Psychological Research (2019) 83:308–320
https://doi.org/10.1007/s00426-017-0947-6

ORIGINAL ARTICLE

Fleeting reliability in the dot-probe task

Angus Chapman[1] · **Christel Devue**[1] · **Gina M. Grimshaw**[1]

Received: 14 May 2017 / Accepted: 13 November 2017 / Published online: 20 November 2017

Abstract
In a dot-probe task, two cues—one emotional and one neutral—are followed by a probe in one of their locations. Faster responses to probes co-located with the emotional stimulus are taken as evidence of attentional bias. Several studies indicate that such attentional bias measures have poor reliability, even though ERP studies show that people reliably attend to the emotional stimulus. This inconsistency might arise because the emotional stimulus captures attention briefly (as indicated by ERP), but cues appear for long enough that attention can be redistributed before the probe onset, causing RT measures of bias to vary across trials. We tested this hypothesis by manipulating SOA (stimulus onset asynchrony between onset of the cues and onset of the probe) in a dot-probe task using angry and neutral faces. Across three experiments, the internal reliability of behavioural biases was significantly greater than zero when probes followed faces by 100 ms, but not when the SOA was 300, 500, or 900 ms. Thus, the initial capture of attention shows some level of consistency, but this diminishes quickly. Even at the shortest SOA internal reliability estimates were poor, and not sufficient to justify the use of the task as an index of individual differences in attentional bias.

Keywords Attention · Dot probe · Emotion · Attentional bias · Threat bias · Reliability

Introduction

A number of theories of emotion posit that attention is biased towards threatening stimuli (Öhman, Lundqvist, & Esteves, 2001; Okon-Singer, Lichtenstein-Vidne, & Cohen, 2013; Pourtois, Schettino, & Vuilleumier, 2013; Yiend, 2010). A common paradigm used for measuring such attentional bias is the dot-probe task (MacLeod, Mathews, & Tata, 1986; Mogg & Bradley, 1999), in which a pair of stimuli are presented on opposite sides of a display: one neutral and one emotional (commonly threatening). Subsequently, a probe is presented in the location previously occupied by one of the stimuli, and participants must respond to a feature of the probe. Attentional biases are inferred in the dot-probe task when participants respond faster to probes that replace emotional than neutral stimuli.

While anxious participants regularly show an attentional bias to threat, this bias is not always shown by non-anxious participants (Bar-Haim et al., 2007). However, indices of attentional allocation from EEG studies reveal that even non-anxious participants preferentially attend to threatening images (Eimer & Kiss, 2007; Grimshaw, Foster, & Corballis, 2014; Holmes, Bradley, Kragh Nielsen, & Mogg, 2009; Kappenman, Farrens, Luck, & Proudfit, 2014), although they may not sustain engagement with those images (Kappenman, MacNamara, & Proudfit, 2015). These findings suggest that although threatening images capture attention, non-anxious individuals may be able to effectively and quickly disengage from such stimuli.

One proposed explanation for the discrepancy between behavioural and neural measures of attentional bias is that behavioural measures simply lack sufficient internal reliability to uncover these biases (Kappenman, MacNamara, & Proudfit, 2015). Over the past decade, the psychometric properties of response time (RT) measures of attentional biases have been explicitly investigated, and several studies, using a wide range of stimuli, have shown that biases in the dot-probe task have poor internal reliability (Amir, Zvielli, & Bernstein, 2016; Cooper et al., 2011; Dear, Sharpe, Nicholas, & Refshauge, 2011; Kappenman et al., 2014, 2015;

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00426-017-0947-6) contains supplementary material, which is available to authorized users.

✉ Gina M. Grimshaw
gina.grimshaw@vuw.ac.nz

[1] School of Psychology, Victoria University of Wellington, PO Box 600, Wellington 6012, New Zealand

Reutter, Hewig, Wieser, & Osinsky, 2017; Schäfer et al., 2016; Schmukle 2005; Staugaard 2009; Van Bockstaele et al., 2011; Waechter, Nelson, Wright, Hyatt, & Oakman, 2014; Waechter & Stolz, 2015; Zvielli, Bernstein, & Koster, 2015; see Rodebaugh et al., 2016, for a discussion of reliability in the dot-probe task). In other words, participants do not show a consistent bias towards (or away from) emotional stimuli on a trial-by-trial basis. Although some fluctuation is to be expected, these studies have shown that the reliability of attentional biases often approaches zero (see Table 1 for a summary of findings from published studies of internal reliability). Such poor reliability calls into question the validity of attentional bias measured by the dot-probe task: if these biases are so inconsistent, then how can this task provide any insight into attention to threat?

While some researchers suggest that poor reliability indicates that the dot-probe task is inherently flawed and unsuited for assessing attentional bias (e.g., Kappenman et al., 2014), it might in fact reflect the variability in the time course of different attentional processes. Many models conceptualise attention as a series of component processes including selection, orienting, engagement, disengagement, and shifting (Koster, Crombez, Verschuere, & De Houwer, 2004; Koster, Verschuere, Crombez, & Van Damme, 2005; Posner, Cohen, & Rafal, 1982); some of these processes may be deployed with consistent timing from trial to trial, while others may be more variable. In line with this idea, recent electrophysiological studies in non-anxious populations have revealed dissociations in the reliability of two ERP components that are used to measure attention to emotional stimuli (Kappenman et al., 2014, 2015): the earlier N2pc, indexing initial attentional selection, and the later LPP, indexing sustained engagement with a stimulus. The N2pc reveals an attentional bias towards threatening images in the dot-probe task that has moderate reliability (much greater than behavioural measures, r's = .5–.9; Kappenman et al., 2014, 2015; Reutter et al., 2017). On the other hand, the LPP shows no evidence of sustained engagement with the threat stimulus, and has poor internal reliability (Kappenman et al., 2015). These findings suggest that individuals are consistent in their initial attentional selection, while the duration of engagement with the emotional stimulus before disengagement takes place is much less consistent. Because behavioural measures of attentional bias are likely an index of multiple attentional processes (Koster et al., 2004, 2005), the stability of these underlying processes is crucial to the reliability of the behavioural measures.

The poor internal reliability of attentional biases in the dot-probe task could, therefore, be explained by dissociations in the consistency with which selection, engagement, and disengagement mechanisms are deployed. Importantly, most studies investigating reliability of the dot-probe task present cue stimuli 500 ms prior to the appearance of the probe (stimulus onset asynchrony, SOA; see Table 1). Given the speed of covert attentional shifts (< 100 ms, Buschman & Miller, 2009; Müller & Rabbitt, 1989), attention could be redirected or redistributed a number of times across the 500 ms SOA interval. There could also be variability in the timecourse of disengagement processes, as suggested by the poor reliability of the LPP (Kappenman et al., 2015), making the locus of attention at the time of probe onset highly variable across trials. In light of these multiple degrees of freedom, it is perhaps unsurprising that attentional biases have poor reliability; by measuring bias with an SOA of 500 ms, reaction time measures are likely tapping into different stages of attentional processing on each trial. This time course hypothesis leads to the prediction that reliability should be much higher if the task is able to directly tap the initial selection or orienting stage of attention, before more variable processes occur. This could be accomplished by reducing the SOA (e.g., to 100 ms), so that attention is still biased to the preferred stimulus when the probe appears.

