Behavior Research Methods (2019) 51:2748–2760
https://doi.org/10.3758/s13428-018-1147-z

The reliability of attentional biases for emotional images measured using a free-viewing eye-tracking paradigm

Christopher Sears[1] · Leanne Quigley[1] · Amanda Fernandez[1] · Kristin Newman[1] · Keith Dobson[1]

Published online: 22 October 2018

Abstract

Cognitive theories of anxiety disorders and depression posit that attentional biases play a role in the development, maintenance, and recurrence of these disorders. Several paradigms have been used to examine attentional biases in anxiety and depression, but information on the reliability of different attentional bias indices is limited. In this study we examined the internal consistency and 6-month test–retest reliability of attentional bias indices derived from a free-viewing eye-tracking paradigm. Participants completed two versions of an eye-tracking task—one that used naturalistic images as stimuli, and one that used face images. In both tasks, participants viewed displays of four images, each display consisting of one threat image, one sad image, one positive/happy image, and one neutral image. The internal consistency of the fixation indices (dwell time and number of fixations) for threat, sad, and positive images over the full 8-s display was moderate to excellent. When the 8-s display was divided into 2-s intervals, the dwell times for the 0- to 2-s and 2- to 4-s intervals showed lower reliability, particularly for the face images. The attentional bias indices for the naturalistic images showed adequate to good stability over the test–retest period, whereas the test–retest reliability estimates for the face images were in the low to moderate range. The implications of these results for attentional bias research are discussed.

Keywords Attentional bias · Eye tracking · Reliability · Test-retest

Researchers studying selective attention in psychopathology have documented biases in the allocation of attention to emotional information. Most commonly, attentional biases have been studied in the context of anxiety disorders and depression. Many studies have shown that anxious individuals exhibit preferential orienting to, and facilitated detection of, threatening stimuli (Armstrong & Olatunji, 2012; Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van IJzendoorn, 2007). In contrast, for depressed individuals, attentional biases typically manifest when emotional stimuli are attended to over longer intervals (Armstrong & Olatunji, 2012; Yiend, 2010). Depression-associated attentional biases are characterized by increased attention to dysphoric stimuli (e.g., depression-related words and images related to sadness) and decreased attention to positive stimuli (e.g., positively valenced words and images; Armstrong & Olatunji, 2012; Peckham, McHugh, & Otto, 2010). According to cognitive models of anxiety and depression, attentional biases are stable, trait-like characteristics that contribute to the development, maintenance, and recurrence of these disorders (Beck & Clark, 1997; Gotlib & Joormann, 2010).

Studies of attentional biases in anxiety and depression have primarily relied on response time tasks to examine attention. These include the dot-probe task (e.g., Bradley, Mogg, White, Groom, & De Bono, 1999; Joormann & Gotlib, 2007; see Bar-Haim et al., 2007, and Peckham et al., 2010, for reviews), the emotional Stroop task (see Epp, Dobson, Dozois, & Frewen, 2012, and Williams, Mathews, & MacLeod, 1996, for reviews), the dichotic listening task (e.g., Ingram, Steidtmann, & Bistricky, 2008; Mathews & MacLeod, 1986), visual search tasks (e.g., Rinck, Becker, Kellermann, & Roth, 2003; Wenzlaff, Rude, Taylor, Stultz, & Sweatt, 2001), and the deployment-of-attention task (e.g., McCabe, Gotlib, & Martin, 2000). A great deal has been learned about attentional biases using response time tasks, but they have several important limitations. For one, these tasks do not provide direct measures of attention, but rather infer where the focus of attention is at a single moment in time on the basis of response latencies to stimuli. In most dot-probe tasks, for

✉ Christopher Sears
sears@ucalgary.ca

[1] Department of Psychology, University of Calgary, Calgary, Alberta, Canada

example, the probe is presented 500–1,000 ms following the presentation of a pair of words or images, with faster responding to probed stimuli being interpreted as reflecting the spatial focus of attention at the time of the probe presentation. The design of many of these tasks precludes the possibility of measuring prolonged engagement with stimuli or changes in attention over time, given their reliance on speeded responses to stimuli measured at a specific point in time.

Another limitation is that the psychometric properties of response time measures of attentional biases have rarely been evaluated or reported. Only recently have researchers considered this issue, and as we discuss below, several studies have found that attentional bias scores based on commonly used tasks have poor reliability (Schmukle, 2005; Staugaard, 2009; Waechter, Nelson, Wright, Hyatt, & Oakman, 2014; Waechter & Stolz, 2015). *Reliability* refers to the consistency of a measure across time, raters, methods, or items comprising the measure (e.g., Crocker & Algina, 1986). According to classical test theory, reliability is the proportion of total variance in a measure that is due to true score variance (i.e., variability in the construct of interest), as opposed to measurement error (Lord & Novick, 1968).

Two commonly reported types of reliability are internal consistency and test–retest reliability. *Internal consistency* is the interrelatedness of items of the test or measure. It is typically estimated using Cronbach's alpha, which varies as a function of the number of items in a test, the average covariance between items, and the total variance (Cronbach, 1951). For behavioral measures such as attentional bias tasks, the "items" are the individual trials of the task. Another common measure of internal consistency is *split-half reliability*, which involves splitting the items into two halves and then computing the correlation between the item halves (Cronbach, 1946). *Test–retest reliability* is the stability of a measure over time. It is estimated by correlating scores from the same measure administered at different time points. Test–retest reliability is relevant to the trait versus state nature of a construct: If a test measures a trait-like variable, then the scores on the test at different time points should correlate highly and instability would reflect measurement error; if a test measures a state-like variable, then the correlation between scores on the test at different time points will be limited by the extent to which the construct is expected to vary over time. Instability may, therefore, reflect real changes in the construct of interest and/or measurement error.

