HHS Public Access

Author manuscript

Psychol Assess. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

Psychol Assess. 2015 June ; 27(2): 365–376. doi:10.1037/pas0000036.

Empirical Recommendations for Improving the Stability of the Dot-Probe Task in Clinical Research

Rebecca B. Price, Ph.D.[1], **Jennie M. Kuckertz, B.A.**[2], **Greg J. Siegle, Ph.D.**[1], **Cecile D. Ladouceur, Ph.D.**[1], **Jennifer S. Silk, Ph.D.**[1], **Neal D. Ryan, M.D.**[1], **Ronald E. Dahl, M.D.**[1,3], and **Nader Amir, Ph.D.**[2]

[1]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA

[2]Joint Doctoral Program in Clinical Psychology, San Diego State University/University of California, San Diego, San Diego, CA

[3]School of Public Health, University of California, Berkeley, Berkeley, CA

Abstract

The dot-probe task has been widely used in research to produce an index of biased attention based on reaction times (RTs). Despite its popularity, very few published studies have examined psychometric properties of the task, including test-retest reliability, and no previous study has examined reliability in clinically anxious samples or systematically explored the effects of task design and analysis decisions on reliability. In the current analysis, we utilized dot-probe data from three studies where attention bias towards threat-related faces was assessed at multiple (≥5) timepoints. Two of the studies were similar (adults with Social Anxiety Disorder, similar design features) while one was much more disparate (pediatric healthy volunteers, distinct task design). We explored the effects of analysis choices (e.g., bias score calculation formula, methods for outlier handling) on reliability and searched for convergence of findings across the three studies. We found that, when considering the three studies concurrently, the most reliable RT bias index utilized data from dot-bottom trials, comparing congruent to incongruent trials, with rescaled outliers, particularly after averaging across more than one assessment point. Although reliability of RT bias indices was moderate to low under most circumstances, within-session *variability* in bias (attention bias variability; ABV), a recently proposed RT index, was more reliable across sessions. Several eyetracking-based indices of attention bias (available in the pediatric healthy sample only) showed reliability that matched the optimal RT index (ABV). On the basis of these findings, we make specific recommendations to researchers using the dot probe, particularly those wishing to investigate individual differences and/or single-patient applications.

Keywords

stability; reliability; attentional bias; dot-probe task; psychometric

Correspondence concerning this article should be addressed to Rebecca Price, Western Psychiatric Institute and Clinic, 3811 O'Hara St., Pittsburgh, PA 15213, Phone: 412-383-5443, Fax: 412-383-5426, rebecca.price@stanfordalumni.org..

Attention bias has been the topic of considerable research efforts over the past 30 years. The tendency to selectively allocate attention towards particular types of information (e.g., negatively valenced, threat-related, or disorder-relevant information) has been demonstrated repeatedly across a wide range of clinical populations, with a particularly large literature in anxiety disorders (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van IJzendoorn, 2007). Interest in the topic has only accelerated in the past decade, with the advent of methods to experimentally manipulate and mechanistically intervene on attention bias as a method for reducing clinical symptoms (for a review see Beard, Sawyer, & Hofmann, 2012).

One of the most widely used research tools for objective measurement of attention bias is the visual dot-probe task (MacLeod, Mathews, & Tata, 1986), which has been used in hundreds of studies to date (e.g., as of 10/2014, PubMed database search of keyword 'dot-probe' retrieves 377 studies). The task yields a reaction time index of attention bias and is one of the most widely used performance-based measures in clinical research, with a particularly well-established history in anxiety (Bar-Haim et al., 2007) and depression research (Gotlib, McLachlan, & Katz, 1988), and growing applications in a wider range of disorders as well (e.g., Cardi, Di Matteo, Corfield, & Treasure, 2013; Forestell, Dickter, & Young, 2012). Despite an extensive literature utilizing the dot-probe task as an experimental paradigm, very little research has investigated the psychometric properties of this task. However, two published studies in healthy samples (Schmukle, 2005; Staugaard, 2009), as well as studies in substance users (Ataya et al., 2012), chronic pain patients (Dear, Sharpe, Nicholas, & Refshauge, 2011), and undergraduates with high and low scores on a social phobia scale (Waechter, Nelson, Wright, Hyatt, & Oakman, In press), have called the reliability of this task into question.

As reliability sets a theoretical upper limit on the task's validity (i.e., its ability to covary with and/or predict other outcomes), this issue is of critical importance. However, reliability and validity are not synonymous, and cases may exist where a less reliable measure is more valid than a more reliable measure (e.g., height is a reliable measure but perhaps not a valid indicator of anxiety, whereas the dot-probe may be a more valid index in spite of having less reliability). A large literature supports the validity of the dot-probe task in distinguishing between groups (e.g., anxious from non-anxious samples; Bar-Haim et al., 2007). However, there are numerous published exceptions (e.g., Mohlman, Price, & Vietri, 2013; Price et al., 2013; Waters, Lipp, & Spence, 2004), as well as issues related to the "file-drawer" problem (e.g., positive studies are more likely to be published than null findings). Sub-optimal reliability of the measure may contribute substantially to such inconsistencies (and concomitant waste of research resources). Recently, the issue of reliability has become even more critical given growing interest in applying the task to answer clinical questions. For instance, researchers are often interested in whether attention bias changes in response to specific treatments (e.g., Amir, Beard, Taylor, et al., 2009; Waters, Wharton, Zimmer-Gembeck, & Craske, 2008). However, if test-retest reliability for a specific index of bias has not been established, statistical tests for such changes in bias may be invalid. Furthermore, recent studies suggest individual differences in performance on the dot-probe task can predict outcome to specific interventions (Amir, Taylor, & Donohue, 2011; Legerstee et al., 2009; Price, Tone, & Anderson, 2011; Waters, Mogg, & Bradley, 2012). These findings indicate that the task could be useful in making personalized treatment prescriptions for

individual patients, but only if the reliability of an individual assessment is sufficient to allow for accurate prediction.

Although reliability of the dot-probe has been low and non-significant in limited extant reports, reliability may vary as a function of a) sample characteristics (e.g. clinical vs. healthy populations, age group) and b) specific task design and analysis decisions made by the experimenter. In spite of the task's popularity, no previous study has examined the impact of these study parameters in order to provide an empirical basis for improving reliability of the task. As a result, experimenters are often left in the dark and must make relatively arbitrary design and analysis decisions.

By way of illustration, one decision facing dot-probe researchers pertains to the formula used to assess attention bias. The dot-probe task (Figure 1) presents pairs of stimuli (e.g., threat-related and neutral words or pictures). One stimulus is then replaced by a probe requiring participants to indicate a response (e.g., dots in horizontal or vertical arrangement). A 'bias score' in reaction times is then calculated and used to infer information about preferential allocation of attention to one type of stimulus over another. The most widely-used formula for bias score calculation, proposed by Mathews & MacLeod (MacLeod et al., 1986), contrasts 'threat incongruent' trials, in which the dot replaces the neutral item in a neutral/non-neutral pair, with 'threat congruent' trials, in which the dot replaces the threat-related item in a neutral/non-neutral pair. Increased scores on this measure indicate that either a) attention was more readily oriented towards non-neutral items (which would speed responses to congruent trials) and/or b) disengagement of attention from non-neutral items was more difficult (which would slow responses to incongruent trials). A more recent proposal by Koster and colleagues (Koster, Crombez, Verschuere, & De Houwer, 2004) suggests that attentional bias can also be indexed by comparing reaction times across 'incongruent' neutral/non-neutral trials and trials presenting neutral/neutral pairs of stimuli. This alternate index may specifically measure difficulties with disengagement from non-neutral information, which is required on incongruent trials but not on neutral/neutral trials. Initial findings suggest that the new index, like the classic index, is relevant to anxiety symptoms and outcomes (Amir et al., 2011; Klumpp & Amir, 2009; Salemink, van den Hout, & Kindt, 2007). If one of these two indices provides more reliable information than the other, there would be important implications for both the design (which trial types should be included?) and the analysis (which bias score formula should be used?) of dot-probe studies.

