant ($F(5.7, 308) = 2.27$, $p = .038$, $\eta_p^2 = .014$).

Scores decreased from pre to postassessment in the two training groups (CBM-A, $t(157) = -7.84$, $p < .001$, $d = -0.51$; CON, $t(140) = -7.32$, $p < .001$, $d = -0.34$) but not significantly in the WL group ($t(26) = -1.14$, $p = .26$, $d = -0.15$). Declines in CBM-A and CON did not differ ($t(292.3) = -1.36$, $p = .17$, $d = -0.16$), but declines were greater in CBM-A versus WL ($t(37.5) = -2.19$, $p = .034$, $d = -0.43$). Declines were not significantly greater in CON versus WL ($t(33.7) = -1.50$, $p = .14$, $d = -0.35$). A contrast on pre–post change scores with weights coded as 1 (CBM-A), 1 (CON), and −2 (WL) suggested that the two training groups showed a trend toward greater declines versus WL ($t(30.7) = -1.93$, $p = .063$). From postassessment to 2-month follow-up, scores did not change significantly within CBM-A or CON ($ps > .29$).

PSWQ scores for CBM-A and CON did not differ at postassessment ($t(279.6) = -0.08$, $p = .94$, $d = -0.01$) or at 2-month follow-up ($t(280.1) = 0.02$, $p = .99$, $d = 0.002$). Likewise, WL scores did not significantly differ, at postassessment, from CBM-A scores ($t(35.5) = -1.69$, $p = .10$, $d = -0.35$) or from CON scores ($t(33.7) = -1.50$, $p = .14$, $d = -0.35$).

DASS-Depression

On the DASS-Depression, scores at preassessment did not significantly differ among the groups ($F(2, 69.9) = 1.43$, $p = .25$). The 3 Group × 5 Time mixed ANOVA yielded a main effect of time ($F(3.3, 308) = 9.84$, $p < .001$, $\eta_p^2 = .030$) and a main effect of group ($F(2, 323) = 3.85$, $p = .022$, $\eta_p^2 = .023$). The Group × Time interaction was not significant ($F(6.6, 308) = 1.28$, $p = .26$, $\eta_p^2 = .008$).

Because we had a priori hypotheses requiring group comparisons, we analyzed pre–post change scores despite the nonsignificant Group × Time interaction effect. Scores decreased from pre to postassessment in both training groups (CBM-A, $t(157) = -6.45$, $p < .001$, $d = -0.45$; CON, $t(140) = -5.82$, $p < .001$, $d = -0.37$), but not significantly in the WL group ($t(26) = -0.41$, $p = .69$, $d = -0.05$). Declines in CBM-A and CON did not differ ($t(296.7) = -0.57$, $p = .57$, $d = -0.065$), but declines were greater in CBM-A versus WL ($t(36.6) = -2.21$, $p = .034$, $d = -0.44$). There was a trend of a greater decline in CON versus WL ($t(36.0) = -1.90$, $p = .066$, $d = -0.41$). A contrast on pre–post change scores with weights coded as 1 (CBM-A), 1 (CON), and −2 (WL) showed that the two training groups showed greater declines than did WL ($t(31.0) = -2.14$, $p = .040$). From postassessment to 2-month follow-up, scores did not change significantly within CBM-A or CON ($ps > .88$).

DASS-Depression scores for CBM-A and CON did not differ at postassessment ($t(290.7) = 0.36$, $p = .72$, $d = 0.04$) or at 2-month follow-up ($t(291.7) = 0.46$, $p = .64$, $d = 0.05$). However, CBM-A and WL scores differed at postassessment ($t(33.6) = -2.72$, $p = .010$, $d = -0.61$), as did CON and WL scores ($t(34.6) = -2.35$, $p = .025$, $d = -0.53$).

Smartphone-Delivered Dot-Probe Attention Bias Assessments

Data Reduction

We performed data reduction on all dot-probe attention bias assessments collected, including participants whose SIAS preassessment scores fell below the cutoff needed for inclusion in the ITT sample. The 429 participants who began the study completed $M = 5.37$ assessments out of the six possible. The days of assessment varied, as participants completed assessments the first time they visited the training website after a given assessment day.

We eliminated outlier reaction times through several steps, defined a priori before we examined the results. From 334,480 trials, we removed inaccurate responses (3.51 %), then responses under 200 ms (0.6 % of accurate trials) or above 1,500 ms (2.4 % of accurate trials), and, finally, responses more than two standard deviations below (0.16 % of the remainder) or above (4.3 % of the remainder) the mean response time (calculated ideographically within individual sessions). After these steps, 90.1 % of responses remained. We also eliminated and treated as missing any sessions in which fewer than 75 % of trials remained after outlier removal, resulting in the elimination of 11.2 % of practice/preassessment sessions, for which responses tended to be more variable (likely due to this session's instructions as practice trials and the novelty of the task), and the elimination of 2.3 % of later sessions. Within the remaining 301,287 trials considered together, the descriptive statistics were $M = 706.9$ (SD = 173.1) across all trials, $M = 701.5$ (SD = 171.2) for neutral trials, and $M = 706.1$ (SD = 170.7) for threat trials.