Reliability of measurement is essential for attentional bias research, since reliability is a precondition for validity, defined as the degree to which a test measures what it purports to measure (Kelley, 1927). Although no tests are free of measurement error, the tests used in psychological research must demonstrate some minimum level of reliability in order to justify their use. Generally accepted guidelines suggest that reliability is excellent if values are ≥ .90, good if values are ≥ .80, moderate if values are ≥ .70, questionable if values are ≥ .60, and poor/inadequate if values are < .60. These guidelines usually refer to the internal consistency (i.e., Cronbach's alpha) of psychological tests and self-report questionnaires. As was noted by Waechter et al. (2014), there are no established guidelines for the reliability of behavioral paradigms such as attentional bias tasks, and thus it is unclear what standards should be used to judge the minimum acceptable value for their reliability. Similarly, standards for test–retest reliability, even for psychological tests and questionnaires, are unclear. Although some researchers have proposed a minimum acceptable value of ≥ .70, test–retest reliability depends on the time elapsed between test and retest and the construct being measured, among other factors. The magnitude of the correlation between two administrations of the same measure would be expected to decrease as the time between administrations increases, due to factors that may be related or unrelated to the construct of interest (Allen & Yen, 1979). Thus, the test–retest reliability of a measure must be interpreted with these considerations in mind.

Evaluations of the reliability of attentional bias measures

As we noted above, there have been only a few attempts to estimate the reliability of attentional bias measures. Schmukle (2005) evaluated the reliability of attentional bias scores based on the dot-probe task in an unselected undergraduate student sample. Two versions of the dot-probe task were used—one that used threat-related and neutral word pairs as stimuli, and one that used threat-related and neutral image pairs as stimuli. Schmukle reported very low estimates of both internal consistency (− .16 to .28) and 1-week test–retest reliability (− .22 to .32) for the attentional bias scores for both versions of the dot-probe task. Staugaard (2009) reported similarly low estimates of internal consistency (− .58 to .37) and test–retest reliability over 1 to 2 weeks (− .24 to .26) for attentional bias scores for a version of the dot-probe task that used emotional (angry or happy) and neutral face image pairs as stimuli. Waechter et al. (2014) examined the internal consistency of attentional bias scores for a dot-probe task using angry, disgusted, and happy face images paired with neutral face images, in a sample of university students who were high or low in social anxiety. They reported reliability estimates that ranged from − .16 to .30 for the various bias indices. Waechter and Stolz (2015) assessed the internal consistency of the dot-probe task using angry and happy face images paired with neutral face images. Their sample comprised high and low trait-anxious undergraduate students. On the basis of their analyses, they concluded that the reliability estimates for the bias scores for angry and happy faces were unacceptably low (.04 to .42).

Psychometric evaluations of attentional bias scores from response time tasks other than the dot-probe task are also rare in the literature. A few studies have examined the test–retest reliability of the emotional Stroop task (e.g., Eide, Kemp, Silberstein, Nathan, & Stough, 2002; Strauss, Allen, Jorgensen, & Cramer, 2005). Eide et al. examined the 1-week test–retest reliability of attentional bias scores from the emotional Stroop task in a nonclinical sample. They reported low reliability for attentional bias scores for both depression-related (.24) and positive (− .11) words. Similarly, Strauss et al. reported poor 1-week test–retest reliability (− .27 to .23) of attentional bias scores for happy, sad, angry, and anxiety-related words in a nonclinical undergraduate student sample. Taken together, although only a small number of psychometric evaluations are present in the literature, the evidence suggests that attentional bias scores derived from response time tasks have very low reliability.

An alternative methodology to measure attentional biases to words and images is eye-gaze tracking. Several researchers have used eye-gaze tracking to measure attention to emotional images and words in anxious individuals (e.g., Nelson, Purdon, Quigley, Carriere, & Smilek, 2015; Quigley, Nelson, Carriere, Smilek, & Purdon, 2012; Schofield, Johnson, Inhoff, & Coles, 2012) and in depressed and dysphoric individuals (e.g., Arndt, Newman, & Sears, 2014; Duque & Vázquez, 2015; Eizenman et al. 2003; Kellough, Beevers, Ellis, & Wells, 2008; Leyman, De Raedt, Vaeyens, & Philippaerts, 2011; Newman & Sears, 2015; Sears, Newman, Ference, & Thomas, 2011; Sears, Thomas, LeHuquet, & Johnson, 2010; Soltani et al., 2015). Many of these studies have used a free-viewing paradigm that involves tracking and recording participants' eye movements as they view displays of images of different emotional valence. A meta-analysis of these studies found evidence of an orienting bias toward threatening stimuli in anxious relative to nonanxious individuals, as well as increased attention to negative stimuli and decreased attention to positive stimuli in depressed relative to nondepressed individuals (Armstrong & Olatunji, 2012). The major advantage of eye tracking is that it provides a direct measure of the allocation of attention, by recording fixations to stimuli as they are examined, since the direction of gaze and the focus of attention are tightly coupled (Wright & Ward, 2008). The continuous nature of eye-tracking data allows for the measurement of changes in attention to stimuli over time, as opposed to the single "snapshot" of attention captured in response time tasks. An additional advantage is that, because manual responses are not required, potential differences in response speed due to aging and other individual differences are eliminated.

A few studies have focused on the reliability of attentional bias indices measured using eye-tracking tasks (Lazarov, Abend, & Bar-Haim, 2016; Skinner et al., 2018; Waechter et al., 2014). Waechter et al. reported the internal consistency of attentional bias indices in high and low socially anxious undergraduate students. Their participants viewed pairs of images consisting of one emotional face (displaying an expression of happiness, anger, or disgust) paired with a neutral face. The pairs of images were presented vertically for 5,000 ms. Waechter et al. used the proportion of viewing time to measure attentional bias, which reflects the proportion of time spent attending to a particular image type relative to the other image in the display. Waechter et al. reported Cronbach's alpha estimates of .94 to .96 for the proportions of viewing time for the different image types, indicating excellent reliability of this measure.