Notably, both of these index calculation methods presume that a stable, trait-like bias towards (or away) from emotional stimuli exists as an individual differences variable that can be accurately summarized in a single dot-probe score. However, contextual factors such as mood state can also influence dot-probe performance (e.g., Bar-Haim et al., 2007; Broadbent & Broadbent, 1988). Furthermore, attention bias may diminish over the course of an experiment (Amir, Najmi, & Morrison, 2009). Such effects might lead to expected, theoretically warranted fluctuations in attention bias indices both within and between sessions, setting a necessary ceiling on the reliability of summative measures. An Attention Bias Variability (ABV) index has recently been proposed to explicitly quantify intrasession variability in attention bias (Iacoviello et al., 2014), providing a possible marker of

attentional control impairment and inconsistency of response when faced with emotional stimuli (e.g., fluctuations between vigilance and avoidance). This index has been validated in two experiments (Iacoviello et al., 2014), where it was shown to be elevated 1) in individuals with post-traumatic stress disorder (PTSD) when compared to both trauma-exposed and unexposed controls and 2) following combat deployment in soldiers who developed significant PTSD and/or depression symptoms, compared to those who did not. Because this index explicitly takes into account within-subject variability occurring within the task, its stability across sessions might exceed that of the more widely-used attention bias indices, which collapse all trials to create a single summary measure.

As a second example, dot-probe researchers must also decide how to identify and handle outlying RT values (e.g., due to distraction, premature responding, delays in response selection, or failure of equipment to record responses). The dot-probe literature suggests the most common approach is to set discretionary RT cut-points *a priori* based around expectations of what a typical, valid response window might look like for a given population (e.g., 100-2500ms) and to eliminate from analysis all trials falling outside of these boundaries. However, the influence of these relatively arbitrary cut-points has not been systematically assessed and may vary from study to study and/or from sample to sample. A distinct strategy, advocated by statisticians as a modern robust method for data analysis (Erceg-Hurn & Mirosevich, 2008), is to rescale (Winsorize) outliers by reassigning outlying values to the nearest value that lies within the valid (non-outlying) distribution. This approach maximizes accuracy and power simultaneously by maintaining all data points, while effectively reducing the influence of outliers. Typically, outliers are defined in a data-driven manner, using the observed distribution of values (e.g., making use of the median, or other specific percentile values, and/or interquartile ranges) to flexibly adjust definitions of outliers on the basis of a specific sample and task design. As a result, rescaling outliers might promote dot-probe reliability in a more robust manner across studies, although no previous report has tested this contention.

Finally, we considered the possibility that RT measures of attention bias, even when analyzed in an optimal fashion, may be limited in their reliability due to sources of error inherent to the RT measurement itself. Tasks such as the dot-probe were originally developed in the cognitive psychology domain to infer information about cognition based on overt behavior (RT). However, recently developed methods may provide a more proximal measure of the constructs of interest, which include visual and cognitive processing of threatening information and attentional allocation, rather than speed of reaction *per se*, which may be influenced by a variety of irrelevant factors (e.g., response selection latency, delays in registration of response due to imprecise button pressing or equipment failure). One such technology is eyetracking, which provides a direct measurement of participants' eye gaze patterns as an index of overt visual attention to threat. When used in the context of the dot-probe, preliminary evidence suggests eyetracking indices of attention towards threat may correlate with RT measures of bias, suggesting possible convergent validity across measurement modalities (Mogg, Garner, & Bradley, 2007). To our knowledge, no previous study has examined the reliability of eyetracking indices collected during the dot-probe. One recent study compared internal consistency reliability of eyetracking indices collected

during a free viewing task (i.e., a distinct task) with dot-probe RT indices (Waechter et al., In press). During later viewing periods (e.g., after 1000 ms), reliability of the eyetracking indices far exceeded the reliability of the dot-probe RT bias measures. Therefore, we sought to explore whether, given a sufficiently long face presentation interval of 2000 ms, the potential of eyetracking indices collected during the dot probe itself may surpass that of RT indices.

In the current study, we sought to develop a set of empirically-grounded recommendations for maximizing the test-retest reliability of the dot-probe. Given that the task is routinely used to study clinical samples of small to modest size [e.g. metaanalytic mean n=~20 per group; (Bar-Haim et al., 2007)], we were particularly interested in maximizing reliability in the context of such small samples. All samples were selected based on the availability of multi-session, test-retest dot-probe data collected in the absence of any explicit intervention. We utilized data from a clinical sample (29 adults with Social Anxiety Disorder) to explore the impact of design features (e.g., number of trials) and analysis features (e.g., identification and handling of reaction time outliers, formula for calculating attention bias scores) on test-retest reliability across 12 twice-weekly repeated measurements. We then assessed for convergence of reliability findings across one similar clinical sample assessed in the same laboratory setting (15 adults with Social Anxiety Disorder, assessed 8 times twice-weekly) and one more disparate sample (28 pediatric healthy controls, assessed 5 times bi-weekly). Inclusion of the pediatric sample allowed for preliminary exploration of whether recommendations would apply across development, clinical and non-clinical groups, and distinct laboratories using distinct task paradigms. Finally, using data available in the pediatric healthy control sample only, we compared the reliability of reaction time indices to the reliability of eyetracking indices of attention bias (collected concurrently during the dot-probe task) to guide recommendations regarding the inclusion of additional data collection modalities when available.

To assess reliability, we focused on test-retest reliability (stability) as a measure of the degree to which the dot-probe task can provide stable information regarding attention bias towards threat in relatively homogeneous samples receiving no explicit intervention. Measuring test-retest reliability allows us to assess whether the dot-probe, administered in its entirety under varying design conditions, can accurately measure a hypothesized construct (e.g., attention bias to threat; Attention Bias Variability) over a clinically meaningful period of time (i.e., several days-several weeks), an important prerequisite for many clinical, single-subject, and individual differences applications.

Methods

The present study utilized existing datasets from three studies of attentional bias that each used a dot-probe task with emotional and neutral face stimuli. Consistent with the typical usage of the dot-probe in affective research, the speed with which responses to the “probe” were made was examined as a function of the stimulus type (i.e., threat or neutral) that was presented and replaced by the probe. This analysis of RT data is designed to provide an index of attentional bias towards threat. Across all studies, three trial types were available and defined as follows (also see Figure 1): “Incongruent” trials = trials where the dot

replaced the neutral face in a threat/neutral face pair; “Congruent” trials = trials where the dot replaced the threat face in a threat/neutral face pair; and “Neutral” trials = trials containing neutral stimuli only, with no threat face [e.g., two images of the same neutral face (Studies 1 & 2) or two blank ovals (Study 3)]. Details of the samples, task parameters, and study procedure differed across the three studies, allowing for an examination of the robustness of patterns in stability across studies. Procedures for all three studies were approved by the Institutional Review Board of the relevant institution. Informed consent was obtained from adult participants while informed parental consent and child assent were obtained for pediatric participants (Study 3).

Study 1

Participants—Adults (age 19-62) with Social Anxiety Disorder were recruited for a randomized controlled trial comparing an attention retraining procedure to a sham training procedure.

Diagnostic assessment was based on a diagnostic interview using the Structured Clinical Interview for the DSM-IV (SCID) (First, Spitzer, Williams, & Gibbon, 1995). To be included in the study, participants met a principal DSM-IV diagnosis of Social Anxiety Disorder. Exclusionary criteria included: (a) evidence of suicidal intent, (b) evidence of current substance abuse or dependence, (c) evidence of current or past schizophrenia, bipolar disorder, or organic mental disorder, (d) any concurrent psychotherapy (e) change in pharmacological treatments during the 12 weeks prior to study entry, and (f) Cognitive Behavioral Therapy within the past 6 months. Data for the current study were utilized from individuals in the sham training procedure, which consisted of 12 repeated dot-probe task administrations in a laboratory setting, given twice-weekly over a 6-week period. Data from the attention retraining group were not used for the current analyses, as, by design, attention bias was expected to change over time. A final total of 29 participants completed all assessment points and were included in reliability analyses. Clinical and demographic characteristics of this sample are shown in Table 1.