Effects of Training Group on Bias Scores

In the multilevel linear modeling analyses tailored to the present study's data, we used three models to test our hypotheses. In Model 1, time was the sole predictor, and it showed a significant decrease in scores over time, in CBM-A and CON combined ($b = -1.50$, $t(980) = -3.11$, $p = .0019$). To explore this effect, we constructed Model 1-CBM-A and Model 1-CON, using only data from CBM-A and CON groups, respectively. The predictor coefficients from these models showed a significant decrease over time within CBM-A ($b = -1.82$, $t(516) = -2.59$, $p = .0098$), and a similar, nonsignificant trend within CON ($b = -1.12$, $t(463) = -1.72$, $p = .087$).

To assess whether training group (CBM-A vs. CON) affected bias scores, we then created Model 2, identical to the first, but with the group main effect included as a predictor. Based on a likelihood ratio test, Model 2 showed a significantly better fit to the data than Model 1 ($\chi^2(1) = 4.42$, $p = .036$), with group as a significant predictor ($b = 4.21$, $t(294) = 2.11$, $p = .036$), showing the group main effect, that scores significantly differed between groups, considering all time points during and at the end of training. To test the direction of the effect, we examined the coefficient for the group effect in the second model. Its value ($b = 4.21$, with CBM-A as the reference group) reflected a model-estimated mean 4.21 ms higher for the CON group compared to CBM-A, thus confirming that training acted in the intended direction.

To test the interaction effect, addressing the hypothesis that scores would increasingly diverge between groups over time, we created Model 3, identical to Model 2, but with the Group × Time interaction included as a predictor. Based on a likelihood ratio test, Model 3 did not show a significantly better fit than Model 2 ($\chi^2(1) = 0.53$, $p = .47$), indicating that there was no significant interaction effect. Taken together, the nonsignificant interaction and the significant group main effect suggest that scores diverged (with a small effect) very soon after training commenced, but did not diverge further throughout the remainder of training.

For traditional analyses, we used the preassessment/practice score, the Post Mean score, and LOCF for missing data. We conducted a two Group (between-subjects factor; CBM-A, CON) × 2 Time (within-subjects factor; Preassessment, Post Mean) mixed ANOVA. The main effect of time was significant ($F(1, 266) = 6.65$, $p = .010$,

$\eta_p^2 = .024$), indicating that scores decreased over time (collapsing across groups). The Group × Time interaction was nonsignificant ($F(1, 266) = 0.27$, $p = .60$, $\eta_p^2 = .001$), indicating that the degree of change in scores over time did not depend on group.

In between-groups t tests, at preassessment, scores for CBM-A and CON did not differ in the ITT sample ($t(239.6) = -0.11$, $p = .91$, $d = -0.013$). The Post Mean scores comparison fell short of significance ($t(294.0) = -1.77$, $p = .078$, $d = -0.20$), although a one-tailed version of the test indicated lower scores in CBM-A than in CON ($p = .039$). Additionally, a two-tailed test on protocol completers (thus obviating LOCF) was significant ($t(211.0) = -2.86$, $p = .0046$, $d = -0.37$).

In within-groups t tests, preassessment scores did not differ from zero in either CBM-A ($t(142) = -0.51$, $p = .61$, $d = -0.042$) or CON ($t(125) = -0.52$, $p = .61$, $d = -0.046$), but Post Mean scores were significantly lower than zero in both CBM-A ($t(155) = -4.58$, $p < .001$, $d = -0.37$) and CON ($t(139) = -2.24$, $p = .026$, $d = -0.19$). Change scores (Post Mean minus preassessment) revealed significant decreases in bias scores from pre to post within the CBM-A group ($t(141) = -2.31$, $p = .022$, $d = -0.19$) but not CON ($t(125) = -1.38$, $p = .17$, $d = -0.12$).

See Fig. 4 for a plot of means and standard errors of bias scores.

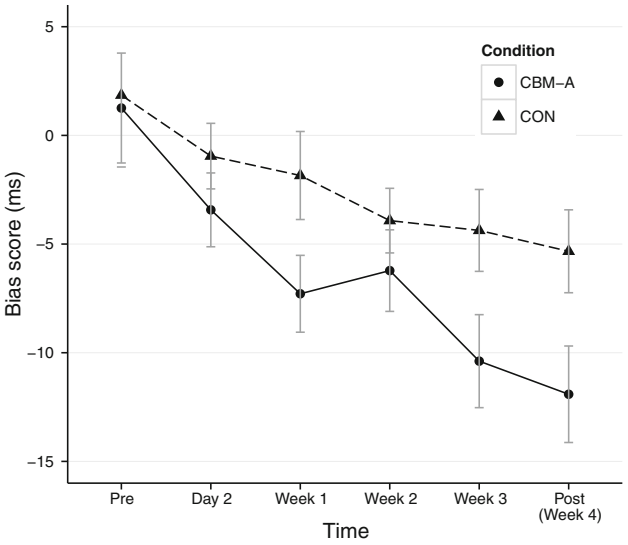

Fig. 4 Smartphone-delivered dot-probe attention bias assessment scores. This figure depicts the means and standard errors. For missing data, we used a multilevel linear model (described as Model 3 in the "Results" section) for post hoc prediction of each value