Interestingly, when Waechter et al. (2014) divided the data from the 5,000-ms presentation into 1,000-ms intervals, they found that the reliability estimates for proportion of viewing time were low (negative, in fact) for the 0- to 1,000-ms interval, but increased for the subsequent 1,000-ms intervals. More specifically, for the 0- to 1,000-ms interval, Cronbach's alpha estimates ranged from − .72 to − .48. Reliability estimates increased to moderate levels for the 1,001- to 2,000-ms interval (.47 to .60). For each 1,000-ms interval between 2,001 and 5,000 ms, the reliability estimates for the proportion of viewing time ranged from .58 to .69 for the different face image types. These results suggest that attentional bias indices measured in eye-tracking tasks have excellent reliability overall when indices are averaged over time, but their reliability varies substantially over the presentation time of the display, with indices based on early time intervals showing poor reliability. Waechter et al. attributed the poor reliability of the attentional bias indices for the first few intervals to an artifact that they had observed in their study—namely, the tendency of participants to look at the top image of the two-image displays first, regardless of its emotional valence (a "look up" bias). Previous eye-tracking research using two-image horizontal displays has similarly revealed a stable "look left" bias, such that participants tend to fixate first on the left image (Nelson, Quigley, Carriere, Purdon, & Smilek, 2010).

Lazarov et al. (2016) evaluated the internal consistency and 1-week test–retest reliability of viewing times on threat and neutral stimuli in participants diagnosed with social anxiety disorder, as well as in high and low socially anxious undergraduate students. Participants' eye movements were tracked as they viewed 6,000-ms displays of 16 face images (4 × 4 matrices), consisting of equal numbers of disgust and neutral emotional expressions. Cronbach's alpha estimates for the total viewing times on threat faces and the total viewing times on neutral faces were high, both at Time 1 (.95 and .95, respectively) and at Time 2, 1 week later (.89 and .92, respectively). One-week test–retest reliability was moderate for both total viewing time on the threat faces (r = .68) and total viewing time on the neutral faces (r = .62). Lazarov et al. did not evaluate the internal consistency of the viewing time indices

across the display period (i.e., by dividing the presentation time into shorter intervals). They did, however, report test–retest reliabilities for first-fixation indices (i.e., latency to first fixation, first-fixation location, and first-fixation dwell time), which were all nonsignificant (*r*s = .06, .26, and .08, respectively). Thus, despite the more visually complex display (16 images) used by Lazarov et al., early attentional bias indices showed low reliability, similar to the results reported by Waechter et al. (2014).

Most recently, Skinner et al. (2018) examined the reliability of attentional bias indices to general, affective, and sensory threat words in an unselected sample of undergraduate students. Participants viewed pairs of threat and neutral words, presented on the left and right of a display. Each pair of words was presented for 4,000 ms while participants' eye movements were tracked and recorded. Participants completed a retest of the eye-tracking task in the same session after a 30-min break. Skinner et al. reported the test–retest reliability (measured by the intraclass correlation coefficient; ICC), measurement error (measured by the standard error of measurement), and internal consistency (measured by Cronbach's alpha) of several attentional bias indices, reflecting overall attention, early attention, and late attention. Test–retest reliability estimates varied across the different attentional bias indices and categories of threat words. The ICCs were moderate for overall viewing time on general and affective threat words during the 4,000-ms viewing period (.71 and .61, respectively), whereas the viewing times on sensory threat words showed lower reliability (ICC = .20). Low test–retest reliability was observed for attentional bias indices based on intervals of shorter duration and early time intervals (e.g., the ICCs ranged from − .31 to .12 for overall viewing time during the first 500 ms for the different categories of threat words). Measurement error was relatively low across the attentional bias indices, indicating consistency across the test and retest sessions. Skinner et al. therefore suggested that the low test–retest reliability observed for several of the attentional bias indices reflected low variance between participants for these indices, rather than high measurement error. Cronbach's alpha values were generally high across the attentional bias indices, including those reflecting early attention. In contrast to the results of Waechter et al. (2014), Skinner et al. reported high internal consistency for viewing times on threat words during the first 500 ms of the viewing period (Cronbach's alphas ranged from .98 to .99). The lowest Cronbach's alpha values were observed for first-fixation duration on threat words, but even these were moderate (.57 to .70). It is unclear why early attentional bias indices would show low internal consistency for Waechter et al. and Lazarov et al. (2016), yet moderate to high internal consistency in Skinner et al.'s study. Differences in the task stimuli (words vs. face images) and methodologies are likely contributors to these inconsistent results.

The present study

Although the studies reviewed above provide important information about the reliability of attentional bias indices derived from eye-tracking tasks, there are significant gaps in the existing literature. For one, there are no published data on the stability of eye-tracking indices over extended test–retest intervals. Recall that Skinner et al. (2018) examined test–retest reliability within a single testing session, and Lazarov et al. (2016) examined 1-week test–retest reliability. If attentional biases reflect a stable, trait-like phenomenon, then attentional bias indices should have adequate test–retest reliability over longer test–retest intervals. In the present study we evaluated the 6-month test–retest reliability of attentional bias indices obtained from a free-viewing eye-tracking task. Another novel feature of our study was that we examined the internal consistency and test–retest reliability of attentional bias indices for both faces images and naturalistic images. Most of the previous studies have used either face images (Lazarov et al., 2016; Waechter et al., 2014) or words (Skinner et al., 2018), and no study has compared the reliability of attentional biases for face images and naturalistic images within the same sample of participants. The two previous studies that used face images (Lazarov et al., 2016; Waechter et al., 2014) used samples selected to include either individuals scoring high or low on measures of social anxiety or individuals with social anxiety disorder, whereas in our study participants were not selected on the basis of their social anxiety. Our study also differed from previous studies by using displays of four images instead of the two-image or two-word displays used in previous research (with the exception of Lazarov et al., 2016, who used displays of 16 face images). Displays of four images are more complex and create more competition between stimuli for attention than do two-image displays, which likely creates a more natural viewing situation. We also used two different eye-tracking measures to measure attentional bias (dwell time and number of fixations), which allowed us to compare their reliability and increased the usefulness of our findings for other investigators. Together, these design features allow our study to contribute significantly to the small existing literature on the reliability of attentional bias indices derived from eye-tracking tasks.