Dot-probe task—Task stimuli consisted of 16 pictures from eight different individuals (four male, four female) with either a disgust or neutral expression. These faces were selected from a standardized facial stimuli set (Matsumoto & Ekman, 1989). In brief, participants saw two faces simultaneously presented on the computer screen for 500ms, one above the other. For consistency with prior research (e.g., Asmundson & Stein, 1994), participants were specifically instructed to fixate on the top face. The faces then disappeared, and a probe (the letter ‘E’ or ‘F’) appeared in place of either picture. Participants were instructed to respond as to whether the probe was an ‘E’ with a left mouse click or ‘F’ with a right mouse click. The probe appeared on the screen until participants responded. Trials consisted of either a disgust-neutral picture pairing (80% of trials) or a neutral-neutral picture pairing (20%). The probe appeared with equal frequency in place of the disgust and neutral faces in the disgust-neutral pairings. Each session consisted of 2 blocks of 160 trials per block (320 trials in total).

Study 2

Participants—Adults (age 18-54) with Social Anxiety Disorder were recruited for a randomized controlled trial comparing an attention retraining procedure to a sham training procedure. Inclusion and exclusion criteria were identical to those of Study 1. Data were utilized from individuals in the sham training procedure, which consisted of 8 repeated dot-probe task administrations in a laboratory setting, given twice-weekly over a 4-week period. A final total of 15 participants completed all assessment points and were included in reliability analyses. Clinical and demographic characteristics of this sample are shown in Table 1.

Dot-probe task—The dot-probe task was identical in Study 2 to that in Study 1, with the exception that each administration consisted of one block of 160 trials only.

Study 3

Participants—Youth (age 9-13) with no clinical diagnoses were recruited as healthy controls in the context of a larger study of pediatric anxiety. Data were concurrently collected from a sample of youth with anxiety disorder diagnoses; however, as all anxious youth participated in active treatment that had measurable effects on dot-probe indices over time, data from these participants were not included in the current analysis of test-retest stability. Diagnoses were made by trained interviewers using the Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version [K-SADS-PL; (Kaufman et al., 1997)]. Participants were excluded if they met diagnostic criteria for any lifetime DSM-IV disorder, exhibited IQ below 70 as assessed by the Wechsler Abbreviated Scale of Intelligence (WASI; (Wechsler, 1999), were taking psychoactive medications including anxiolytics and antidepressants, or had a parent with current or lifetime DSM-IV diagnosis of anxiety or mood disorders (assessed via SCID; (First et al., 1995). Interviews were conducted separately with parents and children, with clinicians integrating data from both sources. The dot-probe task was administered in a laboratory setting at 5 timepoints over a 14–week period: weeks 1, 3, 6, 10 & 14. A final total of 28 participants completed all assessment points and were included in reliability analyses. Clinical and demographic characteristics of this sample are shown in Table 1.

Dot-probe task—Participants completed the task alone in a dimly lit room. After an initial fixation cross presented in the middle of the screen (500ms), a fearful and a neutral face were presented simultaneously on the top and bottom of the screen for a 2000ms interval, followed by a probe (dot) replacing either the fearful face (“congruent” trials) or the neutral face (“incongruent” trials). Long “neutral” trials consisting of two blank ovals presented for 2000ms followed by a dot were included as a control condition. The dot remained on-screen for the remainder of the trial (8.8s), making a total duration of 11.3s for each trial. This unusually long trial duration was selected in order to optimize the task design for collection of slow event-related functional Magnetic Resonance Imaging (fMRI) data at a separate assessment point (data not included in this report). Additional trials with short (200ms) fearful-neutral face pair presentations were also administered, but were excluded from the current analysis given that a) no “neutral” (ovals-only) trials of this shorter duration were administered, meaning that bias scores comparing neutral to incongruent trials could not be

calculated for short trials, and b) the duration was too brief to allow for reliable collection of eyetracking data, precluding direct comparisons of RT and eyetracking reliability. Participants were instructed to respond as quickly as possible to the probe, indicating its location on the screen by pressing a key for up or down, and were instructed to keep eyes on the screen for the duration of the task. Faces were grayscale conversions of the well-validated NimStim battery (Tottenham et al., 2009), half male and half female, with the same actor presented in both images in each pair. Hair was cropped from the images to reduce distraction from irrelevant information. Participants completed 16 randomly interspersed trials of each type (congruent, incongruent, neutral) for a total of 48 trials (80 trials total, when including unanalyzed short-duration trials).

Eyetracking—Eyetracking data were collected during the task using a table-mounted RK-768 eye-tracker, consisting of a video camera and infrared light source pointed at a participant's eye and a device that tracked the location and size of the pupil and corneal reflection at 60 Hz (every 16.7 ms). The resolution for a typical participant was better than .05mm pupil diameter. Eye position was calculated based on the x- and y-coordinates of the recorded eye-gaze minus a corneal-reflection signal, which accounts for small head movements, and individually scaled and offset based on each individual's calibration parameters collected at the start of the session. Eye fixations were defined as eye positions stable within 1° of visual angle for at least 100ms and were used to calculate the following gaze pattern indices: percentage of trials with initial fixations falling within regions of interest defined by the fearful and neutral face boundaries (an index of initial attentional capture); percentage of time spent fixating on fearful and neutral faces (an index of overall attentional preference); mean duration of individual fixations on fearful vs. neutral faces (an index of attentional maintenance); disengage latency to fear vs. neutral faces (latency until the next eye movement, based on just those trials where the dot appeared in the opposite location from the subject's point of fixation at the end of the face pair presentation—an index of disengagement difficulty).

Reliability Analysis Strategy

To identify factors that increased reliability for dot-probe RT indices, we began by jointly examining the effects of a) outlier handling and b) bias index calculation strategies. For all RT analyses, data were first cleaned by removing trials with inaccurate responses and the first three trials of each block (to allow time for participants to acclimate to the task and for experimenters to exit the room). Additional outlier handling fell under two main strategies: (a) discretionary cut-offs to exclude outlier trials from analysis, and (b) rescaling (Winsorizing) outliers, using the observed distribution of RTs to define outliers in a data-driven manner. For discretionary cut-offs, two specific types of threshold were used, consistent with thresholds widely reported in the dot-probe literature. The first set of thresholds entailed deleting trials with specific RT values (RTs <200-300 and >2500-3000). The second set of thresholds entailed deleting trials that deviated from the distribution of each individual's RTs in that session (e.g., ±2 or ±3 SDs from an individual's mean RT). Each of these thresholds applied in isolation, as well as all possible combinations of the two sets of thresholds (absolute values and per-subject SDs), were applied to the data from the three studies. This allowed for assessment of the convergent or divergent effects of specific

combinations of thresholds across studies. For the rescaling outliers approach, a specific distribution-based strategy was applied to all three datasets, as described in previous dot-probe studies and in robust statistical procedures. A Winsorizing procedure was used to eliminate extreme values while minimizing missing values in the data. Values outside 1.5 interquartile ranges from the 25th or 75th percentiles (the "Tukey Hinges") of the full distribution of RT values (across all individuals and all sessions) were rescaled to the last valid value within that range, and then maintained as datapoints at these new, non-outlying values. For example, for a RT distribution with a 25^{th} percentile value of 600 ms, a 75^{th} percentile value of 800 ms, and an interquartile range of 200 ms, values >1100 would be rescaled to 1100 ms (the largest value in the distribution that is within the valid range) whereas values <300 ms would be rescaled to 300 ms (the smallest value in the distribution that is within the valid range).