PC-Delivered Modified Posner Cueing Task Attention Bias Assessments

Data Reduction

We eliminated outlier reaction times through several steps, using a similar trimming process as for data from the smartphone-delivered dot-probe task, adjusted for the faster responses observed on this task. Of the 429 participants who began the study (disregarding the SIAS cutoff criterion needed for inclusion in the ITT sample), 62 visited the laboratory to complete a modified Posner cueing task assessment on the computer. There were $n = 62$ preassessment sessions, from Day 1 of training, and $n = 53$ postassessment sessions from Week 4 of training. From 33,298 trials, we removed inaccurate responses (1.49 %), then responses under 100 ms (0.80 % of accurate trials) or above 1,200 ms (0.28 % of accurate trials), and, finally, responses more than two standard deviations below (1.03 % of the remainder) or above (4.1 % of the remainder) the mean response time (calculated ideographically within individual sessions). After these steps, 94.88 % of responses remained. No session had less than 82 % of trials remaining after trimming, thus we did not eliminate any sessions due to the quality of responses. Within the remaining 30,787 trials considered together, the mean and standard deviation across all trials were $M = 391.3$ (SD = 98.5).

Effects of Training Group on Bias Scores

First, we performed analyses of session data to confirm that spatial cueing had the intended effect on response times, namely, that validly cued trials prompted faster responses than invalidly cued trials. Collapsing across all three word types, we calculated each session's mean response time for validly cued trials, for invalidly cued trials, and for uncued trials. A one-way ANOVA on the three trial types was significant for pre ($F(2, 83.5) = 12.32$, $p < .001$) and postassessment ($F(2,71.3) = 16.68$, $p < .001$). As expected, mean responses to invalidly cued trials were slower than to validly cued trials at pre ($t(82.8) = 3.67$, $p < .001$, $d = 0.79$) and postassessment ($t(70.1) = 4.31$, $p < .001$, $d = 1.00$). Mean responses to uncued trials were slower than to valid trials at pre ($t(81.6) = 4.62$, $p < .001$, $d = 1.00$) and postassessment ($t(68.9) = 5.30$, $p < .001$, $d = 1.23$), but their mean response times did not significantly differ from invalidly cued trials at pre ($t(83.7) = 1.00$, $p = .32$, $d = 0.22$) or postassessment ($t(71.8) = 1.04$, $p = .30$, $d = 0.24$), though they were nominally slower (by 17.1 ms at preassessment and 15.3 ms at postassessment).

Forty-three participants included in our ITT analyses visited our laboratory and completed the modified Posner cueing task. None of these individuals dropped out of the study. At preassessment, scores did not significantly differ

from zero in either group (CBM-A: $t(20) = -0.72$, $p = .48$, $d = -0.16$); CON: $t(21) = 1.07$, $p = .30$, $d = 0.23$), nor did they significantly differ between CBM-A and CON ($t(40.1) = -1.28$, $p = .21$, $d = -0.39$). A comparison of pre–post change scores between groups revealed no significant differences in change in bias scores over time (CBM-A vs. CON: $t(33.2) = 1.40$, $p = .17$, $d = 0.46$). The nominal difference in change scores was in the opposite direction than the intended effect of training, as CBM-A group scores nominally increased, whereas CON group scores nominally decreased.

Discussion

In a double-blind RCT, we tested the effects of CBM-A training, CON training, and waitlist on participants' symptoms of social anxiety, worry, and depression, while also assessing attention bias with a remote smartphone-delivered dot-probe task and an in-laboratory PC-delivered modified Posner cueing task. We demonstrated the feasibility of delivering CBM-A via smartphones in short, frequent sessions, as well as the feasibility of conducting a relatively large, low-cost, minimal contact, web-based RCT. The internal consistency and test–retest reliability findings for the dot-probe assessments demonstrated that smartphones are viable for reaction time task assessments. We found no greater symptom declines in CBM-A than in CON on measures of social anxiety, worry, and depression, although both CBM-A and CON training groups showed significantly greater symptom declines than did WL on measures of social anxiety. The effects of CBM-A versus CON training on attention yielded only small differences in dot-probe attention bias scores and no significant differences in modified Posner cueing task scores.