While the predominant presentation time in the dot-probe task is 500 ms, manipulations of SOA are commonly used to isolate different attentional processes (Bradley, Mogg, Falla, & Hamilton, 1998; Koster et al., 2005; Mogg, Bradley, De Bono, & Painter, 1997). However, few studies have included multiple SOAs when assessing reliability of attentional biases, and those that include multiple SOAs do not directly compare reliabilities across them (see e.g., Schmukle, 2005; Staugaard, 2009). In three experiments, we investigated the internal reliability of attentional biases across a range of SOAs (100, 300, 500, and 900 ms), to determine whether reliability is related to the time course of attentional processing. At very short SOAs, bias measures should primarily reflect the initial (perhaps automatic) selection and orienting of attention to the threat-related stimulus. Given the electrophysiological evidence reviewed above, we might expect these bias measures to be reliable. At mid-range SOAs, attentional deployment might be expected to vary more widely from trial to trial, resulting in lower reliability. We included the much longer (900 ms) SOA in Experiment 1 to test a further hypothesis. Attentional biases at very long SOAs are thought to reflect an eventual "resting" place of attention, and may show either bias towards threat (vigilance), or alternatively away from threat (avoidance; Booth, 2014; Koster, Crombez, Verschuere, Van Damme, & Wiersema, 2006; Onnis, Dadds, & Bryant, 2011). If this propensity for vigilance versus avoidance is a stable individual difference, reliability at the 900 ms SOA should again improve.

In all three experiments, we recruited participants from a general student population and not from a clinically anxious population. However, in Experiment 2, we also assessed state anxiety, to determine whether it modulates the reliability of attentional biases. In Experiments 1 and 2, we

Table 1 Estimates of internal reliability in existing published studies

Authors	Condition	Stimuli	*n*	SOA (ms)	# trials	Task	Method	Reliability	Bias
Amir et al. (2016)	–	Threat scenes	59	500	80	Location	F/S	.19[a]	Vigilance
Cooper et al. (2011)				500	16	Discrimination	SH		
Expt. 1	Control	Fearful faces	20 (ws)					.08	ns
	CO_2 challenge							−.08	Vigilance
Expt. 2	Control	Fearful faces	30 (ws)					−.22	Vigilance
	CO_2 challenge							.26	ns
	Control	Angry faces						−.01	Vigilance
	CO_2 challenge							.09	ns
Dear et al. (2011)				500	40	Discrimination	SH		
	Control	Pain-related pictures	100					.05	Not reported
		Pain-related words						−.05	Not reported
	Chronic pain	Pain-related pictures	139					.25	Not reported
		Pain-related words						.10	Not reported
Kappenman et al. (2014)	–	Threat scenes	96	500	360	Discrimination	SH	.030[a]	ns
Kappenman et al. (2015)	–	Threat scenes	30	500	400	Discrimination	SH	.35[a]	ns
Reutter et al. (2017)	–	Angry faces	92	500	288	Discrimination	SH	.056	ns
Schäfer et al. (2016)	–	Angry faces	144	500	48	Discrimination	F/S	.12[a]	Not reported
Schmukle (2005)	–					Detection	SH		
Exp 1		Social/physical threat words	80	100	64			.03	Not reported
				450–675	64			−.15	Not reported
Exp 2	–	Threat scenes	40	500	48			−.08	Not reported
Staugaard (2009)		Angry faces	39		96	Discrimination	F/S		
	Standard dot-probe			100				.174	Vigilance
				500				−.054	Vigilance
	Modified dot-probe			100				−.290	Vigilance
				500				.074	ns
Van Bockstaele et al. (2011)	–	Spiders	55	500	64	Discrimination	SH	.15	ns
Waechter et al. (2014)		Angry faces		500	64	Discrimination	Comp		
	Low anxious		40					−.10	ns
	High anxious		41					−.10	ns
Waechter and Stolz (2015)		Angry faces		500	64	Location	Comp		
	Control		76					.59[a]	ns
	Anxious		82					.15[a]	ns[b]
Zvielli et al. (2015) Exp 2	–	Smoking-related	45	500	80	Location	F/S	.06[a]	

Table includes studies that assessed reliability in a dot-probe task to measure threat-related biases in a non-selected (or non-clinical) population. Condition refers to participant subgroups or experimental manipulations. Task refers to the probe task. Method of calculation: *SH* even–odd or random split-half method, *F/S* first-half/second-half split method, *Comp*. computational methods. For reliability, [a]indicates values with Spearman–Brown correction. For bias measures, vigilance indicates bias towards threat, ns indicates no significant bias. [b]In Waechter & Stolz (2015), a high anxious subgroup showed an attentional bias away from angry faces (i.e., avoidance)

followed the common practice in the literature of presenting the faces throughout the SOA period. However, this practice confounds SOA with stimulus duration. Therefore, in Experiment 3, we eliminated this confound by presenting all faces for a short fixed stimulus duration regardless of SOA, to determine whether extended exposure to the faces might account for poor reliability at mid-range SOAs.

Experiment 1

In the first experiment, to cover a broad range of SOAs, we presented threatening and neutral faces to the left and right of fixation for 100, 300, 500, or 900 ms prior to the onset of the probe stimulus. The experiment was conducted in two parts, such that participants completed trials with just three SOAs: 100, 300, and 500 ms in Experiment 1A and 100, 500, and 900 ms in Experiment 1B.

Method

Participants

One-hundred and thirty-three first-year psychology students participated in Experiment 1; 68 in Experiment 1A, and 65 in Experiment 1B. Six participants in Experiment 1A were excluded from analysis for poor accuracy (> 3 *SD*s below mean accuracy on any block), one for slow response times (> 3 *SD*s above mean on any block) and one for not maintaining position in the chin rest during the task. Three participants in Experiment 1B were excluded for poor accuracy. The two experiments, therefore, comprised a total of 122 participants, aged between 18 and 25 years (15 men and 45 women, $M_{age} = 18.68$ years, $SD = 1.24$, in Experiment 1A; 15 men and 47 women, $M_{age} = 18.42$ years, $SD = 1.05$, in Experiment 1B), with normal or corrected-to-normal vision. These sample sizes were sufficient to detect reliabilities of $r > .30$ in each experiment, and greater than $r > .22$ when participants from both experiments were combined for the 100 and 500 ms SOAs, given 80% power.[1] Participants completed the task in groups of up to four, seated in separate cubicles. All participants gave informed consent, and the study was approved by the School of Psychology Human Ethics Committee, Victoria University of Wellington.

[1] Power was calculated assuming a one-tailed test, as negative reliabilities are meaningless. This method underestimates the true power of our Experiments, as it is based on a single correlation coefficient and not the distribution of reliabilities that our computational method produces.

Materials and apparatus

The experiment was presented on a Dell Precision T1700 computer running OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) using a 24″ inch AOC monitor with a resolution of 1920 by 1080 pixels and a 120 Hz vertical refresh rate. Participants were seated at a viewing distance of 57.3 cm from the monitor, maintained by use of a chin rest.

Angry and neutral face pairs of six male actors were taken from the NimStim facial expressions set (#20, 21, 23, 24, 25, 34; Tottenham et al., 2009). Angry faces showed open mouths, while neutral faces had closed mouths. All faces were greyscaled and edited to include only the face and hair, then superimposed on a grey rectangle, such that the point between the eyes was at the rectangle's centre. The pairs of faces from each model were equated for average pixel luminance and root mean square contrast with the SHINE Image Processing Toolbox for MATLAB (Willenbockel et al., 2010). Resulting images had a mean luminance of 40 cd/m^2, and were presented against a dark grey background with mean luminance of 16 cd/m^2.