After applying outlier handling, a mean RT was calculated for trials of a given type, for each participant at each session. Bias indices were then calculated using two general strategies described in the literature. For Incongruent vs. Congruent (ICvC) indices, the following formula was applied: Bias Score = mean RT to incongruent trials – mean RT to congruent trials. For Incongruent vs. Neutral (ICvN) indices, the following formula was applied: Bias Score = mean RT to incongruent trials – mean RT to neutral trials. Within each of these general strategies, we further examined the effect of including only trials with a dot appearing either on the bottom or the top of the screen in the calculation. This allowed us to examine the effect of eliminating variance related to a procedural variable (dot-top vs. dot-bottom), which was not of theoretical interest, prior to comparing trials based on variables of theoretical interest (e.g., congruence).

Finally, after identifying the optimal outlier-handling and bias index calculation strategies from these primary analyses, a further exploratory analysis was conducted to investigate whether stability could be further improved by averaging together indices acquired at two successive timepoints. Bias scores from every two timepoints in a given study were averaged (e.g., mean of timepoints 1 & 2, mean of timepoints 3 & 4, mean of timepoints 5 & 6, etc.) and then these averages were subjected to test-retest reliability analysis. This analysis allowed us to ask whether administering repeated dot-probe assessments, separated by roughly 2-5 days, would be helpful in attempting to promote valid inferences, e.g., for single-subject clinical applications.

Attention bias variability index—Attention bias variability (ABV) was quantified as described previously (Iacoviello et al., 2014). Dot-probe trials were separated into "bins" of 20 sequential trials each and an ICvC index was calculated for each bin (neutral-neutral trials within each bin were insufficient to calculate ICvN indices). The standard deviation of the ICvC indicesw was then calculated across all bins and divided by the individual's mean RT (across all threat-neutral trials in the experiment) to correct for variance in RTs. Thus, the index quantifies the within-subject, within-session variability (standard deviation) of bias, rather than quantifying attentional bias itself.

Reliability Index

The Intraclass Correlation Coefficient (ICC) was used to provide a summary test-retest reliability score that incorporates data from all timepoints simultaneously. The ICC is widely used as an index of reliability (Weir, 2005) and quantifies the proportion of total variance in a measurement that is attributable to between-subjects variance. In theory, ICCs range from 0 to 1, with 0 indicating no reliability and 1 indicating perfect reliability. In practice, it is possible for empirical estimates to be negative, as estimates all have upper bounds of 1, but no lower bounds. Negative ICCs are akin to a value of 0 in indicating no reliability. ICCs were calculated in SPSS using a 2-way random effects model (with random factors for assessment point and subject), allowing systematic variability due to assessment point to contribute to the error variance in the denominator so that interchangeability of timepoints is not assumed (an 'absolute' agreement definition). Under these conditions, ICC values can be roughly interpreted as a correlation coefficient (Bartko, 1966). Two distinct forms of ICC were calculated: the single measure ICC provides an assessment of the proportion of variance within a single measurement that is attributable to between-subjects variance, suggesting what the reliability level would be if a single assessment point was taken for each individual (the most likely scenario in clinical research); the average measures ICC provides an overall assessment of the proportion of variance that is attributable to between-subjects variance across all timepoints, suggesting what the reliability of the index would be if all the assessment points were administered and then averaged together. This value can be thought of as indicating the degree to which all assessment points measured the same trait, and is largely akin to the internal consistency index α (and, in all cases, has a nearly-identical value to α). The *p*-value for the F-test of the null hypothesis that ICC is not greater than 0 is identical whether single-measure or average-measures ICCs are being considered and is therefore reported once per analysis.

Results

Effects of Outlier Handling on Bias Score Reliability

When applying discretionary cut-offs to exclude outlier trials from analysis, examination of all possible combinations of thresholds revealed no consistent benefit for a single strategy across the three datasets. As illustrated in Table 2 (using two specific examples of threshold combinations), even minor changes to thresholding cut-offs had substantial effects on reliability that were inconsistent across the three studies. Whereas ICCs were generally better for Studies 1 and 2 using the first set of thresholds shown (delete RTs 300 and 2,500 and 3 SD from individual's session mean), which maximized reliability in Study 1 compared with all other threshold combinations, ICCs were generally better for Study 3 using the second set of thresholds (delete RTs 300 and 3,000 and 2 SD from individual's session mean), which maximized reliability in this study compared to all other threshold combinations.

By contrast, when an outlier rescaling (Winsorizing) approach was used, a more consistent benefit was observed across all three studies. ICCs using this approach tended to be similar to the maximum values obtained through trial-and-error searching for discretionary cut-points; i.e., ICCs were largely similar to those obtained by applying the best possible

discretionary threshold combination for a particular dataset (though ICCs decreased slightly in some cases).

Effects of Bias Score Calculation

The ICvC index based on dot-bottom trials alone was the index with the most significant and near-significant ($p<.1$) ICC results across the three studies. As shown through the highlighted text in Table 2, this index had the highest overall ICCs of all indices in 6 out of the 9 scenarios (study/outlier-handling combinations), and the highest ICCs in all three studies when outlier rescaling was applied. The ICvN index, when using all available trials (dot-bottom and dot-top), was significant for Study 1 only, which was the study with the largest number of trials administered. No other indices had significant ICCs across more than one scenario in Table 2.

Effects of Averaging Two Subsequent Timepoints

The analyses presented in Table 2 suggested that rescaling outliers and using the ICvC dot-bottom index was the optimal strategy for simultaneously improving reliability across the three studies. Using this index (rescaled ICvC dot-bottom indices), new ICCs were calculated using the average bias across pairs of two subsequent timepoints (in place of bias scores obtained from a single timepoint). ICC-single measure indices were marginally improved compared to the results in Table 2 for all three studies (Study 1: ICC-single measure=.19, ICC-average measures=.58, p=.001; Study 2: ICC-single measure=.12, ICC-average measures=.36, p=.13; Study 3: ICC-single measure=.24, ICC-average measures=.48, p=.02).

Reliability of Attention Bias Variability (ABV) Index

As shown in Table 3, ICCs for the ABV index greatly exceeded those for any index of attention bias *per se*. Stability was greatest in Study 1, which had the largest number of trials per session as well as the largest number of sessions. Unlike attention bias measures, maximal reliability was obtained for ABV when using a single set of discretionary cut-offs (excluding trials <300 and >3000ms and ±2 SDs from individual's session mean), rather than using a data-driven Winsorizing approach.

Reliability of Eyetracking Measures

As shown in Table 4, ICCs for eyetracking indices (obtained in Study 3 only) tended to be higher than those obtained from RT measures of attention bias. Three of the four indices had ICCs significantly greater than zero, and two of the indices had ICCs that surpassed that of any RT attention bias measure in any of the three studies.

Discussion

The current study examined effects of several analysis strategies on the test-retest stability of dot-probe task attentional bias indices. From these data we present a set of empirical recommendations for researchers using the dot-probe task. Several consistent, robust effects were observed across three diverse studies of attentional bias to threat-related faces, in spite of multiple disparities in task design and sample characteristics (clinical and non-clinical,

adult and pediatric). Thus these recommendations may be broadly applicable to a wide range of research studies. Given that reliability can strongly promote validity, these recommendations may improve researchers' capacity to detect effects of interest.

Attention Bias Score Calculation

For measures of attention bias *per se*, calculating bias based on just one dot-location tended to be more reliable than averaging both dot-locations, especially when dot-bottom trials were used and when an incongruent vs. congruent (ICvC) bias index was calculated. The dot-bottom ICvC bias score generally outperformed all other bias indices examined, including all variants of the ICvN index (Table 2). This specific benefit for dot-bottom ICvC indices was consistent across all three studies, and also relatively robust across multiple methods of outlier handling. This finding is particularly striking given that stability improved in spite of cutting the number of trials in half prior to bias score calculation. Homogeneity of dot-location may have reduced error variance, while between-subject differences in bias either remained unchanged or increased, resulting in an increased proportion of the total variance attributable to between-subject differences and, consequently, an improved ICC and improved reliability.