Method

Participants

The final sample consisted of 85 women with a mean age of 26.31 years (*SD* = 10.84; range from 18 to 63); two participants in the original sample were excluded due to poor calibration that led to significant missing data (> 25%). The participants consisted of community members (38 participants,

45% of the total sample) and university students (47 participants, 55% of the total sample). Recruitment was carried out using an online research participation system and by placing posters on campus and in the community. Participants were recruited as part of a longitudinal study examining predictors of depression relapse. For the purpose of the longitudinal study, participants were classified as currently depressed, previously depressed, or never depressed during their first lab visit. The sample for the present study consisted of 60 participants classified as previously depressed (71% of the total sample) and 25 participants classified as never depressed participants (29% of the total sample).[1] We selected participants who were asymptomatic throughout the duration of the study (i.e., who reported minimal symptoms of depression at the first and second laboratory visits, as assessed by the measures described below), to minimize potential mood state effects on attentional biases. In exchange for taking part in the study, for each laboratory visit participants received either bonus credit in a psychology course (1% of their final grade) or a $25 (CAN) gift card.

Measures

Table 1 lists descriptive statistics for the measures administered during the first and second laboratory visits. Participants completed the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996), a 21-item self-report measure that is used to assess depression symptom severity over the previous 2 weeks. Each item is rated from 0 to 3, with a total possible score of 63. Higher scores indicate more symptoms of depression. BDI scores greater than 20 are generally interpreted as reflecting higher than average depression symptoms, whereas scores less than 12 correspond to minimal depression symptoms (Dozois, Dobson, & Ahnberg, 1998). Participants also completed the Patient Health Questionnaire-9 (PHQ-9; Spitzer, Kroenke, Williams, & the Patient Health Questionnaire Primary Care Study Group, 1999), a nine-item depression scale based on the diagnostic criteria for major depressive disorder, as described in the *Diagnostic and Statistical Manual*, 5th edition (DSM-5; American Psychiatric Association, 2013). As can be seen in Table 1, mean BDI-II and PHQ scores for the sample were very low, reflecting minimal symptoms of depression.

[1] The primary internal consistency and test–retest reliability analyses were conducted separately for participants classified as previously depressed or as never depressed, to explore whether this classification influenced the results. No systematic differences between the samples were observed. The same comparisons between the student and community samples revealed minor differences for threat images and threat faces (for threat images, reliability was higher for the student sample; for threat faces, reliability was higher for the community sample). The results are reported for the full sample, given the lack of consistent substantive differences. Additional research with larger samples may be warranted, however, especially for researchers interested in potential differences between student and community samples.

Table 1 Descriptive statistics for the measures administered during the first (Time 1) and second (Time 2) laboratory visits

	Time 1	Time 2	*t* Statistic	*p*
BDI-II	3.7 (3.4)	3.2 (3.2)	1.22	.22
PHQ	2.0 (2.1)	2.1 (2.6)	0.59	.55
BAI	6.1 (6.6)	6.4 (6.7)	0.45	.65
Mood rating	2.8 (1.4)	3.0 (1.2)	1.02	.30

BDI-II = Beck Depression Inventory. PHQ = Patient Health Questionnaire. BAI = Beck Anxiety Inventory. Mood rating = current mood rating from – 5 (*very negative*) to + 5 (*very positive*). Standard deviations are in parentheses

Participants' mood state was measured using an 11-point mood rating scale. The scale consists of a horizontal line with endpoints ranging from – 5 (labeled *very negative*) to + 5 (labeled *very positive*), with a midpoint of 0 (labeled *neutral*). Participants chose one of the 11 points on the scale to represent their current mood. The 11-point rating scale has been used successfully in several studies (e.g., Newman & Sears, 2015; Frayn, Sears, & von Ranson, 2016; Speirs, Belchev, Fernandez, Korol, & Sears, 2018).

Stimuli for eye tracking

The naturalistic stimuli consisted of 256 color images, divided equally among four categories: sad, threat, positive, and neutral. The majority of images were collected from the Internet, and the remainder from the International Affective Picture System database (Lang, Bradley, & Cuthbert, 1997). These were the same set of images used by Newman and Sears (2015). The sad images included scenes of people appearing sad and unhappy, neglected animals, scenes of poverty and illness, and gloomy landscapes. The threat images included themes of threat and injury, such as people being threatened with weapons, people with physical injuries (e.g., a burn on an arm), dangerous situations (e.g., a person walking along a cliff), vehicle accidents, and menacing animals. The positive images showed people smiling and laughing, children playing, rabbits and kittens, and vacation activities and destinations (e.g., a beach at a tropical resort). The neutral images were selected to include people in various activities and to have no obvious positive or negative theme (e.g., a woman talking on the telephone; a group of people having a meeting). They also included pictures of objects (e.g., a bicycle, a computer) and a variety of landscapes (e.g., office buildings). Care was taken to ensure that there were no systematic differences between the image categories (e.g., more landscapes in the neutral category than in the other categories, or more people in the sad than in the positive category).

The images were categorized by a separate group of undergraduate students ($N = 107$) prior to the study; for each image they were asked to choose one of four categories that best

described each image: (1) positive/happy, (2) sad/depressing/gloomy, (3) threatening/dangerous/fearful, or (4) neutral/no emotion. They were also asked to provide a valence rating using an 11-point scale, with – 5 representing *very negative* and + 5 representing *very positive*, with a midpoint of zero representing *neutral*. An image was chosen for use in the study only if at least 85% of the raters agreed as to its category. The mean valence ratings for the positive, sad, threat, and neutral images were 2.63, – 2.53, – 2.88, and 0.12, respectively. The mean valence ratings were significantly different from one another ($p < .001$), with the exception of the difference between the ratings for sad and threat images, $t(118) = 1.26$, $p = .21$.