Our primary test was of whether CBM-A would reduce social anxiety more than would CON. It did not: Symptom change in the two groups showed a very similar pattern across all time points. Several RCTs testing similar CBM-A and control protocols with individuals diagnosed with social anxiety disorder have likewise found no significant differences in symptom change between conditions. Four studies employed home-based PC training (Boettcher et al. 2012, 2013; Carlbring et al. 2012; Neubauer et al. 2013), and two used laboratory-based PC training (Bunnell et al. 2013; Sawyer et al. 2012). Given the number of studies and their larger sample sizes finding no significant differences, the evidence now suggests that the benefits of CBM-A over control training for treating social anxiety are very limited or highly inconsistent. Indeed, another multi-session, lab-based study revealed indistinguishably significant reductions for CBM-A and CON groups on self-report, behavioral, and physiological measures of anxiety in participants who feared public speaking (McNally et al. 2013).

However, we did find that both CBM-A and CON training reduced social anxiety more than did WL. Furthermore, the decreased symptom levels for CBM-A and CON were stable from postassessment through follow-up. What caused these reductions? Perhaps the data permit only one firm conclusion: The active ingredient was *not* the contingency of probe placement, as the no-contingency CON was as helpful as CBM-A. Medication treatments involve placebo effects, and researchers label some beneficial aspects of psychotherapy as "nonspecific factors." The present study's CBM-A and CON participants presumably benefitted from some nonspecific effects; however, such effects are not easily demarcated from specific ones. For example, if, hypothetically, one's use of *any* distracting mobile app when feeling anxious were beneficial, then we could define it to be either a specific factor for mobile app treatments or a nonspecific factor for CBM-A or CON training compared with other mobile app treatments. Perhaps active use of any computerized treatment tool, such as a mobile app treatment, could instill participants with confidence that their social anxiety will improve. Bolstered confidence, in turn, may foster greater comfort in social interaction, thereby attenuating social anxiety. On the other hand, previous laboratory studies did find superiority for CBM-A over CON training. By employing several comparison conditions to isolate treatment factors, future investigations may clarify which specific aspects are clinically useful and, importantly, replicable in studies of computerized cognitive training.

In this study, the effects of training on CBM-A compared to CON dot-probe attention bias scores were significant when we employed data analytic methods tailored to the study's five assessments during and after training, addressing missing data via multilevel linear modeling. Yet, the effects of training on dot-probe bias scores were small, and some traditional statistical analyses revealed a nonsignificant group by time interaction. The modified Posner cueing task detected no attentional effects of training. Thus, the training conditions may not have had the intended impact on attention bias. It is possible that weekly dot-probe assessment could have interfered with training in the CBM-A group, though the training sessions were far more frequent than assessment sessions.

Pragmatically, any factors that confer symptom reduction are important. Hence, elements common to CBM-A and CON may warrant further investigation. They may reduce anxiety, but not necessarily through reducing attention bias for threat. The magnitude of the LSAS-SR decline in the CON group was substantial, alongside the decline in the CBM-A group. Declines in LSAS and LSAS-SR scores have varied widely across CBM-A RCTs, and

Table 3 LSAS/LSAS-SR scores in CBM-A trials targeting social anxiety symptoms (or disorder)

	CBM-A				CON				Difference in change
	n	Pre	Post	Change	*n*	Pre	Post	Change	
Amir et al. (2009b)	22	74.5	46.1	−28.4	22	68.1	60.0	−8.1	−20.3
Schmidt et al. (2009)	18	80.8	68.5	−12.3	18	80.7	78.0	−2.6	−9.6
Boettcher et al. (2012)	33	83.1	64.7	−18.4	35	80.5	64.6	−15.9	−2.5
Carlbring et al. (2012)	40	73.8	66.0	−7.8	39	73.0	60.5	−12.5	4.7
Heeren et al. (2012)[a]	20	82.1	61.0	−21.1	18	79.5	62.9	−16.6	−4.5
Neubauer et al. (2013)	30	69.9	65.8	−4.0	29	63.4	65.6	2.2	−6.2
Sawyer et al. (2012)	15	72.8	61.9	−10.9	16	79.6	65.0	−14.6	3.7
Boettcher et al. (2013)	43	74.7	62.1	−12.6	43	73.2	57.6	−15.6	3.0
Bunnell et al. (2013)	15	86.7	59.9	−26.7	16	76.8	66.4	−10.4	−16.3
Enock & McNally (2010)	16	47.6	35.4	−12.2	NA	NA	NA	NA	NA
Present study, ITT analysis	158	74.2	57.6	−16.6	141	72.8	58.2	−14.6	−2.0
Present study, protocol completers only	120	73.4	53.9	−19.5	104	73.0	56.1	−16.9	−2.6

Scores are from the LSAS-clinician administered version (Amir et al. 2009b; Sawyer et al. 2012; Schmidt et al. 2009) or LSAS-SR (all other studies) in trials employing dot-probe CBM-A tasks with socially anxious individuals. Scores reflect group means at preassessment and postassessment, as well as the pre–post change. The column labeled "difference in change" shows the difference between LSAS change in CBM-A versus in CON. There are various differences across studies in their procedures and handling of missing data