The benefit for dot-bottom trials over dot-top trials may be related to the instructions given in Studies 1 and 2 to fixate on the top face in the pair. Although this explicit instruction was not given in Study 3, typically eye gaze may be preferentially drawn to the top half of the screen, irrespective of stimulus valence (Waechter et al., In press). Fixation on the top face means that dot-bottom trials are more likely to require a saccade prior to response. Given that dot-probe bias scores have sometimes been argued to primarily index a bias in attentional disengagement [rather than an initial orienting bias; (Koster et al., 2004)], this need for a saccade away from the top face might elicit a stronger and more robust signal of attentional bias (i.e., indexing the efficiency with which attention is disengaged and redirected from faces of a certain type) when comparing congruent and incongruent trials.

As a practical matter, researchers with existing dot-probe datasets utilizing a vertical task orientation (as in the current studies) may wish to consider analyzing dot-bottom trials alone. Researchers designing new dot-probe studies may consider including slightly more dot-bottom than dot-top trials or utilizing horizontal stimulus presentation to eliminate sources of noise in the data pertaining specifically to top/bottom attentional allocation. Our data cannot directly address whether such designs would have a beneficial effect on stability, but provide tentative suggestions and hypotheses for future research. By contrast, our data do clearly suggest that inclusion of both congruent and incongruent trials is important, given that optimal reliability was obtained from an ICvC index rather than an ICvN index. In attention bias modification protocols, where incongruent trials are predominantly presented in order to induce a bias away from threat through training (e.g., Amir, Beard, Burns, & Bomyea, 2009; Amir, Beard, Taylor, et al., 2009; MacLeod, Rutherford, Campbell, Ebsworthy, & Holker, 2002), it would appear beneficial to include at least a small number of congruent ("catch") trials in instances where the researcher's goals include ongoing assessment of changes in bias over the course of training.

Outlier Handling

A standard approach utilized by most dot-probe researchers to handle irrelevant RT outliers (e.g., subject became distracted, looked away from the screen, responded prematurely) has been to set discretionary definitions of outlying RTs *a priori* and exclude all trials outside these thresholds. The present analyses suggest that, when outlier trials are defined and excluded in this manner, even minor adjustments in the precise thresholds can substantially impact reliability of attention bias scores, and these effects are inconsistent from one study to the next. Thus, no simple rule-of-thumb was evident that could be applied broadly across studies. By contrast, rescaling (Winsorizing) outliers, and using the observed distribution of RTs to define outliers in a data-driven manner, performed more consistently across the three studies. Rescaling outliers produced ICCs that approximated the best possible threshold combination for each study, but eliminated the need for trial-and-error searching for optimal thresholds on a per-study basis.

Number of Dot-Probe Trials

The three studies included in this analysis ranged widely in the number of available trials for analysis, from 320 (Study 1) to 48 trials (Study 3). It is therefore notable that test-retest stability metrics for attention bias measures were fairly consistent across the three studies when optimal procedures were applied (ICvC dot-bottom index, rescaling outliers). The stability benefit observed using dot-bottom only trials further suggests that the number of trials used for calculation was far less influential in determining stability in comparison to other factors (e.g., reducing noise induced by extraneous design features). These findings suggest that considerable streamlining of the dot-probe assessment procedure may be possible without sacrificing test-retest stability. However, given that few studies have utilized such abbreviated versions of the dot-probe, and the current study focused exclusively on reliability, the validity of such an approach requires further examination.

Reliability of Attention Bias Variability (ABV) Index

While the extant dot-probe literature has focused almost exclusively on the derivation of a single summary score of attention bias, this approach may mask within-subjects, intrasession variability in bias that is of both theoretical and practical import. An explicit index of such variation, ABV, exhibited vastly superior stability across all three datasets compared to attention bias indices. Unlike attention bias indices, the number of trials included did influence ABV reliability, with optimal stability achieved with 320 trials. In addition, unlike with bias indices, rescaling outliers in a data-driven matter was not the optimal outlier handling strategy. Instead, a single discretionary set of exclusions was optimal. This distinction may be related to the nature of the index, which capitalizes on variability and therefore might be less accurately assessed when the data are forced into a more normal distribution (as occurs during Winsorizing). These promising reliability results, coupled with initial findings linking ABV to increased PTSD symptoms (Iacoviello et al., 2014), suggest further validation of the ABV index is warranted. Given that the index can be readily calculated from virtually any dot-probe dataset, a large store of relevant data already exists that could be used for this purpose.

Reliability of Dot-Probe RT Measures for Single-Subject Clinical Applications

Our analyses suggest specific recommendations to improve the reliability of dot-probe RT indices. Nevertheless, even in the best of circumstances, single-measure ICCs for RT attention bias indices remained low (<.20). Although no specific cut-points are generally prescribed for ICCs, measurement error reflected in any ICC less than 1.0 will attenuate correlations between that measure and any other measure, with the size of this detrimental impact increasing as the ICC approaches zero (Weir, 2005). It has been noted that the attenuating effect of measurement error on correlations becomes minimal as ICCs increase above 0.80 (Nunnally & Bernstein, 1994), a threshold not approached by any single-measure ICC of attention bias in the current study. Overall, our findings suggest that a single assessment point of dot-probe bias score is likely to have low power for detection of relationships with other constructs of interest. Averaging across two subsequent assessment points did improve single-measure ICCs across all three studies, but only marginally (ICCs remained <.25). By contrast, the average-measures ICCs in Table 2, which fell in the .5-.6 range under the best of circumstances, suggest that averaging across multiple (5-12) distinct sessions improves reliability. Unfortunately, it is likely impractical for many dot-probe researchers to collect data in this manner.

While suboptimal stability of bias indices could derive from dot-probe RTs providing an unreliable assessment (e.g., due to measurement noise), there may be both theoretical and analytic constraints that place a necessary upper limit on the stability of bias. For example, RT difference score calculation (obtained through subtraction of 2 correlated measurements) may necessarily constrain reliability (see Sipos, Bar-Haim, Abend, Adler, & Bliese, In press). Furthermore, attention bias may not in fact be a stable, trait-like construct, but may fluctuate due to changing contextual factors, even within the course of a single assessment session. The vastly improved stability observed for the ABV index, which explicitly quantifies such intrasession fluctuations, suggests this variability itself exhibits trait-like qualities, possibly related to impairment of attentional control (Iacoviello et al., 2014), and represents an important source of information that cannot be ignored if stability is to be optimized.

In spite of these constraints, and the empirical limits we and others have observed on the reliability of dot-probe bias indices, a substantial literature supports the task's capacity to differentiate between anxious and non-anxious samples (Bar-Haim et al., 2007). One possible explanation is that a "file-drawer" problem masks the number of negative studies that have been conducted and never published. In addition, distinctions across disparate samples (e.g., anxious and non-anxious samples) in attention bias towards threat may be sufficiently large so as to be detected even when substantial measurement error exists. Since the ICC reflects the proportion of between-subjects variance to total variance, it increases as between-subjects differences increase, reflecting the intuitive conclusion that small differences between individuals are more difficult to detect than large ones. However, our data suggest that dot-probe RT bias scores collected at a single assessment point within a fairly homogeneous sample (e.g., socially anxious adults, healthy youth) are likely to be underpowered for individual differences analyses, as their relationship with other variables of interest (e.g., neural measures of brain function; clinical symptoms) will be constrained.

In particular, single-subject applications (i.e., prescribing treatments based on dot-probe performance) may require further improvements to reliability. Alternate (RT or non-RT) measures of attention bias (e.g., explicit quantification of ABV) and/or more sophisticated modeling of dissociable RT influences (Ratcliff, 2008) may be required in order to make reliable inferences about individual patients. Notably, the ABV index exhibited a single-measure ICC of .65 when a relatively large number of trials were administered (Study 1), making it potentially a much more attractive candidate for further study of single-subject clinical applications and individual differences research.