The face stimuli consisted of 120 face images taken from the NimStim Database (Tottenham et al., 2009), which consists of face images depicting a variety of expressions, created for use in studies of face and emotion recognition. Four categories of faces were used, to correspond to the four types of naturalistic images: sad, threat, happy, and neutral. These were the same set of faces used by Soltani et al. (2015). The sad faces depicted individuals frowning or looking upset. The threat faces showed individuals that appeared either angry or fearful (e.g., furrowed brows and snarling mouths). The happy faces consisted of individuals who were smiling and friendly-looking. The neutral faces showed individuals with blank expressions (neither smiling nor frowning). There were equal numbers of open and closed mouths for each face type. Likewise, for each face type, equal numbers of male and female faces were presented (15 male and 15 female for each type). Participants were shown displays of four faces during data collection, each of a different actor with a different expression (sad, threat, happy, or neutral).

Apparatus

Eye movements were recorded using an EyeLink 1000 eye-tracking system (SR Research Ltd., Ottawa, Ontario), which uses infrared video-based tracking technology. The system has a 1000-Hz sampling rate, a temporal resolution of 2 ms, and an average gaze error of less than 0.5 degrees of visual angle. Stimuli were shown on a 24-in. LED monitor positioned approximately 60 cm away from the participant. Participants used a chin rest to minimize head movement while they viewed the images, to maximize tracking accuracy.

Procedure

The study was approved by the university research ethics board, and participants provided informed consent at the beginning of their first laboratory visit. Participants attended two laboratory visits, the second visit taking place approximately 6 months following the first ($M = 28.49$ weeks, $SD = 8.26$ weeks). The study procedure was the same for both visits. Participants first completed a battery of questionnaires, including the measures described previously (BDI-II, PHQ), as well as additional measures not relevant to the present hypotheses. They rated their current mood using the 11-point mood scale. The eye-tracking system was then calibrated for the participant, a procedure that required approximately 5 min. Following calibration, participants completed the eye-tracking tasks. The procedure was the same for the naturalistic image and face image versions of the task. For each task, participants were shown 30 sets of four images, with one sad image, one positive image, one threat image, and one neutral image in each set. One image was placed in each of the four corners of the display (top left, top right, bottom left, bottom right). Images were randomly assigned to the four corners of the display, and each image type was equally likely to appear in each corner across all 30 trials. Each set of images was presented for 8 s, and eye-tracking data were collected continuously throughout this interval. Each participant was shown a randomized sequence of 30 displays within each set. The order in which the naturalistic image and face image sets were presented was randomly determined across participants.

At the start of each presentation, participants were instructed to fixate on a black dot in the center of the display, to ensure proper gaze measurement and a central focus of attention before the images were presented. Participants were asked to look at the images as if they were watching a slide show, and they were told that there was no right or wrong way to view the images. Two displays of images were presented prior to data collection, to familiarize participants with the procedure and to confirm the accuracy of the calibration. Viewing the two sets images (30 four-image displays of naturalistic images and 30 four-image displays of faces) required approximately 10 min. Different face and naturalistic images were shown during the first and second laboratory visits, to prevent familiarity with the images from affecting participants' viewing behavior.

Results

Eye-tracking data preparation

The fixation data were processed using the EyeLink Data Viewer analysis software (SR Research) to filter for blinks, missing data, and other recording artifacts (using the default settings). To be included in the analyses, a fixation had to be at least 100 ms in duration; adjacent, sequential fixations that lasted less than 100 ms were merged into a single fixation. The first dependent variable was the total dwell time for each image type (in milliseconds) during the 8-s presentation. The total dwell time was computed for each image type on a trial-by-trial basis by summing the individual fixation times to each

image over the 8-s presentation duration and then averaging over the 30 trials. The mean total dwell times for each of the image types (naturalistic images and face images), for the first and second laboratory visits, are listed in Table 2. The second dependent variable was the total number of fixations to each image type during the 8-s presentation. The total number of fixations was computed for each image type on a trial-by-trial basis by summing the individual fixations to each image over the 8-s presentation and then averaging over the 30 trials. The mean total numbers of fixations for each of the image types (naturalistic images and face images), for the first and second laboratory visits, are listed in Table 3.

Analysis of mean total dwell time

Table 2 shows the mean total dwell times for the naturalistic and face images. These data were analyzed using repeated measures analysis of variance (ANOVA) and *t* tests to determine whether the differences in total dwell times across image types were consistent with the patterns observed in other eye-tracking studies. Only the Time 1 data were analyzed, given that they were collected under conditions directly comparable to the data collected in previous studies. For the total dwell time data from the naturalistic images, the ANOVA produced an effect of image type (threat, sad, positive, neutral), $F(3, 252) = 55.15$, $p < .001$, $\eta_p^2 = .39$. Follow-up comparisons using protected *t* tests showed that the total dwell times for positive images ($M = 2{,}083$ ms) were significantly longer than the total dwell times for threat images ($M = 1{,}689$ ms), sad images ($M = 1{,}761$ ms), and neutral images ($M = 1{,}000$ ms): $t(84) = 3.39$, $p < .01$; $t(84) = 2.72$, $p < .01$; and $t(84) = 15.68$, $p < .001$, respectively. We observed no significant difference between the total dwell times for threat and sad images, $t(84) = 1.58$, $p = .11$. These differences mirror those observed in the never depressed/control groups of previous studies, with the longest total dwell times for positive images and minimal differences between sad and threat images (e.g., Eizenman et al. 2003; Kellough et al., 2008; Speirs et al., 2018).

For the total dwell time data from the face images, the ANOVA produced an effect of image type (threat, sad, happy, neutral), $F(3, 252) = 34.37$, $p < .001$, $\eta_p^2 = .29$. As expected, follow-up comparisons showed that the total dwell times for happy faces ($M = 2{,}045$ ms) were significantly longer than the total dwell times for threat faces ($M = 1{,}442$ ms), sad faces ($M = 1{,}471$ ms), and neutral faces ($M = 1{,}529$ ms): $t(84) = 6.52$, $p < .001$; $t(84) = 5.82$, $p < .001$; and $t(84) = 6.34$, $p < .001$, respectively. There was no significant difference between the total dwell times for threat faces and sad faces, $t(84) = 1.11$, $p = .26$. Again, this pattern of results is consistent with the differences in total dwell times to emotional face images observed by other investigators in samples of never depressed/control participants (e.g., Leyman et al., 2011; Soltani et al., 2015).