Procedure

Each trial began with a 25% grey fixation cross (0.6° × 0.6°) presented in the centre of the screen for a random interval between 400 and 800 ms. The fixation cross remained on screen until probe offset, and participants were instructed to fixate on the cross while it remained on screen. Two faces, one angry and one neutral (from the same model), subtending 6.9° × 9.0° of visual angle each, were presented with the inner edge of the face rectangle appearing 2.4° to the left or right of fixation. Immediately following the offset of the faces, a probe was presented for 100 ms consisting of two 25% grey dots in either a vertical or horizontal orientation (: or. .) in the location previously occupied by the centre of the angry face (angry cue) or neutral face (neutral cue), 5.9° to the left or right of fixation. Participants were required to report the orientation of the dot pair with their dominant hand using the '1' or '2' keys on the number pad, with key mapping counterbalanced across participants. Trials ended after a response was made, or 1700 ms following the offset of the dots. Incorrect or omitted responses resulted in a short tone presented through headphones as feedback.

Participants completed three blocks of 96 trials, one for each SOA (100, 300, or 500 ms in Experiment 1A; 100, 500, or 900 ms in Experiment 1B). Block order was counterbalanced across participants. Blocks consisted of two subblocks of 48 trials, separated by a break, each consisting of each possible combination of angry face location (left, right), probe location (left, right), probe orientation (vertical, horizontal), and image model (one of six). Thus, the probe was presented an equal number of times in the location of

the angry and neutral face, such that faces were not predictive of probe location. Prior to the experimental blocks, participants completed a block of 48 practice trials with an SOA of 500 ms, using angry and neutral faces of a model not included in the experimental blocks (NimStim #37; Tottenham et al., 2009).

Statistical analysis

Attentional bias Only response times for correct responses were used in analyses. Individual response times for each participant on each trial were inverse transformed (1/RT) to minimize the effects of slow RTs on means (Ratcliff, 1993). Mean response times for each cue type were calculated from the transformed data, and these were then reverted to ms, as bias scores are best interpreted in ms units. Bias scores were calculated by subtracting mean response times for angry cues from mean response times for neutral cues ($RT_{neutral} - RT_{angry}$).

Reliability Attentional bias is a difference score, and so a standard Cronbach's alpha can only be calculated by artificially arranging individual trials into different subsets based on specific faces or locations (e.g., Schmukle, 2005; Staugaard, 2009; Dear et al., 2011). Alternatively, a split-half reliability can be calculated (see Table 1), in which two bias scores are calculated for each participant (for example, based on odd or even trials, or the first and second halves of an experiment; see Williams & Kaufmann, 2012, for a discussion of different methods), and then, these are correlated. However, this method provides only two samples of the bias for each participant, and an estimated reliability that lies somewhere along a distribution of possible reliabilities. A better estimate can be calculated using computational methods (MacLeod et al., 2010; Williams & Kaufmann, 2012; Waechter et al., 2014), in which each participant's response time data are randomly split in half many times. Our analysis was carried out on 1000 random splits of the data using R (Version 3.1.1; R Core Team, 2015). Angry and neutral cue trials were split in half independently, and bias scores were calculated as described above. The pair of bias measures obtained in each iteration were then correlated, providing a distribution of reliabilities from which a mean and standard deviation were calculated. An important advantage of this method is that it reduces sampling variability that can occur when just one random split of the data is done, and, therefore, minimizes the impact of other sources of RT variability. Where noted, reliabilities are corrected for test length using the Spearman–Brown prophecy formula (Spearman, 1910). This value (indicated as r_{SB}) corrects for the fact that any test is divided in half to calculate a split-half reliability, which reduces any correlation relative to a test–retest reliability (in which the entire test is repeated; DeVellis, 1991; Miller & Ulrich, 2013). The formula uses the relationship between test length and reliability to provide a predicted reliability for the test, were it to be presented at its full length. We use uncorrected *r* values for our statistical analyses, but also present corrected values to facilitate comparison with the other measures in the literature. For psychometric purposes, corrected reliability is conventionally considered "respectable" for *r*'s > .70 and "unacceptable" for *r*'s < .60 (DeVellis, 1991).

Similarly, we estimated the reliability of RTs for each cue valence: data from angry and neutral cue trials were split in half independently, mean RT in each half was calculated for each participant, and the halves were correlated. Reliability was calculated as the mean correlation across 1000 random splits of the data.

Significance of each reliability estimate was calculated using permutation tests. The data set was subjected to the same procedure as described above, except that trials were randomly labelled as having "angry" or "neutral" cues. For each SOA, labels were randomly permuted prior to reliability analysis. One-thousand mean reliability estimates (each consisting of 1000 splits) from random permutations of the data formed the null distribution, and the *p* value was calculated as the proportion of the null distribution with reliability greater than the mean of the original, non-permuted data. This is equivalent to a one-tailed significance test.

Because we were also interested in whether reliability was greater at some SOAs than others, we statistically compared reliability estimates for each pair of SOAs. For each SOA, 10,000 samples (with replacement) were drawn from the computational reliability distribution. For each pairwise comparison, the *p* value was calculated as the proportion of samples for one SOA in which reliability was greater than for the other SOA.

Results and Discussion

Experiments 1A and 1B were analysed separately. Accuracy and response time data were subjected to 3 (SOA: 100, 300, and 500 ms in Experiment 1A; 100, 500, and 900 ms in Experiment 1B) × 2 (cue valence: angry, neutral) repeated-measures Analyses of Variance (ANOVAs).

Accuracy

Overall, participants were highly accurate in identifying the orientation of dots (M = 93.0%, SD = 3.5% in Experiment 1A; M = 92.3%, SD = 5.1% in Experiment 1B). Analysis of each experiment revealed no main effects or interaction, all *p*'s > .17.

Table 2 Mean response time (ms) and bias scores as a function of SOA in each experiment

Condition	Experiment 1A	Experiment 1B	Experiment 2	Experiment 3
100 ms SOA				
Neutral	465 (65)	477 (80)	478 (68)	453 (80)
Angry	474 (63)	480 (82)	477 (63)	456 (82)
Bias Score	− 9* (25)	− 3 (29)	1 (17)	− 3 (24)
300 ms SOA				
Neutral	473 (67)			
Angry	471 (71)			
Bias Score	2 (21)			
500 ms SOA				
Neutral	474 (73)	482 (88)	484 (74)	426 (82)
Angry	472 (68)	481 (86)	487 (71)	425 (19)
Bias Score	2 (23)	1 (23)	− 3 (15)	1 (19)
900 ms SOA				
Neutral		493 (90)		
Angry		492 (90)		
Bias Score		1 (24)		

Standard deviations appear in parentheses. Bias scores are calculated as RT(neutral cue) − RT(angry cue)
*$p < .05$

Response times

Mean response times and bias scores for each SOA in all experiments are presented in Table 2. To assess whether participants showed group-level behavioural biases at any SOA, ANOVAs were conducted on mean correct response times in each experiment. In Experiment 1A, this revealed no main effects of SOA, $F(2, 118) = 0.28$, $p = .756$, $\eta_p^2 = .005$, nor cue valence, $F(1, 59) = 0.92$, $p = .342$, $\eta_p^2 = .015$, but an SOA × cue valence interaction, $F(2, 118) = 4.64$, $p = .011$, $\eta_p^2 = .073$. To further explore this interaction, bias scores were calculated and one-sample *t*-tests revealed that biases at 100 ms significantly differed from zero, $t(59) = 2.81$, $p = .007$, $d = 0.14$. Unexpectedly, participants showed a significant bias away from the angry face at 100 ms. Bias scores at 300 and 500 ms did not significantly differ from zero, both t's < 1.