[a] In Heeren et al. (2012), due to the brief training period, the follow-up scores (CBM-A, 51.1; CON, 71.2; difference in change, −22.7), are highly relevant, as well as the postassessment scores shown below for consistency with the table columns

the means from other trials and the present one are shown in Table 3, for comparison. Our study alone has a waitlist group, essential for evaluating symptom decline without treatment. Future reviews, meta-analyses, and experiments should probe whether there are moderators of the treatment effects of CBM-A, control training, and waitlist. For example, participants may harbor differing expectations. They may commence training with the belief that the procedure is merely experimental and unproven as effective, or they may commence training with the perspective that the procedure is a high-tech, powerful new treatment. How clinical researchers recruit participants may shape the perceptions of those enrolled in the study. In our study, participants were treatment-seekers, motivated to attempt a novel treatment to better their condition, receiving no financial compensation, as in some other CBM-A studies. As another contextual issue, perhaps in-laboratory situations affect participants' anxiety and attentional bias in contrast to remote-delivery situations. Any of these factors may mobilize beneficial nonspecific effects.

The reliability data for the attention bias assessment tasks were informative. We found that attention bias scores from the modified Posner cueing task were unreliable even though the cue validity scores across all word types were highly reliable. This highlights the challenges of measuring attention bias. It is unclear what factors may cause individual differences in attention bias to emerge and then to be detectable or not in different studies, but one issue may be the tendency of attention bias scores to shift towards threat-avoidance when a sample is exposed to acute stress (e.g., Wald et al. 2011). Several reliability estimates had negative values. Exploration of this phenomenon is beyond the scope of this article, but this issue has seemingly surfaced in past studies of reliability (e.g., Schmukle 2005; Staugaard 2009) where some negative reliability estimates emerged, and the problem exists for Cronbach's alpha as well as split-half reliability.

The finding of substantial dot-probe attention bias score reliability in later weeks demonstrates that smartphones are capable of delivering reliable reaction-time based assessments that are at least as good as those administered via PC (Ataya et al. 2012; Browning et al. 2011; Schmukle 2005; Staugaard 2009; Waechter et al. 2013). Unfortunately, few dot-probe studies have reported reliability estimates. The small differences in bias scores between training groups, in the context of high test–retest reliability from Week 3 to 4 ($r(215) = .63$, $p < .001$, missing data excluded), suggests that participants may not respond to training as predictably as researchers might expect. The reliability suggests consistency across the two sessions: Participants were likely to maintain similar scores at these two times. Hence, their differential responses to the task's threat and neutral trials maintained a similar pattern. However, for Week 4 alone, CBM-A group scores showed only a small, nonsignificant difference compared to CON. Despite having performed approximately 4,000 trials with probes appearing in the location of the same eight neutral face stimuli, CBM-A participants' responses did not reflect optimal detection and rapid response to probes in their predictable locations. People may not behave so uniformly and conveniently as to

optimize their reaction time task performance based on contingencies within long bouts of repetitive training.

Our study leveraged advantages of web-based research and treatment. In contrast to traditional RCTs with in-clinic treatment and clinician assessments, requiring extensive resources and staff, our trial needed only one doctoral student and two research assistants to run hundreds of volunteers. The resulting sample size exceeded that of any published trial concerning cognitive bias modification. In total, participants performed over 20,000 training sessions and tapped on their screens approximately 2 million times, as we recorded their reaction times remotely. The training sessions were more evenly distributed over time than in previous studies, and the amount of training overall was high. Participants who completed the study performed an average of 3,450 critical training trials, a greater number than other 4-week RCTs of CBM-A to reduce social anxiety: 1,024 critical trials across eight sessions were in training protocols for Amir et al. (2009b), Boettcher et al. (2012), Bunnell et al. (2013), Carlbring et al. (2012), Neubauer et al. (2013), Sawyer et al. (2012), and Schmidt et al. (2009). The amount in the present study was comparable to more concentrated training designs (Li et al. 2008, used 3,360 critical trials across seven sessions; Heeren et al. 2012, used 2,976 across four sessions) but the smartphone-delivered sessions were spread across numerous sessions (54, on average, in protocol completers). Web-based research methods hold great potential for large, inexpensive trials, with frequent treatment sessions. Such methods facilitate high statistical power and fidelity of treatment delivery, which may lead to an advantageous cycle of treatment development, testing, and clinical use (Enock and McNally 2013).

Our study has limitations. We did not conduct formal psychiatric diagnostic assessments of any mental disorders, and assessments were confined to questionnaires. Symptom scores based on participants' responses suggest that most would be diagnosable with social anxiety disorder, given the proportion exceeding diagnostic screening cutoffs on the LSAS-SR. Also, participants had no incentive to exaggerate symptoms, as there was no financial compensation, and we informed recruits that symptoms were not required for enrollment. Still, the lack of diagnostic interview is an important limitation of this experimental design, a choice driven chiefly by feasibility concerns. The fact that participants were seekers of self-directed treatment makes the study clinically relevant with respect to this population, as many people seek such treatment irrespective of any diagnosis. The waitlist group was small as we added 9 months after launching the study, but its inclusion provided a vital baseline for assessing change in the two training groups.