Table 2 Mean total dwell times for each image type for the first (Time 1) and second (Time 2) laboratory visits, and correlations between Time 1 and Time 2

	Time 1	Time 2	*r*	*p*
Naturalistic images				
Threat images	1,689 (49.2)	1,858 (52.9)	.62	<.001
Sad images	1,761 (51.2)	1,583 (49.8)	.77	<.001
Positive images	2,083 (73.5)	2,075 (68.3)	.80	<.001
Neutral images	1,000 (33.7)	992 (33.7)	.61	<.001
Face images				
Threat faces	1,442 (31.7)	1,465 (34.0)	.47	<.001
Sad faces	1,471 (36.7)	1,389 (33.9)	.57	<.001
Happy faces	2,045 (68.8)	2,093 (74.6)	.48	<.001
Neutral faces	1,529 (23.2)	1,539 (30.1)	.36	<.001

Standard errors are in parentheses

Internal consistency of the fixation data

Internal consistency was estimated using Cronbach's alpha and the Spearman–Brown split-half reliability coefficient. To compute the alpha and split-half reliability coefficients, each of the 30 trials of the naturalistic images and the 30 trials of the face images was considered an item. For the split-half reliability calculation, the two item halves comprised odd trials and even trials. The internal consistency coefficients were calculated for total dwell time during the 8-s presentation duration for the different image types (threat, sad, positive/happy, and neutral) for the naturalistic and face image sets, separately. Cronbach's alpha and split-half reliability values for total dwell times on the naturalistic images are presented in Table 4. Cronbach's alpha and split-half reliability values for total dwell times on the face images are presented in Table 5.

As can be seen in Table 4, Cronbach's alpha for total dwell times for the naturalistic images at Times 1 and 2 ranged from

Table 3 Mean total numbers of fixations for each image type for the first (Time 1) and second (Time 2) laboratory visits, and correlations between Time 1 and Time 2

	Time 1	Time 2	*r*	*p*
Naturalistic images				
Threat images	6.59 (0.18)	7.17 (0.20)	.70	<.001
Sad images	7.26 (0.19)	6.55 (0.19)	.78	<.001
Positive images	8.15 (0.22)	8.06 (0.21)	.72	<.001
Neutral images	4.08 (0.12)	4.12 (0.13)	.71	<.001
Face images				
Threat faces	5.59 (0.14)	5.54 (0.12)	.60	<.001
Sad faces	5.68 (0.13)	5.36 (0.13)	.65	<.001
Happy faces	7.25 (0.18)	7.41 (0.23)	.48	<.001
Neutral faces	5.92 (0.12)	5.75 (0.11)	.39	<.001

Standard errors are in parentheses

Table 4 Cronbach's alpha and split-half reliability for the dwell time data for threat, sad, positive, and neutral naturalistic images, for each time interval during the first (Time 1) and second (Time 2) laboratory visits

	Total dwell time		0–2 s		2–4 s		4–6 s		6–8 s	
	α	S-h	α	S-h	α	S-h	α	S-h	α	S-h
Threat images										
Time 1	.85	.87	.63	.55	.67	.64	.57	.55	.63	.74
Time 2	.84	.86	.28	.43	.67	.61	.62	.63	.74	.80
Sad images										
Time 1	.86	.84	.14	.08	.59	.56	.62	.50	.77	.72
Time 2	.87	.85	.26	.35	.57	.50	.69	.71	.76	.75
Positive images										
Time 1	.92	.90	.58	.56	.71	.67	.79	.78	.86	.84
Time 2	.91	.91	.39	.35	.62	.71	.74	.72	.86	.83
Neutral images										
Time 1	.82	.78	.66	.74	.67	.66	.63	.68	.49	.52
Time 2	.82	.80	.60	.56	.63	.65	.56	.54	.50	.41

α = Cronbach's alpha. S-h = split-half reliability

.82 to .92. Reliability was highest for the positive images, followed by the sad, threat, and neutral images. Table 5 shows Cronbach's alpha for total dwell times for the face images at Times 1 and 2, which ranged from .59 to .94. Reliability was highest for happy faces, followed by sad, threat, and neutral faces. As can be seen in Tables 4 and 5, the split-half reliabilities were similar to the Cronbach's alpha values for naturalistic and face images.

The total number of fixations to the images was analyzed in the same manner and produced very similar results.

Table 5 Cronbach's alpha and split-half reliability for dwell times on threat faces, sad faces, happy faces, and neutral faces, and for each 2-s time interval during the first (Time 1) and second (Time 2) laboratory visits

	Total dwell time		0–2 s		2–4 s		4–6 s		6–8 s	
	α	S-h	α	S-h	α	S-h	α	S-h	α	S-h
Threat faces										
Time 1	.81	.83	−.07	−.19	.50	.43	.48	.53	.66	.66
Time 2	.80	.78	−.01	.19	.34	.37	.53	.47	.68	.54
Sad faces										
Time 1	.87	.84	−.15	.13	.54	.40	.50	.42	.69	.69
Time 2	.83	.79	.13	.28	.54	.48	.58	.57	.57	.61
Happy faces										
Time 1	.94	.95	.33	.46	.70	.62	.80	.83	.88	.88
Time 2	.94	.93	.26	.31	.73	.76	.84	.84	.89	.90
Neutral faces										
Time 1	.59	.62	−.11	−.14	.31	.27	.21	.36	.43	.33
Time 2	.72	.80	−.05	.06	.20	.28	.53	.53	.51	.62

α = Cronbach's alpha. S-h = split-half reliability

Cronbach's alphas for number of fixations on the naturalistic images at Time 1 were .86, .87, .89, and .82, for the threat, sad, positive, and neutral images, respectively. For the Time 2 data, Cronbach's alphas were .86, .87, .87, and .84 for the threat, sad, positive, and neutral images, respectively. For the face images at Time 1, Cronbach's alphas for number of fixations were .87, .88, .90, and .80 for the threat, sad, happy, and neutral faces, respectively. For the face images at Time 2, Cronbach's alphas were .83, .85, .92, and .79, for the threat, sad, happy, and neutral faces, respectively. The split-half reliabilities for number of fixations for the naturalistic images at Time 1 were .87, .83, .88, and .79 for the threat, sad, positive, and neutral images, respectively. For the Time 2 data, the split-half reliabilities were .87, .85, .86, and .82 for the threat, sad, positive, and neutral images, respectively. For the face images at Time 1, the split-half reliabilities for number of fixations were .89, .87, .93, and .81 for the threat, sad, happy, and neutral faces, respectively. For the face images at Time 2, the split-half reliabilities were .82, .79, .91, and .85 for the threat, sad, happy, and neutral faces, respectively.