Reliability of Eyetracking Indices

In a pediatric sample of healthy volunteers, eyetracking indices appeared generally more reliable than RT indices of bias, even with a very small number of trials. Although single-measure ICCs remained suboptimal, for two of the indices examined (percentage of trials with initial fixation on fear face; time spent fixating on fear vs. neutral faces), single-measure ICCs surpassed .3, falling in the “moderate” correlation range. When possible, incorporation of eyetracking into dot-probe assessment may therefore be beneficial. Many commercially available “smartphones” now come with the ability to perform basic eyetracking functions through front-facing cameras, suggesting eyetracking may become an increasingly feasible way to assess attention bias in the future, even in non-laboratory settings. Eyetracking measures collected during the dot-probe task have not been widely validated in the same manner as dot-probe RTs, although initial data suggest convergence across measurement modalities (Mogg et al., 2007). However, it is important to note that eyetracking measures of attention bias may only be sensitive to certain components of attention (Petersen & Posner, 2012; Posner & Petersen, 1990) that may or may not be critical in psychopathology. For example, sensitivity to errors (Hajcak & Foti, 2008), which can be readily assessed using RT measures (e.g. the flanker task), would not be assessed by eyetracking.

Notably, a recent study reported ICCs for fMRI activation during the dot-probe task in the .7-.8 range (Britton et al., 2013). Future attention bias research would likely benefit from incorporation of additional assessment modalities (e.g., eyetracking, fMRI) that might provide improved measurement stability and allow for further cross-validation of the information acquired through newer technologies.

Limitations

The current report explicitly assessed test-retest reliability, an important determinant of validity, but not validity *per se*. Future studies should expand on these findings through a focus on simultaneous maximization of reliability and validity. All three datasets utilized threat-related (fearful or disgusted) and neutral face stimuli and thus we were unable to assess the generalizability of findings to dot-probe procedures utilizing words, other types of pictures, and other types of stimulus contrasts (e.g., positive/appetitive vs. neutral, disorder-relevant vs. non-disorder-relevant), nor can we comment on the comparative reliability of alternative RT measures of attention bias (e.g., exogenous cuing tasks). We are also unable to assess generalizability to samples that differ substantially from those included here, although convergence of findings across distinct clinical and non-clinical, and adult and

pediatric, samples appears to support generalizability of findings across disparate groups. Finally, we intentionally conducted our analyses in small samples representative of those common in extant dot-probe studies. A complimentary approach for future research would be to assess the degree to which these recommendations are necessary or beneficial when larger sample sizes are available.

Conclusion

The present analysis leads to several concrete, empirically based recommendations for dot-probe data analysis that may maximize the stability of attention bias scores. These are: (1) calculate bias scores as Incongruent – Congruent mean RT, using only dot-bottom trials; (2) rescale outliers rather than excluding them from analysis, defining outliers in a data-driven manner rather than based on discretionary thresholds; (3) administer as many repeated dot-probe assessments as possible (two assessments are marginally better than one, whereas >=5 assessments is substantially more reliable). When applying these strategies to RT data across three distinct studies, reliability of bias scores tended to improve, but remained below levels typically recommended for psychometric adequacy. Thus, researchers utilizing the dot-probe should be mindful of the decreased power and upper limits on validity that may stem from inadequate reliability, particularly when individual differences and/or single-subject applications are the focus of research. Furthermore, it is recommended that researchers (4) quantify within-sessions variability in bias (ABV), as this appears to represent a stable, trait-like component of dot-probe RTs and (5) when possible, incorporate complementary indices of attentional bias into dot-probe assessment, such as eyetracking, which may have improved reliability over RTs. As research on attention bias continues to advance from relatively straight-forward between-groups comparisons (e.g., do anxious individuals exhibit increased attention towards threat compared to non-anxious individuals?) to addressing more complex, nuanced, and clinically relevant issues (e.g., do those anxious individuals with the greatest attention bias benefit more from treatment X than treatment Y?; do individual differences in attention bias correlate with specific neurobiological dimensions of function?), attending to the psychometric properties of attention bias indices--and utilizing methods to maximize them--will become increasingly critical.

Acknowledgments

Disclosure: Supported by National Institutes of Health grants MH080215, MH082998, MH018269, and 5R01MH087623-03. Dr. Price is supported by a Career Development Award from NIMH (1K23MH100259). Dr. Amir is the co-founder of a company that markets anxiety relief products.

References

Amir N, Beard C, Burns M, Bomyea J. Attention modification program in individuals with generalized anxiety disorder. Journal of Abnormal Psychology. 2009; 118:28–33. doi:10.1037/a0012589. [PubMed: 19222311]

Amir N, Beard C, Taylor CT, Klumpp H, Elias J, Burns M, et al. Attention training in individuals with generalized social phobia: A randomized controlled trial. Journal of Consulting and Clinical Psychology. 2009; 77(5):961–973. doi:10.1037/a0016685. [PubMed: 19803575]

Amir N, Najmi S, Morrison AS. Attenuation of attention bias in obsessive-compulsive disorder. Behaviour Research and Therapy. 2009; 47(2):153–157. doi:10.1016/j.brat.2008.10.020. [PubMed: 19046576]

Amir N, Taylor CT, Donohue MC. Predictors of response to an attention modification program in generalized social phobia. Journal of Consulting and Clinical Psychology. 2011; 79:533–541. doi: 10.1037/a0023808. [PubMed: 21707134]

Asmundson GJG, Stein MB. Selective processing of social threat in patients with generalized social phobia: Evaluation using a dot-probe paradigm. Journal of Anxiety Disorders. 1994; 8(2):107–117.

Ataya AF, Adams S, Mullings E, Cooper RM, Attwood AS, Munafo MR. Internal reliability of measures of substance-related cognitive bias. Drug and Alcohol Dependence. 2012; 121(1-2):148–151. doi:10.1016/j.drugalcdep.2011.08.023. [PubMed: 21955365]

Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van IJzendoorn MH. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psychological Bulletin. 2007; 133(1):1–24. doi:10.1037/0033-2909.133.1.1. [PubMed: 17201568]

Bartko J. The intraclass correlation coefficent as a measure of reliability. Psychological Reports. 1966; 19:3–11. [PubMed: 5942109]

Beard C, Sawyer AT, Hofmann SG. Efficacy of attention bias modification using threat and appetitive stimuli. Behavior Therapy. 2012; 43(4):724–740. doi:10.1016/j.beth.2012.01.002. [PubMed: 23046776]

Britton JC, Bar-Haim Y, Clementi M. a. Sankin LS, Chen G, Shechner T, et al. Training-associated changes and stability of attention bias in youth: Implications for Attention Bias Modification Treatment for pediatric anxiety. Developmental Cognitive Neuroscience. 2013; 4:52–64. doi: 10.1016/j.dcn.2012.11.001. [PubMed: 23200784]

Broadbent D, Broadbent M. Anxiety and attentional bias: State and trait. Cognition & Emotion. 1988; 2(3):165–183.

Cardi V, Di Matteo R, Corfield F, Treasure J. Social reward and rejection sensitivity in eating disorders: an investigation of attentional bias and early experiences. World Journal of Biological Psychiatry. 2013; 14(8):622–633. doi:10.3109/15622975.2012.665479. [PubMed: 22424288]

Dear BF, Sharpe L, Nicholas MK, Refshauge K. The psychometric properties of the dot-probe paradigm when used in pain-related attentional bias research. Journal of Pain. 2011; 12(12):1247–1254. doi:10.1016/j.jpain.2011.07.003. [PubMed: 21982721]

Erceg-Hurn DM, Mirosevich VM. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. American Psychologist. 2008; 63(7):591–601. doi: 10.1037/0003-066X.63.7.591. [PubMed: 18855490]

First, MB.; Spitzer, RL.; Williams, JBW.; Gibbon, M. Structured clinical interview for DSM-IV Axis I disorders - patient edition. New York Psychiatric Institute; New York: 1995.

Forestell CA, Dickter CL, Young CM. Take me away: the relationship between escape drinking and attentional bias for alcohol-related cues. Alcohol. 2012; 46(6):543–549. doi:10.1016/j.alcohol. 2012.05.001. [PubMed: 22705274]

Gotlib IH, McLachlan AL, Katz AN. Biases in visual attention in depressed and nondepressed individuals. Cognition & Emotion. 1988; 2:185–200.