In Experiment 1B, analyses revealed a main effect of SOA, $F(2, 122) = 4.11$, $p = .019$, $\eta_p^2 = .063$. Post-hoc Bonferroni corrected comparisons revealed a general speeding in response for trials with shorter SOAs, such that response times were overall faster during 100 ms SOA blocks (M = 478 ms, SD = 80) than 900 ms blocks (M = 493 ms, SD = 89; p = .047); however, neither were significantly different from response times on 500 ms blocks (M = 481 ms, SD = 86; p's > .10). Neither the main effect of cue valence, $F(1, 61) = 0.02$, $p = .900$, $\eta_p^2 < .001$, nor the SOA × cue valence interaction were significant, $F(2, 122) = 0.48$, $p = .623$, $\eta_p^2 = .008$, showing that there were no significant attentional biases in this experiment. Thus, the bias away from angry faces in the 100 ms condition of Experiment 1A was not replicated in Experiment 1B.

Reliability

Reliability of RTs in all conditions in both experiments was high (r's > .86, Supplementary Table S1). Means and *SD*s of bias score reliability distributions for all experiments are presented in Table 3. In Experiment 1A, reliability was marginally significant at 100 ms, $r = .140$, $p = .084$; when adjusted for test length, this measure of reliability was low ($r = .140$; $r_{SB} = .246$). Reliability at 300 and 500 ms did not significantly differ from zero, both p's > .58. Comparing reliabilities to each other shows that reliability at 100 ms was marginally greater than at 300 ms, $p = .078$, but not 500 ms, $p = .112$, which did not differ, $p = .424$. In Experiment 1B, reliability was again marginally significant at 100 ms, $p = .051$. When adjusted for test length, this measure of reliability was still low by conventional standards ($r = .186$; $r_{SB} = .313$). Reliability at 500 and 900 ms did not significantly differ from zero, both p's > .55. At 100 ms, reliability was marginally greater than at both 500 ms, $p = .088$, and 900 ms, $p = .073$, which did not differ, $p = .462$.

Because the 100 and 500 ms SOA blocks were effectively equivalent across experiments 1A and 1B, we combined the two data sets and analysed reliability in the larger sample. Reliability was significantly greater than zero for biases at 100 ms, $p = .020$, although still low after correcting for test length ($r = .162$; $r_{SB} = .279$). Reliability was poor ($r = -.092$) and not significant at 500 ms, $p = .672$. Reliability at 100 ms

Table 3 Mean (SD) reliability of bias scores as a function of SOA in each experiment

Condition	Experiment 1A	Experiment 1B	1A/1B Combined	Experiment 2 (first half)	Experiment 2 (second half)	Experiment 2 (both halves)	Experiment 3
100 ms SOA	.140† (.108)	.186† (.113)	.162* (.082)	− .119 (.091)	.176* (.095)	.091 (.100)	.181* (.097)
300 ms SOA	− .074 (.106)						
500 ms SOA	− .046 (.110)	− .025 (.106)	− .044 (.065)	− .021 (.091)	− .038 (.092)	− .125 (.086)	− .141 (.100)
900 ms SOA		− .042 (.109)					

*$p < .05$
†$p < .10$

was significantly greater than at 500 ms, $p = .029$. This latter finding is consistent with our prediction that reliability would be improved when bias scores primarily reflect attentional selection and/or orienting.

To summarise, reliability of attentional biases was greater than zero at 100 ms but not at any other SOAs tested. However, this estimate of reliability is still considerably lower than is acceptable by conventional psychometric standards. By 300 ms, reliability was not significantly greater than zero, suggesting that the allocation of attention becomes inconsistent shortly following the initial attentional processes. As in previous studies (e.g., Schmukle, 2005; Staugaard, 2009), biases at 500 ms were unreliable. We further show that bias is unreliable at 900 ms, suggesting that even at long SOAs, attentional bias scores do not reliably reflect individual differences in attentional allocation either towards or away from threat.

Experiment 2

In Experiment 2, we attempted to improve the reliability of attentional bias measurements. Reliability is directly linked to the number of trials used for analysis (Miller & Ulrich, 2013; Williams & Kaufmann, 2012). Therefore, we assessed reliability of attentional biases after doubling the number of trials used at the two critical SOAs (100 and 500 ms). Furthermore, we also collected measures of anxiety and of self-reported attentional control. Previous studies have not consistently found a link between individual differences in anxiety and biases in the dot-probe task (Bradley et al., 1997; Kappenman et al., 2015), though such a link is predicted by theories of emotion and attention (Bar-Haim et al., 2007; Bishop, 2008; Eysenck, Derakshan, Santos, & Calvo, 2007; Eysenck & Derakshan, 2011; Okon-Singer et al., 2013). One proposed explanation for this discrepancy is the poor reliability of measures of attentional biases, because correlations based on unreliable measures are themselves unreliable. In this experiment, we expected that increasing the trial numbers would further improve the reliability of the measure of attentional biases at 100 ms, but perhaps not 500 ms, given that the 100 ms SOA most cleanly captures early attentional selection and orienting.

Method

Participants

Eighty-nine participants completed the experiment in exchange for course credit. Three participants were excluded for low accuracy (> 3 *SD*s below mean accuracy at either SOA). The final sample consisted of 86 participants (19 men and 67 women), aged between 18 and 30 years (M = 19.00 years, SD = 2.15, one participant did not provide their age). This sample size is sufficient to detect reliability of $r > .26$ with 80% power.

Procedure

The stimuli and procedure were the same as in Experiment 1 with the following exceptions. Participants completed eight blocks of 48 trials that alternated between 100 and 500 ms probe SOAs (192 total trials at each SOA). The starting block was counterbalanced across participants. Prior to the experimental blocks, participants completed a block of 48 practice trials with face cues presented for 300 ms. Following the experiment, participants completed the State-Trait Anxiety Inventory (Spielberger & Sydeman, 1994), which consists of 42 items assessing symptoms of anxiety, and the Attentional Control Scale (Derryberry & Reed, 2002), which consists of 20 items assessing an individual's self-reported ability to control and focus their attention on real-world tasks. Internal reliability of each scale was acceptable in this sample (α's > .85). Participants also completed the Fear of Spiders Questionnaire (Szymanski and O'Donohue, 1995) as part of recruitment for other studies in our lab; findings are not reported here.

Results and Discussion

Accuracy and response time data were analysed in separate 2 (SOA: 100, 500 ms) × 2 (cue valence: angry, neutral) repeated-measures ANOVAs.

Accuracy

Overall, accuracy was high (M = 93.5%, SD = 4.2), and analyses revealed no main effects or interaction, F's < 1.