The dot-probe task for training and assessment appeared on smartphones' small screens, and we did not attempt to control participants' viewing distance. Distance from eyes to screen, stimulus size, and visual angles have varied widely in computer-based studies, so this concern is not unique to the present study. One potential concern introduced by the smartphone adaptation of the task is the appearance of the E and F response buttons on the sides of the screen, simultaneously appearing with the letter probe. This arrangement could affect participants' attention in unknown ways. Based on the instructions, participants' thumbs may have partially covered the on-screen response buttons.

Dropout rates from the ITT sample were substantial in both CBM-A and CON training groups, whereas there were no dropouts from WL. There could be many reasons for this, but this difference suggests the need for increased tolerability of training. Our attempt to increase tolerability for this study consistent of making sessions shorter and more accessible, to allow participants to work the training into their daily lives less obtrusively than long blocks of training. Future efforts to increase the tolerability, or even the pleasure, of training methods could reduce dropout.

In conclusion, smartphones can deliver frequent cognitive training sessions. Since CBM-A and control training may reduce symptoms equally more than waitlist participation, these methods warrant further experimental testing to isolate active ingredients and evaluate their merit for clinical use by the public.

Acknowledgments The authors thank Imke Vonk, Linh Vu, and Jennifer Yu for their assistance, primarily in data collection.

Conflict of Interest Philip M. Enock, Stefan G. Hofmann and Richard J. McNally declare that they have no conflict of interest.

Informed Consent All participants signed up through the website, which described the study in an easy-to-read format (Online Supplement 1). They electronically provided consent after going through these pages and a detailed consent form. All procedures were approved by Harvard's Committee on the Use of Human Subjects in Research.

Animal Rights No animal studies were carried out by the authors for this article.

References

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

Amir, N., Beard, C., Taylor, C. T., Klumpp, H., Elias, J., Burns, M., et al. (2009b). Attention training in individuals with generalized social phobia: A randomized controlled trial. *Journal of*

Consulting and Clinical Psychology, 77(5), 961–973. doi:10.1037/a0016685.

Amir, N., Elias, J., Klumpp, H., & Przeworski, A. (2003). Attentional bias to threat in social phobia: Facilitated processing of threat or difficulty disengaging attention from threat? *Behaviour Research and Therapy, 41*(11), 1325–1335. doi:10.1016/S0005-7967(03)00039-1.

Amir, N., Weber, G., Beard, C., Bomyea, J., & Taylor, C. T. (2008). The effect of a single-session attention modification program on response to a public-speaking challenge in socially anxious individuals. *Journal of Abnormal Psychology, 117*(4), 860–868. doi:10.1037/a0013445.

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment, 10*(2), 176–181. doi:10.1037/1040-3590.10.2.176.

Ataya, A. F., Adams, S., Mullings, E., Cooper, R. M., Attwood, A. S., & Munafò, M. R. (2012). Internal reliability of measures of substance-related cognitive bias. *Drug and Alcohol Dependence, 121*(1–2), 148–151. doi:10.1016/j.drugalcdep.2011.08.023.

Baker, S. L., Heinrichs, N., Kim, H.-J., & Hofmann, S. G. (2002). The Liebowitz social anxiety scale as a self-report instrument: A preliminary psychometric analysis. *Behaviour Research and Therapy, 40*(6), 701–715. doi:10.1016/S0005-7967(01)00060-2.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133*(1), 1–24. doi:10.1037/0033-2909.133.1.1.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Allyn and Bacon.

Beard, C., Sawyer, A. T., & Hofmann, S. G. (2012). Efficacy of attention bias modification using threat and appetitive stimuli: A meta-analytic review. *Behavior Therapy, 43*(4), 724–740. doi:10.1016/j.beth.2012.01.002.

Boettcher, J., Berger, T., & Renneberg, B. (2012). Internet-based attention training for social anxiety: A randomized controlled trial. *Cognitive Therapy and Research, 36*(5), 522–536. doi:10.1007/s10608-011-9374-y.

Boettcher, J., Leek, L., Matson, L., Holmes, E., Browning, M., MacLeod, C., et al. (2013). Internet-based attention bias modification for social anxiety: A randomised controlled comparison of training towards negative and training towards positive cues. *PLoS ONE, 8*(9), e71760. doi:10.1371/journal.pone.0071760.

Browning, M., Grol, M., Ly, V., Goodwin, G. M., Holmes, E. A., & Harmer, C. J. (2011). Using an experimental medicine model to explore combination effects of pharmacological and cognitive interventions for depression and anxiety. *Neuropsychopharmacology, 36*(13), 2689–2697. doi:10.1038/npp.2011.159.

Bunnell, B. E., Beidel, D. C., & Mesa, F. (2013). A randomized trial of attention training for generalized social phobia: Does attention training change social behavior? *Behavior Therapy*. doi:10.1016/j.beth.2013.04.010.