To establish the reliability of the dwell time data across the 8-s presentation, each 8-s presentation was divided into 2-s intervals, and Cronbach's alpha and split-half reliability were calculated for the dwell times during each 2-s interval, for each image type. A 2-s interval was selected because several studies have used dwell times within 2-s intervals as the dependent variables when examining changes in attention over time (e.g., Arndt et al., 2014; Soltani et al., 2015). (Note that this analysis was not practical for the number-of-fixations data, because of the small number of fixations within each 2-s interval.)

Dividing the 8-s presentation into 2-s intervals revealed variation in the reliability of the dwell time indices over time. For sad images, the reliability estimates at Times 1 and 2 were low during the 0- to 2-s interval and generally increased over time (see Table 4). Positive images showed a similar pattern of results, although the reliability estimates were generally higher at each time interval for positive images than for the other image types. For neutral images, the reliability estimates at Times 1 and 2 were lowest during the 6- to 8-s interval. The reliability estimates for threat images were also generally stable across each 2-s interval at Times 1 and 2, except for the 0- to 2-s interval at Time 2, for which reliability was low.

Table 5 shows that reliability was generally low for dwell times for all face types in the 0- to 2-s interval, although it was relatively higher for the happy faces. The reliability estimates for all face types generally increased over time, although there were pronounced differences in the magnitudes of the estimates for the different face types. For example, for the 4- to 6-s and 6- to 8-s intervals, Cronbach's alpha ranged from .48 to .68 for threat faces, .50 to .69 for sad faces, .80 to .89 for happy faces, and .21 to .53 for neutral faces.

Test–retest reliability of the fixation data

Table 2 lists the Pearson correlations between the Time 1 and 2 total dwell times for the different image types (sad, threat, positive, and neutral) for the naturalistic and face images. For the naturalistic images, the test–retest correlations ranged from .61 to .80. For the face images, the test–retest correlations ranged from .36 to .57. These relationships are shown in Figs. 1 and 2. Table 3 lists the Pearson correlations between Times 1 and 2 for the numbers of fixations on the different image types. For the naturalistic images, the test–retest correlations ranged from .70 to .78, and for the face images they ranged from .39 to .65.

Discussion

In this study we evaluated the internal consistency and 6-month test–retest reliability of attentional bias indices for emotional face images and naturalistic images obtained from a free-viewing eye-tracking paradigm. The overall dwell time indices measured during the 8-s presentation generally showed good to excellent reliability according to commonly accepted criteria (i.e., Cronbach's alpha and split-half reliability > .80), with only a few exceptions. The split-half reliabilities for neutral images at Time 1 and threat and sad faces at Time 2 fell just below the .80 cutoff. The internal consistency estimates for total dwell times on neutral faces were lower, ranging from .59 to .80. Similar results were obtained in the analyses of the number of fixations, although Cronbach's alpha for the neutral face images was not substantially lower than it was for the other face images. Taken together, these results are consistent with prior investigations that have reported high internal consistency for total dwell times over extended viewing periods (4,000–6,000 ms; Lazarov et al., 2016; Skinner et al., 2018; Waechter et al., 2014).

As we noted, the total dwell time indices for neutral naturalistic images and neutral face images showed lower reliability relative to the other image types. Examination of the descriptive statistics for the total dwell time indices suggests that the lower reliability of attentional bias indices for neutral images appears to correspond to the lower variability in these indices relative to other image types. Recall that reliability is the proportion of true score variance relative to the total variance of a measure (Lord & Novick, 1968). Lower true score variance will, therefore, result in lower reliability, holding measurement error constant. The standard errors of the mean total dwell times for neutral naturalistic and face images were less than half of the standard errors of the mean total dwell times for the corresponding positive images (see Table 2). Given the neutral content of the images, this outcome makes sense. We would not expect to observe systematic individual differences in total dwell times for neutral images (i.e., biases toward neutral images). Instead, we would expect that most individuals would consistently spend less time viewing the neutral image in a display relative to the other images, and thus variability in viewing time on the neutral images would be limited. The lower reliability of the total dwell time indices for neutral images should not be a major concern for researchers, however, since hypotheses typically relate to attentional biases for specific types of emotional stimuli (e.g., sad images in depression, threatening images in anxiety).

Dividing the 8-s presentation duration into 2-s intervals revealed that the reliability of the dwell time indices for the different image types increased over time. This pattern of findings replicates that observed by Waechter et al. (2014). For the naturalistic image set, reliabilities for the attentional bias indices for each image type at each interval were generally similar across Times 1 and 2, with a few exceptions. For instance, the Cronbach's alpha and split-half reliability values were notably lower for positive and threat images for the 0- to 2-s interval at Time 2 than at Time 1. Examination of the descriptive statistics for dwell times on positive and threat images during the 0- to 2-s interval revealed that the variance of these indices was lower at Time 2 than at Time 1, which likely accounts for their lower reliability.[2] A probable explanation is that participants' knowledge of the different image types acquired during their first lab visit reduced the variability in initial orienting biases at the second lab visit. This may be particularly problematic for studies with short test–retest periods, given that the test–retest period in the present study was relatively long (approximately 6 months).