Response times

Mean response times and bias scores at each SOA are presented in Table 2. To assess whether participants showed group-level behavioural biases at either SOA, we analysed mean correct response times. There was a main effect of SOA, $F(1, 85) = 5.04$, $p = .027$, $\eta_p^2 = .056$, which reflected faster responses in the 100 ms block (M = 478 ms, SD = 64.7) than in the 500 ms block (M = 485 ms, SD = 72.3). However, there was no significant main effect of cue valence, $F(1, 85) = 1.53$, $p = .220$, $\eta_p^2 = .018$, nor a significant interaction between SOA and cue valence, $F(1, 85) = 2.41$, $p = .124$, $\eta_p^2 = .028$, showing that there were no significant attentional biases in this experiment.

Reliability

Reliability of mean RT was high in each of the individual conditions (r's > .91, Supplementary Table S1). Means and SDs of reliability estimates for bias scores are presented in Table 3. In contrast to Experiment 1, reliability was not greater than zero at either 100 ms, $p = .200$, or 500 ms, $p = .900$, although reliability was marginally greater at 100 ms than at 500 ms, $p = .050$. Because of the increased number of trials, participants may have adjusted the way that they control their attention to the stimuli during the experiment, leading to changes in reliability over time. To test for this possibility, we split trials from each SOA into halves corresponding to the first and second parts of each SOA block. For the first 92 trials, reliability was not greater than zero at 100 ms, $p = .932$, or 500 ms, $p = .558$, and these estimates did not differ from each other, $p = .213$. For the last 92 trials, however, reliability was significantly above zero at 100 ms, $p = .020$ (providing similar reliability to that in Experiment 1 when corrected for test length; $r = .176$; $r_{SB} = .299$), but not at 500 ms, $p = .636$. Statistically, reliability was marginally greater at 100 than 500 ms, $p = .051$.

Contrary to our expectations, using twice as many trials per SOA as in Experiment 1 did not improve the internal reliability of attentional bias (see Miller & Ulrich, 2013; Williams & Kaufmann, 2012). However, when trials were separated into the first and second half of the experiment (such that trial numbers were equivalent to those in Experiment 1), we found that reliability was greater than zero for the 100 ms SOA, but only in the second half of the experiment. This finding should be considered exploratory, given that we had no hypotheses to suggest that reliability should improve over trials.

Relationship to anxiety

There was a moderate, negative relationship between trait anxiety as assessed with the STAI-T and self-reported attentional control as assessed with the Attentional Control Scale, $r(85) = -.477$, $p < .001$, such that greater reported anxiety was related to poorer self-reported attentional control. Attentional bias at 100 ms was unrelated to both anxiety, $r(85) = .100$, $p = .357$, and attentional control, $r(85) = .026$, $p = .814$. For attentional bias at 500 ms, there was a marginal, weak, positive relationship with anxiety, $r(85) = .184$, $p = .091$, but no relationship with attentional control, $r(85) = -.131$, $p = .230$. Attentional biases at 100 and 500 ms were unrelated, $r(85) = .170$, $p = .117$.

Because anxiety and attentional control are thought to be related to attentional biases (Eysenck et al., 2007; Eysenck & Derakshan, 2011), we expected that improved reliability at shorter SOAs might allow us to observe such relationships. Many studies have failed to find a relationship between these measures and attentional bias (e.g., Bradley et al., 1997; Kappenman et al., 2015), but this may have been due to the poor reliability of bias scores. Because internal reliability was not high in this experiment, it is perhaps unsurprising that no significant relationships were observed.

Experiment 3

In Experiment 3, we made two changes to overcome some limitations in the previous experiments. First, in both Experiments 1 and 2, only six individual faces were presented (with an angry and a neutral expression for each), and therefore, each face was presented many times (48 times in Experiment 1 and 64 times in Experiment 2). Overexposure to the faces may have reduced their novelty, and therefore, the amount of attention they received, reducing our ability to observe an attentional bias. We therefore increased our stimulus set to 16 individuals (with an angry and neutral expression for each), which were only presented 16 times each across the course of Experiment 3.

Second, although most dot-probe studies manipulate the presentation duration of cues alongside SOA (i.e., the cues are present across the whole interval), this procedure confounds stimulus duration and SOA. Because many additional processes will be engaged while faces are present (e.g., memory processes, eye movements, or face perception),

unreliable biases at longer SOA may reflect variable deployment of these processes, and not attentional allocation per se. Therefore, in Experiment 3, we held the stimulus duration constant at 70 ms while changing the cue-to-target interval to manipulate SOA.

Method

Participants

Seventy-nine participants completed the experiment in exchange for course credit. Three participants were excluded for low accuracy (> 3 *SD*s below mean accuracy in either block). The final sample consisted of 76 participants (15 men and 61 women), aged between 18 and 24 years (M = 18.68 years, SD = 1.18). This sample size is sufficient to detect reliability of $r > .27$ with 80% power.

Materials and Procedure

Angry and neutral face pairs of 16 male actors were taken from the NimStim facial expressions set (#20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 34, 35, 36, and 37; Tottenham et al., 2009). One further actor (#33) was used for practice trials only. The images were edited as in Experiment 1 and 2, and pairs were equated for average pixel luminance and root mean square contrast.

The procedure was similar to Experiments 1 with the following exceptions. Participants completed two blocks of 128 trials, one with 100 ms SOA and one with 500 ms SOA. The order of blocks was counterbalanced across participants. Stimulus duration was fixed at 70 ms, and the cue-to-target interval was either 30 or 430 ms. Prior to the experimental blocks, participants completed a block of 48 practice trials with face cues presented for 300 ms.

Results and Discussion

Accuracy and response time data were analysed in separate 2 (SOA: 100, 500 ms) × 2 (cue valence: angry, neutral) repeated-measures ANOVAs.

Accuracy

Overall, accuracy was high (M = 93.1%, SD = 4.3), and analyses revealed no main effects or nor interaction, F's < 1.25, p's > .25.

Response times

Mean response times and bias scores at each SOA are presented in Table 2. To assess whether participants showed group-level behavioural biases at either SOA, we analysed mean correct response times. There was a main effect of SOA, $F(1, 75) = 27.28$, $p < .001$, $\eta_p^2 = .267$, which reflected faster responses in the 500 ms block (M = 426 ms, SD = 80.0) than in the 100 ms block (M = 454 ms, SD = 80.2). This is opposite to the SOA effects observed in Experiments 1 and 2, but is to be expected in this experiment, because attention is not engaged with the face across the face-probe interval, meaning that participants are better able to prepare for the probe in the long SOA condition. Importantly, there was no significant main effect of cue valence, $F(1, 75) = 0.49$, $p = .485$, $\eta_p^2 = .007$, nor a significant interaction between SOA and cue valence, $F(1, 75) = 1.85$, $p = .178$, $\eta_p^2 = .024$, showing that there were no significant attentional biases in this experiment.

Reliability

Reliability was high for RTs in the individual conditions (r's > .91, Supplementary Table S1). Means and *SD*s of reliability estimates for bias scores are presented in Table 3. Consistent with our predictions, reliability was significantly above zero at 100 ms, $p = .036$, but was still lower than conventional standards ($r = .181$; $r_{SB} = .306$). Reliability was not significant at 500 ms, $p = .899$. Furthermore, reliability was statistically greater at 100 than 500 ms, $p = .010$.