Carlbring, P., Apelstrand, M., Sehlin, H., Amir, N., Rousseau, A., Hofmann, S., et al. (2012). Internet-delivered attention bias modification training in individuals with social anxiety disorder—a double blind randomized controlled trial. *BMC Psychiatry, 12*(1), 66. doi:10.1186/1471-244X-12-66.

Cox, B. J., Ross, L., Swinson, R. P., & Direnfeld, D. M. (1998). A comparison of social phobia outcome measures in cognitive-behavioral group therapy. *Behavior Modification, 22*(3), 285–297. doi:10.1177/01454455980223004.

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods, 42*(1), 205–211. doi:10.3758/BRM.42.1.205.

Eldar, S., Apter, A., Lotan, D., Edgar, K. P., Fox, N. A., Pine, D. S., et al. (2012). Attention bias modification treatment for pediatric anxiety disorders: A randomized controlled trial. *The American Journal of Psychiatry, 15*, 213–220. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491316/.

Enock, P. M., & McNally, R. J. (2010). *Feasibility and efficacy of attention bias modification* via *iPhone and other handheld devices to reduce social anxiety and worry*. Unpublished manuscript. Department of Psychology, Harvard University, Cambridge, MA.

Enock, P. M., & McNally, R. J. (2013). How mobile apps and other web-based interventions can transform psychological treatment and the treatment development cycle. *The Behavior Therapist, 36*(3), 56, 58, 60, 62–66. Retrieved from http://www.abct.org/docs/PastIssue/36n3.pdf.

Enock, P. M., Robinaugh, D. J., Reese, H. E., & McNally, R. J. (2012). *Improved reliability estimation and psychometrics of the dot-probe paradigm on smartphones and PC*. Poster session presented at the meeting of The Association of Behavioral and Cognitive Therapies, National Harbor, MD.

Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General, 130*(4), 681. doi:10.1037//0096-3445.130.4.681.

Fresco, D. M., Coles, M. E., Heimberg, R. G., Leibowitz, M. R., Hami, S., Stein, M. B., et al. (2001). The Liebowitz Social Anxiety Scale: A comparison of the psychometric properties of self-report and clinician-administered formats. *Psychological Medicine, 31*(6), 1025–1035. doi:10.1017/S0033291701004056.

Fresco, D. M., Mennin, D. S., Heimberg, R. G., & Turk, C. L. (2003). Using the Penn State Worry Questionnaire to identify individuals with generalized anxiety disorder: A receiver operating characteristic analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 34*(3–4), 283–291. doi:10.1016/j.jbtep.2003.09.001.

Gee, A. (2011). Therapist-free therapy: Cognitive-bias modification may put the psychiatrist's couch out of business. *The Economist*. Retrieved from http://www.economist.com/node/18276234.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (p. 648). New York: Cambridge University Press. Retrieved from http://www.amazon.com/Analysis-Regression-Multilevel-Hierarchical-Models/dp/052168689X.

Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J. C., Fox, N. A., Leibenluft, E., et al. (2010). Attention bias modification treatment: A meta-analysis toward the establishment of novel treatment for anxiety. *Biological Psychiatry*. doi:10.1016/j.biopsych.2010.07.021.

Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J. C., Fox, N. A., Leibenluft, E., et al. (2012). Erratum: Attention bias modification treatment: A meta-analysis toward the establishment of novel treatment for anxiety. *Biological Psychiatry, 72*(5), 429. doi:10.1016/j.biopsych.2012.07.015.

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. doi:10.1037/a0024355.

Hedman, E., Ljótsson, B., Rück, C., Furmark, T., Carlbring, P., Lindefors, N., et al. (2010). Internet administration of self-report measures commonly used in research on social anxiety disorder: A psychometric evaluation. *Computers in Human Behavior, 26*(4), 736–740. doi:10.1016/j.chb.2010.01.010.

Heeren, A., Reese, H. E., McNally, R. J., & Philippot, P. (2012). Attention training toward and away from threat in social phobia: Effects on subjective, behavioral, and physiological measures of

anxiety. *Behaviour Research and Therapy, 50*(1), 30–39. doi:10.1016/j.brat.2011.10.005.

Heimberg, R. G., Mueller, G. P., Holt, C. S., Hope, D. A., & Liebowitz, M. R. (1992). Assessment of anxiety in social interaction and being observed by others: The social interaction anxiety scale and the Social Phobia Scale. *Behavior Therapy, 23*(1), 53–73. doi:10.1016/S0005-7894(05)80308-9.

Li, S., Tan, J., Qian, M., & Liu, X. (2008). Continual training of attentional bias in social anxiety. *Behaviour Research and Therapy, 46*(8), 905–912. doi:10.1016/j.brat.2008.04.005.

Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry, 22*, 141–173.

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd ed.). Sydney: Psychology Foundation.