Hajcak G, Foti D. Errors are aversive: defensive motivation and the error-related negativity. Psychological Science. 2008; 19:103–108. doi: 10.1111/j.1467-9280.2008.02053.x. [PubMed: 18271855]

Iacoviello B, Wu G, Abend R, Murrough J, Feder A, Fruchter E, et al. Attention bias variability and symptoms of posttraumatic stress disorder. Journal of Traumatic Stress. 2014; 27:232–239. doi: 10.1002/jts.21899. [PubMed: 24604631]

Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADSPL): Initial reliability and validity data. Journal of the American Academy of Child & Adolescent Psychiatry. 1997; 36:980–987. [PubMed: 9204677]

Klumpp H, Amir N. Examination of vigilance and disengagement of threat in social anxiety with a probe detection task. Anxiety, Stress, and Coping. 2009; 22:283–296. doi: 10.1080/10615800802449602.

Koster EH, Crombez G, Verschuere B, De Houwer J. Selective attention to threat in the dot probe paradigm: differentiating vigilance and difficulty to disengage. Behaviour Research and Therapy. 2004; 42(10):1183–1192. doi:10.1016/j.brat.2003.08.001. [PubMed: 15350857]

Legerstee JS, Tulen JHM, Kallen VL, Dieleman GC, Treffers P. D. a. Verhulst FC, et al. Threat-related selective attention predicts treatment success in childhood anxiety disorders. Journal of the American Academy of Child and Adolescent Psychiatry. 2009; 48:196–205. doi:10.1097/CHI.0b013e31819176e4. [PubMed: 19127173]

MacLeod C, Mathews A, Tata P. Attentional bias in emotional disorders. Journal of Abnorm Psychology. 1986; 95(1):15–20.

MacLeod C, Rutherford E, Campbell L, Ebsworthy G, Holker L. Selective attention and emotional vulnerability: assessing the causal basis of their association through the experimental manipulation of attentional bias. Journal of Abnormal Psychology. 2002; 111(1):107–123. [PubMed: 11866165]

Matsumoto D, Ekman P. American-Japanese cultural differences in intensity ratings of facial expressions of emotion. Motivation & Emotion. 1989; 13(2):143–157.

Mogg K, Garner M, Bradley BP. Anxiety and orienting of gaze to angry and fearful faces. Biological Psychology. 2007; 76:163–169. doi:10.1016/j.biopsycho.2007.07.005. [PubMed: 17764810]

Mohlman J, Price RB, Vietri J. Attentional bias in older adults: Effects of generalized anxiety disorder and cognitive behavior therapy. Journal of Anxiety Disorders. 2013; 27:585–591. 10.1016/j.janxdis.2013.06.005. [PubMed: 23916715]

Nunnally, J.; Bernstein, I. Psychometric Theory. 3rd ed.. McGraw-Hill; New York: 1994.

Petersen S, Posner M. The attention system of the human brain: 20 years after. Annual Review of Neuroscience. 2012; 35:73–89. doi: 10.1146/annurev-neuro-062111-150525.

Posner M, Petersen S. The attention system of the human brain. Annual Review of Neuroscience. 1990; 13:25–42.

Price M, Tone EB, Anderson PL. Vigilant and avoidant attention biases as predictors of response to cognitive behavioral therapy for social phobia. Depression and Anxiety. 2011; 28:349–353. doi: 10.1002/da.20791. [PubMed: 21308888]

Price RB, Siegle G, Silk JS, Ladouceur CD, McFarland A, Dahl RE, et al. Sustained neural alterations in anxious youth performing an attentional bias task: A pupilometry study. Depression and Anxiety. 2013; 30:22–30. doi: 10.1002/da.21966. [PubMed: 22700457]

Ratcliff R. Methods for dealing with reaction time outliers. Psychological Bulletin. 1993; 114:510–532. [PubMed: 8272468]

Ratcliff R. Modeling aging effects on two-choice tasks: response signal and response time data. Psychology & Aging. 2008; 23(4):900–916. doi:10.1037/a0013930. [PubMed: 19140659]

Salemink E, van den Hout MA, Kindt M. Selective attention and threat: quick orienting versus slow disengagement and two versions of the dot probe task. Behaviour Research and Therapy. 2007; 45(3):607–615. doi:10.1016/j.brat.2006.04.004. [PubMed: 16769035]

Schmukle SC. Unreliability of the dot probe task. European Journal of Personality. 2005; 19:595–605. doi:10.1002/per.554.

Sipos ML, Bar-Haim Y, Abend R, Adler AB, Bliese PD. Postdeployment threat-related attention bias interacts with combat exposure to account for PTSD and anxiety symptoms in soldiers. Depression and Anxiety. In press. doi:10.1002/da.22157.

Staugaard SR. Reliability of two versions of the dot-probe task using photographic faces. Psychology Science Quarterly. 2009; 51:339–350.

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Research. 2009; 168(3):242–249. doi: 10.1016/j.psychres.2008.05.006. [PubMed: 19564050]

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

Waters AM, Lipp OV, Spence SH. Attentional bias toward fear-related stimuli: an investigation with nonselected children and adults and children with anxiety disorders. Journal of Experimental Child Psychology. 2004; 89(4):320–337. doi:10.1016/j.jecp.2004.06.003. [PubMed: 15560877]

Waters AM, Mogg K, Bradley BP. Direction of threat attention bias predicts treatment outcome in anxious children receiving cognitive-behavioural therapy. Behaviour Research and Therapy. 2012; 50:428–434. doi:10.1016/j.brat.2012.03.006. [PubMed: 22542533]

Waters AM, Wharton TA, Zimmer-Gembeck MJ, Craske MG. Threat-based cognitive biases in anxious children: comparison with non-anxious children before and after cognitive behavioural treatment. Behaviour Research and Therapy. 2008; 46(3):358–374. doi:10.1016/j.brat.2008.01.002. [PubMed: 18304519]

Wechsler, D. Wechsler Abbreviated Scale of Intelligence (WASI). Harcourt Assessment; San Antonio: 1999.

Weir J. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. The Journal of Strength & Conditioning Research. 2005; 19:231–240.