Although we created a larger stimulus set, and manipulated SOA independently of stimulus presentation time, this experiment shows similar results to Experiment 1. Reliability at 100 ms was significantly greater than zero, and its magnitude here ($r = .181$) was similar to that observed in Experiment 1 ($r = .162$). As in both the previous experiments, attentional bias was not reliable at the 500 ms SOA, suggesting that confounds associated with extended stimulus duration cannot account for the poor reliability at this longer SOA.

General discussion

This series of experiments was motivated by a simple question: why is the reliability of the dot-probe task so poor, even though neural measures suggest that even non-anxious people preferentially attend to threat-relevant stimuli? We hypothesised that an important factor might be the relatively long SOA (500 ms) at which reliability is most commonly assessed. It may well be the case that the initial selection and orienting of attention to a threat stimulus is reliable, but the long SOA means that there is ample time for attention to be disengaged and reallocated or redistributed before the probe appears. If the timecourse of these various processes varies from trial to trial, internal reliability of the bias should be expected to be low. This hypothesis leads us to predict that the bias should be reliable (or at least more reliable) at

a short (100 ms) SOA when the task more closely captures just the initial attention selection and orienting to the threat stimulus. We further hypothesised that attentional bias measures might become reliable again at very long SOA (900 ms) when attention can be more strategically directed towards a preferred stimulus (whether that bias results in vigilance or avoidance of the threat stimulus).

We, therefore, conducted three experiments to investigate the internal reliability of attentional biases in the dot-probe task across a range of SOAs. Consistent with our predictions, biases in Experiment 1 were significantly reliable at 100 ms, but not at other SOAs (300, 500, and 900 ms). In Experiment 2, in which we doubled the number of trials in an attempt to improve the reliability of biases, we failed to replicate the findings of Experiment 1 (when considering either the total set of trials or the first half). However, biases at 100 ms in the second half of this experiment showed a level of reliability consistent with that in Experiment 1. In Experiment 3, we replicated the findings of Experiment 1, revealing significant reliability at 100 ms but not at 500 ms. Because stimulus duration was fixed at 70 ms in Experiment 3, this replication shows that poor reliability at 500 ms SOA is not the result of non-attentional processes (such as memory, eye movements, or other aspects of face processing) associated with extended stimulus duration. Collectively, these experiments provide evidence that the initial selection/orienting of attention shows some level of trial-to-trial consistency (though much lower than is generally considered to be psychometrically acceptable; DeVellis, 1991) and that SOA is a notable moderator of the internal reliability of attentional bias measures.

In addition to the poor reliability of biases at 500 ms across all experiments, in Experiment 1, reliability was also found to be poor at SOAs of 300 and 900 ms. Unreliable biases at 300 ms might arise because disengagement of attention from the emotional stimulus has already begun by this point. If this is the case, then future experiments could use a higher resolution of SOAs to track the time course of engagement and disengagement processes, as indicated by changes in reliability over time. Because covert spatial attention can be rapidly redirected (Buschman and Miller, 2009; Müller and Rabbitt, 1989), future studies might also investigate the reliability of attentional biases for SOAs shorter than 100 ms.

Previous studies have shown avoidance of threatening stimuli at longer SOAs (i.e., slower response times when probes replace the angry face; Booth 2014; Koster et al., 2006; Onnis et al., 2011). If this avoidance of threat reflects a stable individual difference that is consistent from trial to trial, we would expect the bias to also become reliable at the longer 900 ms SOA. Instead, biases were unreliable still at this very long SOA. Assessments of reliability with such long SOAs are rare, but Bar-Haim and colleagues (2010) report significant split-half reliability of attentional biases to threat words with an SOA of 1000 ms ($r = .45$). However, the authors suggest that their biases might reflect the atypical characteristics of their sample, who lived in dangerous locations near the Gaza Strip and reported very high levels of anxiety, PTSD, and depression. Our sample was drawn from a non-selected undergraduate student population, and therefore our findings may not generalise to a clinically anxious population for whom later attentional processes might be more consistently deployed from trial to trial.

Although we show that attentional biases are statistically reliable at the short 100 ms SOA, the reliability estimates are still not sufficiently high to justify their use as an individual difference measure in studies where the goal is to correlate bias with other traits. In these experiments, we chose to use a specific set of stimulus and task parameters (face size, eccentricity, probe detection task, etc.), which we held constant while we manipulated SOA. It is possible that other stimulus parameters might have led to better reliability at the 100 ms SOA, and the search for those parameters might lead to a better dot-probe task. However, the limitations on reliability might also be statistical and, therefore, difficult to overcome. Several researchers have noted the poor reliability in a number of cognitive tasks (not just the dot-probe task; e.g., Ross, Richler, & Gauthier, 2015; Strauss et al., 2005) that rely on difference scores as a dependent measure. A recent report by Hedge and colleagues (Hedge, Powell, & Sumner, 2017) shows that this problem extends even to well established and robust effects in cognitive psychology including Stroop, flanker, and spatial cueing effects. While *response times* on such tasks can be highly reliable (as they are in our dot-probe experiments reported here), the *difference between two response times*, which is calculated to isolate a specific set of cognitive processes, often has very poor reliability. Hedge and colleagues show that this is because such difference scores, by design, lack inter-individual variability. This is what makes them ideal for experimental research, but makes them poor measures of individual differences. Thus, while we show here that the dot-probe task will produce a more reliable measure of attentional bias when it specifically targets early attentional selection or orienting (i.e., using a short SOA), there may be hard limits on how reliable any difference measure can be.

Across the experiments reported here, we did not observe an attentional bias towards angry faces. Indeed, a significant bias was observed only in the 100 ms SOA condition and only in Experiment 1A and that was a bias away from the angry face. Given that this effect did not replicate in the other experiments, it is most likely a spurious finding. Examination of the literature similarly suggests that biases at 100 ms are inconsistent, with some studies reporting a bias towards threat (Holmes, Green, & Vuilleumier, 2005; Cooper, & Langton, 2006) and some studies showing a bias

away from threat (Koster et al., 2005; Mogg et al., 1997). Our findings are, therefore, consistent with the literature (e.g., Bar-Haim et al., 2007) showing that attentional biases to threat are not consistently observed in non-anxious participants on the dot-probe task, even though neural measures do reveal the existence of biases (Eimer & Kiss, 2007; Grimshaw et al., 2014; Holmes, Bradley, Nielsen, & Mogg, 2009; Kappenman et al., 2014, 2015).

One might question what it means for a bias to be reliable, but not significantly different from zero. It is worth noting that the zero point lies midway on a continuum that lies between attentional bias towards threat, and attentional bias away from threat. Thus, it is possible for individuals to have a stable (i.e., reliable) attentional bias either towards or away from threat, while no systematic directional bias is observed at the group level. Table 1 further shows that in other studies, reliability is not dependent on the existence of an attentional bias. However, such a pattern of biases (both towards and away from threat) is not consistent with neural measures that suggest a population-level attentional bias to threat, even in non-anxious individuals. Our findings provide no obvious explanation for this discrepancy. One possibility is that neural measures are tapping different processes than the behavioural task. For example, the N2pc (which consistently shows an attentional bias to threat in non-anxious participants) might actually measure attentional selection but not orienting (Hilimire, Hickey, & Corballis, 2013; Sawaki & Luck, 2010). Because non-anxious participants are thought to have greater control over their attention (Eysenck et al., 2007; Eysenck & Derakshan, 2011), they might be able to suppress capture by the irrelevant threatening image, leading to an N2pc that reflects "attend-to-me" tagging of the emotional stimulus, without associated behavioural bias (see Sawaki & Luck, 2010, for a similar hypothesis regarding suppression of distraction by salient singletons). Notably, most of the studies that report an emotion-related N2pc in a dot-probe paradigm also report no such bias in RT (Grimshaw et al., 2014; Kappenman et al., 2014, 2015; Reutter et al., 2017).