MacLeod, C. (1995). Training selective attention: A cognitive-experimental technique for reducing anxiety vulnerability? In *World congress of behavioural and cognitive therapies* (p. 118).

MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology, 95*(1), 15–20. doi:10.1037/0021-843X.95.1.15.

MacLeod, C., Rutherford, E., Campbell, L. W., 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, 111*(1), 107–123. doi:10.1037/0021-843X.111.1.107.

MacLeod, C., Soong, L. Y., Rutherford, E. M., & Campbell, L. W. (2007). Internet-delivered assessment and manipulation of anxiety-linked attentional bias: Validation of a free-access attentional probe software package. *Behavior Research Methods, 39*(3), 533–538. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17958165.

Matsumoto, D., & Ekman, P. (1988). *Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and Neutral Faces (JACNeuF)*. San Francisco, CA: Intercultural and Emotion Research Laboratory, Department of Psychology, San Francisco State University.

Mattick, R., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy, 36*(4), 455–470. doi:10.1016/S0005-7967(97)10031-6.

McNally, R. J., Enock, P. M., Tsai, C., & Tousian, M. (2013). Attention bias modification for reducing speech anxiety. *Behaviour Research and Therapy, 51*(12), 882–888. doi:10.1016/j.brat.2013.10.001.

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. doi:10.1016/0005-7967(90)90135-6.

Neubauer, K., von Auer, M., Murray, E., Petermann, F., Helbig-Lang, S., & Gerlach, A. L. (2013). Internet-delivered attention modification training as a treatment for social phobia: A randomized controlled trial. *Behaviour Research and Therapy, 51*(2), 87–97. doi:10.1016/j.brat.2012.10.006.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2010). *nlme: Linear and nonlinear mixed effects models*. Vienna: R Foundation for Statistical Computing. Retrieved from http://cran.r-project.org/web/packages/nlme.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna. Retrieved from http://www.r-project.org/.

Rodebaugh, T. L., Woods, C. M., & Heimberg, R. G. (2007). The reverse of social anxiety is not always the opposite: The reverse-scored items of the social interaction anxiety scale do not belong. *Behavior Therapy, 38*(2), 192–206. doi:10.1016/j.beth.2006.08.001.

Rytwinski, N. K., Fresco, D. M., Heimberg, R. G., Coles, M. E., Liebowitz, M. R., & Cissell, S. (2009). Screening for social anxiety disorder with the self-report version of the Liebowitz Social Anxiety Scale. *Depression and Anxiety, 26*(1), 34–38. doi:10.1002/da.20503.

Sawyer, A. T., Whitfield-Gabrieli, S., Gabrieli, J. D. E., Amir, N., Fang, A., Richey, A., et al. (2012). Attention retraining in SAD: Treatment outcome and neurobiological correlates. In A. Asnaani & A. T. Sawyer (Chairs), *Modifying cognitive biases: Emerging data on applications and effects on clinical disorders*. Symposium conducted at the meeting of The Association of Behavioral and Cognitive Therapies, National Harbor, MD.

Schmidt, N. B., Richey, J. A., Buckner, J. D., & Timpano, K. R. (2009). Attention training for generalized social anxiety disorder. *Journal of Abnormal Psychology, 118*(1), 5–14. doi:10.1037/a0013643.

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality, 19*(7), 595–605. doi:10.1002/per.554.

SPSS Inc. (2009). PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc. http://www-01.ibm.com/support/docview.wss?uid=swg21476197

Staugaard, S. R. (2009). Reliability of two versions of the dot-probe task using photographic faces. *Psychology Science Quarterly, 51*(3), 339–350.

Ulrich, R., & Giray, M. (1989). Time resolution of clocks: Effects on reaction time measurement-Good news for bad clocks. *British Journal of Mathematical and Statistical Psychology, 42*(1), 1–12. doi:10.1111/j.2044-8317.1989.tb01111.x.

Waechter, S., Nelson, A. L., Wright, C., Hyatt, A., & Oakman, J. (2013). Measuring attentional bias to threat: Reliability of dot probe and eye movement indices. *Cognitive Therapy and Research*. doi:10.1007/s10608-013-9588-2.

Wald, I., Lubin, G., Holoshitz, Y., Muller, D., Fruchter, E., Pine, D. S., et al. (2011). Battlefield-like stress following simulated combat and suppression of attention bias to threat. *Psychological Medicine, 41*(4), 699–707. doi:10.1017/S0033291710002308.

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology, 33*(4), 448–457. doi:10.1037/h0027806.

Wells, T., & Beevers, C. G. (2010). Biased attention and dysphoria: Manipulating selective attention reduces subsequent depressive symptoms. *Cognition and Emotion, 24*(4), 719–728. doi:10.1080/02699930802652388.

Zlomke, K. R. (2009). Psychometric properties of internet administered versions of Penn State Worry Questionnaire (PSWQ) and Depression, Anxiety, and Stress Scale (DASS). *Computers in Human Behavior, 25*(4), 841–843. doi:10.1016/j.chb.2008.06.003.