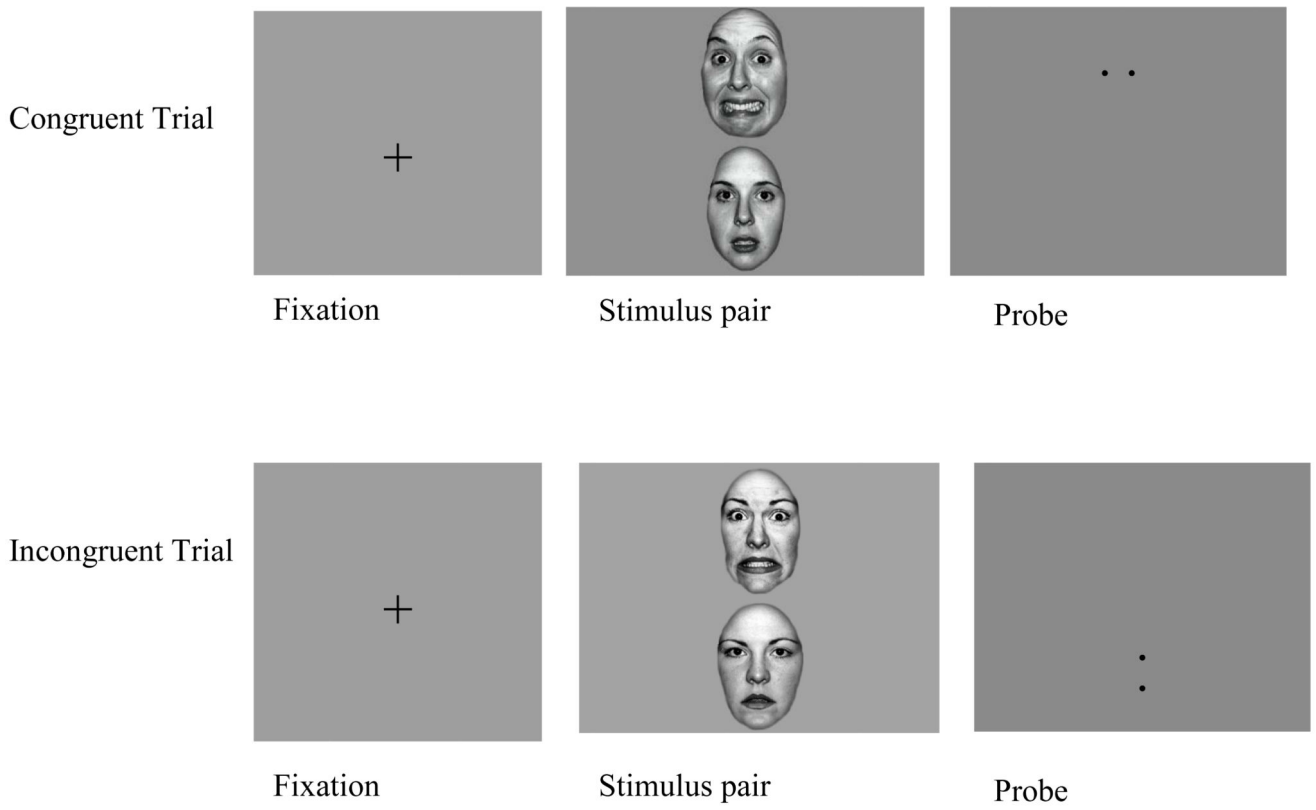

Figure 1.
Schematic of dot-probe task

Table 1

Demographic and clinical features of the three study samples

	Study 1 (*N*=29)	**Study 2 (*N*=15)**	**Study 3 (*N*=28)**
Age	34.14 (11.34)	24.47 (9.16)	10.8 (1.7)
Female, *n* (%)	17 (58.62%)	7 (46.67%)	15 (53.6%)
Caucasian, *n* (%)	11 (37.93%)	14 (93.33%)	23 (82.1%)
Liebowitz Social Anxiety Scale	80.86 (15.73)	64.60 (22.26)	--
Beck Depression Inventory	23.79 (8.67)	18.73 (8.95)	--
Pediatric Anxiety Rating Scale	--	--	1.5 (2.5)
Mood and Feelings Questionnaire	--	--	1.4 (1.9)

Note: Data presented as mean (*SD*) unless otherwise noted. Scores from the parent-report version of the Mood and Feelings Questionnaire, a pediatric depression inventory, are presented.

Table 2

Effects of bias score type, dot location, and outlier handling method on reliability indices

	Study 1 (N=29; 320 trials) Adult Social Anxiety Disorder			*Study 2 (N =15; 160 trials) Adult Social Anxiety Disorder*			*Study 3 (N =28; 48 trials) Pediatric Healthy Volunteers*		
Method/measure	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***P***	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***p***	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***p***
Rescale (Windsorize) RTs									
ICvC Index: All trials	−.01	−.20	.72	−.02	−.13	.59	*.10*	*.35*	*.062*
ICvC Index: Dot-on-bottom trials only	**.09**	**.55**	**.001**	**.11**	**.49**	**.025**	**.19**	**.53**	**.003**
ICvC Index: Dot-on-top trials only	**.04**	**.35**	**.045**	−.07	−1.04	.95	.01	.05	.41
ICvN Index: All trials	**.07**	**.45**	**.008**	−.03	−.31	.70	*.09*	*.33*	*.077*
ICvN Index: Dot-on-bottom trials only	.032	.28	.10	.05	.31	.14	.03	.12	.32
ICvN Index: Dot-on-top trials only	−.01	−.11	.61	.03	.21	.25	**.12**	**.39**	**.036**
Delete RTs <300 & >2500 & ±3 *SD*s from individual's session mean									
ICvC Index: All trials	−.04	−.74	.96	.06	.34	.10	*.09*	*.34*	*.07*
ICvC Index: Dot-on-bottom trials only	**.08**	**.49**	**.003**	**.19**	**.65**	**.001**	*.09*	*.33*	*.09*
ICvC Index: Dot-on-top trials only	.002	.02	.44	−.003	−.03	.49	−.03	−.14	.64
ICvN Index: All trials	**.08**	**.50**	**.002**	−.06	−.8	.89	−.03	−.19	.68
ICvN Index: Dot-on-bottom trials only	.02	.22	.17	.01	.07	.39	−.002	−.02	.49
ICvN Index: Dot-on-top trials only	.02	.23	.15	.06	.34	.13	−.11	−1.0	.98
Delete RTs <300 & >3000 & ±2 *SD*s from individual's session mean									
ICvC Index: All trials	−.05	−1.5	.99	.04	.27	.17	**.19**	**.55**	**.002**
ICvC Index: Dot-on-bottom trials only	*.04*	*.34*	*.052*	**.09**	**.45**	**.044**	**.13**	**.44**	**.025**
ICvC Index: Dot-on-top trials only	.006	.07	.37	−.004	−.04	.50	.08	.29	.11
ICvN Index: All trials	.03	.28	.10	−.09	−2.2	.99	.05	.21	.20
ICvN Index: Dot-on-bottom trials only	.01	.11	.30	.02	.12	.34	**.13**	**.42**	**.03**
ICvN Index: Dot-on-top trials only	−.01	−.04	.66	.001	.007	.45	−.04	−.23	.73

Note: RTs=reaction times; ICvC=Incongruent vs. Congruent; ICvN=Incongruent vs. Neutral; ICC=Intraclass Correlation Coefficient. ICC values less than zero indicate negative average covariance and poor reliability. *p* values are provided for the *F*-test of the hypothesis that ICC >0; values with $p<.05$ are shown in bold and $p<.10$ are shown in italics. For each study and each outlier handling method, the index with the highest overall reliability is highlighted in pink.

Table 3

Reliability of Attention Bias Variability (ABV) index as a function of outlier handling method

	Study 1 (N=29; 320 trials) Adult Social Anxiety Disorder			*Study 2 (N=15; 160 trials) Adult Social Anxiety Disorder*			*Study 3 (N=28; 48 trials) Pediatric Healthy Volunteers*		
Method/measure	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***p***	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***p***	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***p***
Rescale (Windsorize) RTs									
Attention Bias Variability Index	**.39**	**.88**	**<.001**	**.14**	**.57**	**.006**	**.15**	**.47**	**.01**
Delete RTs <300 & >2500 & ±3 *SD*s from individual's session mean									
Attention Bias Variability Index	**.63**	**.95**	**<.001**	**.13**	**.54**	**.02**	**.21**	**.57**	**.001**
Delete RTs <300 & >3000 & ±2 *SD*s from individual's session mean									
Attention Bias Variability Index	**.65**	**.96**	**<.001**	**.22**	**.69**	**.001**	**.30**	**.68**	**<.001**

Note: RTs=reaction times. Significant ICC's obtained in every analysis. Method yielding the highest overall reliability for each study is highlighted in pink.

Table 4

Reliability of eyetracking attention bias indices in pediatric healthy volunteer sample

	Study 3 (N=28; 32 trials) Pediatric Healthy Volunteers		
Bias measure	**ICC (single measure)**	**ICC/alpha (average of all measures)**	***P***
% of trials with initial fixation on fear face	**.33**	**.71**	<**.001**
Time spent fixating on fear vs. neutral faces (across full trial)	**.32**	**.70**	<**.001**
Mean duration of individual fixations on fear vs. neutral faces	.08	.31	.10
Disengage latency to fear vs. neutral faces (based on trials where dot appeared in opposite location from subject's current fixation)	**.14**	**.44**	**.02**

Note: ICC=Intraclass Correlation Coefficient. *p* values are provided for the *F*-test of the hypothesis that ICC >0; values with $p<.05$ are shown in bold.