Importantly, several findings now suggest that more direct (i.e., not response time based) correlates of attentional bias for dot-probe tasks might have better internal reliability. In addition to the good reliability of the N2pc (Kappenman et al., 2014, 2015; Reutter et al., 2017), attentional bias as reflected in lateral prefrontal activation to emotional stimuli has recently been shown to have moderate test–retest reliability (White et al., 2016). Eye-tracking measures might also provide more direct measures. For example, Waechter et al. (2014) tracked gaze in a free-viewing paradigm, finding that total dwell time on an emotional (compared to a neutral) face provided a reliable measure of bias. However, these direct measures do not necessarily correlate with response time measures (Kappenman et al., 2014, 2015; Reutter et al., 2017; Waechter et al., 2014), or with traits such as anxiety (Kappenman et al., 2014, 2015; but see Reutter et al., 2017); therefore, further investigation is needed to understand how these measures are related to attention to threat.

Individual differences in attentional bias have been proposed to account for a range of psychological phenomena, and therefore the reliable measurement of such biases is an important goal. While the poor internal reliability of attentional biases as measured by many tasks (not just the dot-probe) is now well documented (Cisler, Bacon, & Williams, 2009; Strauss, Allen, Jorgensen, & Cramer, 2005), there has been little consideration of how experimental manipulations can affect reliability. Many attentional bias researchers ignore the issue of reliability altogether. We, therefore, encourage a more systematic reporting of reliability in future research, so that we might develop better methods of capturing individual variability in these biases. We show here that the time course of underlying attentional processes is one important factor in determining the reliability of attentional bias measurements. Our findings suggest that early attentional selection and orienting show some level of consistency in the dot-probe task, but that this stability is fleeting.

Acknowledgements We thank Laura Kranz and Emma O'Brien for assistance with data collection. This research was supported by a Grant from the Royal Society of New Zealand Marsden Fund (VUW1307) to GG. Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

Compliance with ethical standards

Ethical standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the Victoria University of Wellington Human Ethics Committee. Informed consent was obtained from all individual participants included in the study.

Data availability The data sets generated and analysed for this study are not publicly available due to ethical constraints, but are available from the corresponding author on reasonable request.

References

Amir, I., Zvielli, A., & Bernstein, A. (2016). De(coupling) of our eyes and our mind's eye: A dynamic process perspective on attentional bias. *Emotion, 16*(7), 978–986. https://doi.org/10.1037/emo0000172.

Bar-Haim, Y., Holoshitz, Y., Eldar, S., Frenkel, T. I., Muller, D., Charney, D. S., Pine, D. S., Fox, N. A., & Wald, I. (2010). Life-threatening danger and suppression of attention bias to threat. *American Journal of Psychiatry, 167*(6), 694–698. https://doi.org/10.1176/appi.ajp.2009.09070956.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and non-anxious individuals: A meta-analytic

study. *Psychological Bulletin, 133*(1), 1–24. https://doi.org/10.1037/0033-2909.133.1.1.

Bishop, S. J. (2008). Neural mechanisms underlying selective attention to threat. *Annals of the New York Academy of Sciences, 1129*, 141–152. https://doi.org/10.1196/annals.1417.016.

Booth, R. W. (2014). Uncontrolled avoidance of threat: Vigilance-avoidance, executive control, inhibition and shifting. *Cognition and Emotion, 28*(8), 1465–1473. https://doi.org/10.1080/02699931.2014.882294.

Bradley, B. P., Mogg, K., Falla, S. J., & Hamilton, L. R. (1998). Attentional bias for threatening facial expressions in anxiety: Manipulation of stimulus duration. *Cognition and Emotion, 12*(6), 737–753. https://doi.org/10.1080/026999398379411.

Bradley, B. P., Mogg, K., Millar, N., Bonham-Carter, C., Fergusson, E., Jenkins, J., & Parr, M. (1997). Attentional biases for emotional faces. *Cognition and Emotion, 11*(1), 25–42. https://doi.org/10.1080/026999397380014.

Buschman, T. J., & Miller, E. K. (2009). Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations. *Neuron, 63*(3), 386–396. https://doi.org/10.1016/j.neuron.2009.06.020.

Cisler, J. M., Bacon, A. K., & Williams, N. L. (2009). Phenomenological characteristics of attentional biases towards threat: A critical review. *Cognitive Therapy and Research, 33*, 221–234.

Cooper, R. M., Bailey, J. E., Diaper, A., Stirland, R., Renton, L. E., Benton, C. P., Penton-Voak, I. S., Nutt, D. J., & Munafò, M. R. (2011). Effects of 7.5% CO_2 inhalation on allocation of spatial attention to facial cues of emotional expression. *Cognition and Emotion, 25*(4), 626–638. https://doi.org/10.1080/02699931.2010.508887.

Cooper, R. M. & Langton, S. R. H. (2006). Attentional bias to angry face using the dot-probe task? It depends when you look for it. *Behavior Research and Therapy, 44*, 1321–1329. https://doi.org/10.1016/j.brat.2005.10.004.

Dear, B. F., Sharpe, L., Nicholas, M. K., & Refshauge, K. (2011). The psychometric properties of the dot-probe paradigm when used in pain-related attentional bias research. *The Journal of Pain, 12*(12), 1247–1254. https://doi.org/10.1016/j.jpain.2011.07.003.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology, 111*(2), 225–236. https://doi.org/10.1037/0021-843X.111.2.225.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park: Sage.

Eimer, M., & Kiss, M. (2007). Attentional capture by task-irrelevant fearful faces is revealed by the N2pc component. *Biological Psychology, 74*(1), 108–112. https://doi.org/10.1016/j.biopsycho.2006.06.008.

Eysenck, M. W., & Derakshan, N. (2011). New perspectives in attentional control theory. *Personality and Individual Differences, 50*(7), 955–960. https://doi.org/10.1016/j.paid.2010.08.019.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336.

Grimshaw, G. M., Foster, J. J., & Corballis, P. M. (2014). Frontal and parietal EEG asymmetries interact to predict attentional bias to threat. *Brain and Cognition, 90*, 78–86. https://doi.org/10.1016/j.bandc.2014.06.008.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0935-1.

Hilimire, M. R., Hickey, C., & Corballis, P. M. (2013). Target resolution in visual search involves the direct suppression of distractors: Evidence from electrophysiology. *Psychophysiology, 49*, 504–509. https://doi.org/10.1111/j.1469-8986.2011.01326.x.

Holmes, A., Bradley, B. P., Nielsen, K., M., & Mogg, K. (2009). Attentional selectivity for emotional faces: Evidence from human electrophysiology. *Psychophysiology, 46*(1), 62–68. https://doi.org/10.1111/j.1469-8986.2008.00750.x.

Holmes, A., Green, S., & Vuilleumier, P. (2005). The involvement of distinct visual channels in rapid attention towards fearful facial expressions. *Cognition and Emotion, 19*(6), 899–922. https://doi.org/10.1080/02699930441000454.

Kappenman, E. S., Farrens, J. L., Luck, S. J., & Proudfit, G. H. (2014). Behavioral and ERP measures of attentional bias to threat in the dot-probe task: Poor reliability and lack of correlation with anxiety. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.01368.

Kappenman, E. S., MacNamara, A., & Proudift, G. H. (2015). Electrocortical evidence for rapid allocation of attention to threat in the dot-probe task. *Social Cognitive and Affective Neuroscience, 10*(4), 577–583. https://doi.org/10.1093/scan/nsu098.

Koster, E. H. W., Crombez, G., Verschuere, B., & De Houwer, J. (2004). Selective attention to threat in the dot probe paradigm: Differentiating vigilance and difficulty to disengage. *Behaviour Research and Therapy, 42*, 1183–1192. https://doi.org/10.1016/j.brat.2003.08.001.

Koster, E. H. W., Crombez, G., Verschuere, B., Van Damme, S., & Wiersema, J. R. (2006). Components of attentional bias to threat in high trait anxiety: Facilitated engagement, impaired disengagement, and attentional avoidance. *Behaviour Research and Therapy, 44*, 1757–1771. https://doi.org/10.1016/j.brat.2005.12.011.

Koster, E. H. W., Verschuere, B., Crombez, G., & Van Damme, S. (2005). Time-course of attentional bias for threatening pictures in high and low trait anxiety. *Behaviour Research and Therapy, 45*, 1087–1098. https://doi.org/10.1016/j.brat.2004.08.004.

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

MacLeod, J. W., Lawrence, M. A., McConnell, M. M., Eskes, G. A., Klein, R. M., & Shore, D. L. (2010). Appraising the ANT: Psychometric and theoretical considerations of the Attention Network Task. *Neuropsychology, 24*(5), 637–651. https://doi.org/10.1037/a0019803.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7.

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect size. *Psychonomic Bulletin and Review, 20*, 819–858. https://doi.org/10.3758/s13423-013-0404-5.

Mogg, K., & Bradley, B. P. (1999). Some methodological issues in assessing attentional biases for threatening faces in anxiety: A replication study using a modified version of the probe detection task. *Behavior Therapy and Research, 37*, 595–604. https://doi.org/10.1016/S0005-7967(98)00158-2.

Mogg, K., Bradley, B. P., De Bono, J., & Painter, M. (1997). Time course of attentional bias for threat information in non-clinical anxiety. *Behaviour Research and Therapy, 35*(4), 297–303. https://doi.org/10.1016/S0005-7967(96)00109-X.

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

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of*

Personality and Social Psychology, 80(3), 381–396. https://doi.org/10.1037/0022-3514.80.3.381.

Okon-Singer, H., Lichtenstein-Vidne, L., & Cohen, N. (2013). Dynamic modulation of emotional processing. *Biological Psychology, 92*, 480–491. https://doi.org/10.1016/j.biopsycho.2012.05.010.

Onnis, R., Dadds, M. R., & Bryant, R. A. (2011). Is there a mutual relationship between opposite attentional biases underlying anxiety? *Emotion, 11*(3), 582–594. https://doi.org/10.1037/a0022019.

Posner, M. I., Cohen, Y., & Rafal, R. D. (1982). Neural systems control of spatial attention. *Philosophical Transactions of the Royal Society B: Biological Sciences, 298*(1089), 187–198. https://doi.org/10.1098/rstb.1982.0081.

Pourtois, G., Schettino, A., & Vuilleumier, P. (2013). Brain mechanisms for emotional influences on perception and attention: What is magic and what is not. *Biological Psychology, 92*, 492–512. https://doi.org/10.1016/j.biopsycho.2012.02.007.

R Core Team (2015). R: A Language and Environment for Statistical Computing [Computer Software]. Retrieved from http://www.R-project.org.

Ratcliff, R. (1993). Methods for dealing with response time outliers. *Psychological Bulletin, 114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510.

Reutter, M., Hewig, J., Wieser, M. J., & Osinsky, R. (2017). The N2pc component reliably captures attentional bias in social anxiety. *Psychophysiology, 54*, 519–527. https://doi.org/10.1111/psyp.12809.

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology, 125*(6), 840–851. https://doi.org/10.1037/abn0000184.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods, 47*, 736–743.

Sawaki, R., & Luck, S. J. (2010). Capture versus suppression of attention by salient singletons: Electrophysiological evidence for an automatic attend-to-me signal. *Attention Perception and Psychophysics, 72*(6), 1455–1470. https://doi.org/10.3758/APP.72.6.1455.

Schäfer, J., Bernstein, A., Zvielli, A., Hölfer, M., Wittchen, H., & Schönfeld, S. (2016). Attentional bias temporal dynamics predict posttraumatic stress symptoms: A prospective-longitudinal study among soldiers. *Depression and Anxiety, 33*, 630–639. https://doi.org/10.1002/da.22526.

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

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x.

Spielberger, C. D., & Sydeman, S. J. (1994). State-trait anxiety inventory and state-trait anger expression inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 292–321). Hillsdale: Lawrence Erlbaum Associates.

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

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional Stroop tasks. *Assessment, 12*(3), 330–337. https://doi.org/10.1177/1073191105276375.

Szymanski, J., & O'Donohue, W. (1995). Fear of Spiders Questionnaire. *Journal of Behavioral Therapy and Experimental Psychology, 26*(1), 31–34. https://doi.org/10.1016/0005-7916(94)00072-T.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research, 168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006.

Van Bockstaele, B., Vershuere, B., Koster, E. H. W., Tibboel, H., De Houwer, J., & Crombez, G. (2011). Differential predictive power of self report and implicit measures on behavioural and physiological fear responses to spiders. *International Journal of Psychophysiology, 79*, 166–174. https://doi.org/10.1016/j.ijpsycho.2010.10.003.

Waechter, S., Nelson, A. L., Wright, C., Hyatt, A., & Oakman, J. (2014). Measuring attention bias to threat: Reliability of dot probe and eye movement indices. *Cognitive Therapy and Research, 38*(3), 313–333. https://doi.org/10.1007/s10608-013-9588-2.

Waechter, S., & Stolz, J. A. (2015). Trait anxiety, state anxiety, and attentional bias to threat: Assessing the psychometric properties of response time measures. *Cognitive Therapy and Research, 39*(4), 441–458. https://doi.org/10.1007/s10608-015-9670-z.

White, L. K., Britton, J. C., Sequiera, S., Ronkin, E. G., Chen, G., Bar-Haim, Y., Pine, D. S. (2016). Behavioral and neural stability of attentional bias to threat in healthy adolescents. *NeuroImage, 136*, 84–93. https://doi.org/10.1016/j.neuroimage.2016.04.058.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*, 671–684. https://doi.org/10.3758/BRM.42.3.671.

Williams, B. J., & Kaufmann, L. M. (2012). Reliability of the go/no-go association task. *Journal of Experimental Social Psychology, 48*, 879–891. https://doi.org/10.1016/j.jesp.2012.03.001.

Yiend, J. (2010). The effects of emotion on attention: A review of attentional processing of emotional information. *Cognition and Emotion, 24*(1), 3–47. https://doi.org/10.1080/02699930903205698.

Zvielli, A., Bernstein, A., & Koster, E. H. W. (2015). Temporal dynamics of attentional bias. *Clinical Psychological Science, 3*(5), 772–788. https://doi.org/10.1177/2167702614551572.