The reliability of the attentional bias indices for all face types was low during the 0- to 2-s interval. For threat, sad, and neutral faces, Cronbach's alpha values were near zero, and the split-half reliability coefficients were low or negative. For happy faces, Cronbach's alpha and split-half reliability were slightly higher (.26 to .45), but still well below acceptable levels. Examination of the descriptive statistics for these variables again suggests that the low reliability was due to relatively low variance across the sample for these indices. That is, there was a lack of reliable individual differences in initial orienting biases for emotional faces across the sample. Skinner et al. (2018) also observed low reliability of early attentional bias indices, which was attributable to low between-participant variance rather than high measurement error. Taken together, these results caution against the use of attentional bias indices from early time intervals in eye-tracking tasks, due to their low reliability. Although we do not suggest that researchers necessarily ignore this type of data (i.e., eye-tracking data collected during the first 1–2 s of a presentation), researchers should take into consideration the low reliability of early attentional bias indices when interpreting their results. In addition, these data highlight

[2] These data are available upon request.

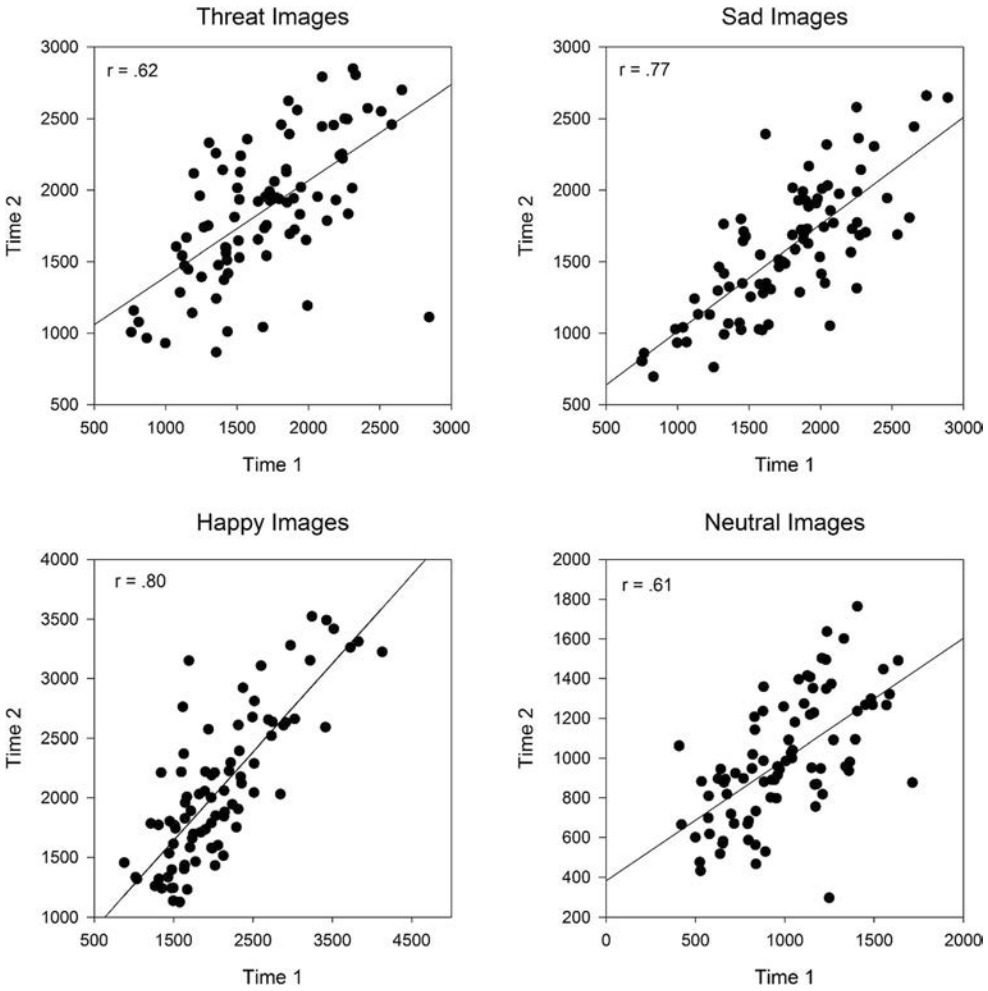

Fig. 1 Correlations for total dwell times for each type of naturalistic image at the first (Time 1) and second (Time 2) laboratory visits (i.e., test–retest correlations)

the utility of analyzing temporal patterns of attention in eye-tracking paradigms. Examining changes in attention over time, and placing greater emphasis on later time intervals at which attentional bias indices are most reliable, has the potential to advance research on the biases associated with various forms of psychopathology. We note that previous eye-tracking studies using displays ranging from two to 16 images have similarly observed low reliability of early attentional bias indices (e.g., Lazarov et al., 2016; Waechter et al., 2014); thus, this phenomenon does not appear to be tied to stimulus competition. However, it would be interesting to test whether early attentional bias indices would show higher reliability in more dynamic and/or naturalistic viewing conditions, such as video or virtual reality displays. Future research should explore modifications to eye-tracking paradigms and displays in order to improve the reliability of early attentional bias indices.

This is the first study to evaluate the stability of attentional bias indices from an eye-tracking task over an extended test–retest interval. Our results revealed differences in test–retest reliability across the two image sets. For naturalistic images, the total dwell times for each image type showed moderate to high stability from Time 1 to Time 2 (.61 to .80), as did the total numbers of fixations (.71 to .78). This level of stability for attentional bias indices is notable, especially considering the behavioral nature of the data and the relatively long test–retest interval. These results provide evidence for the trait-like nature of attentional biases to emotional images. For face images, the fixation indices for each image type showed low to moderate stability from Time 1 to Time 2 for both total dwell times (.36 to .57) and total numbers of fixations (.39 to .65). The lower test–retest reliability for the face images than for the naturalistic images may have been due to the nature of the images. The face image set was presumably less visually engaging to participants than was the naturalistic image set. Thus, participant familiarity or boredom with the face images at retest may have altered the patterns of attention, resulting in reduced stability. Alternatively, attentional biases for face images may be more susceptible to state influences than are attentional biases for naturalistic images, although it is unclear what those state influences might be. Future investigations of the test–retest reliability of attentional bias indices from eye-

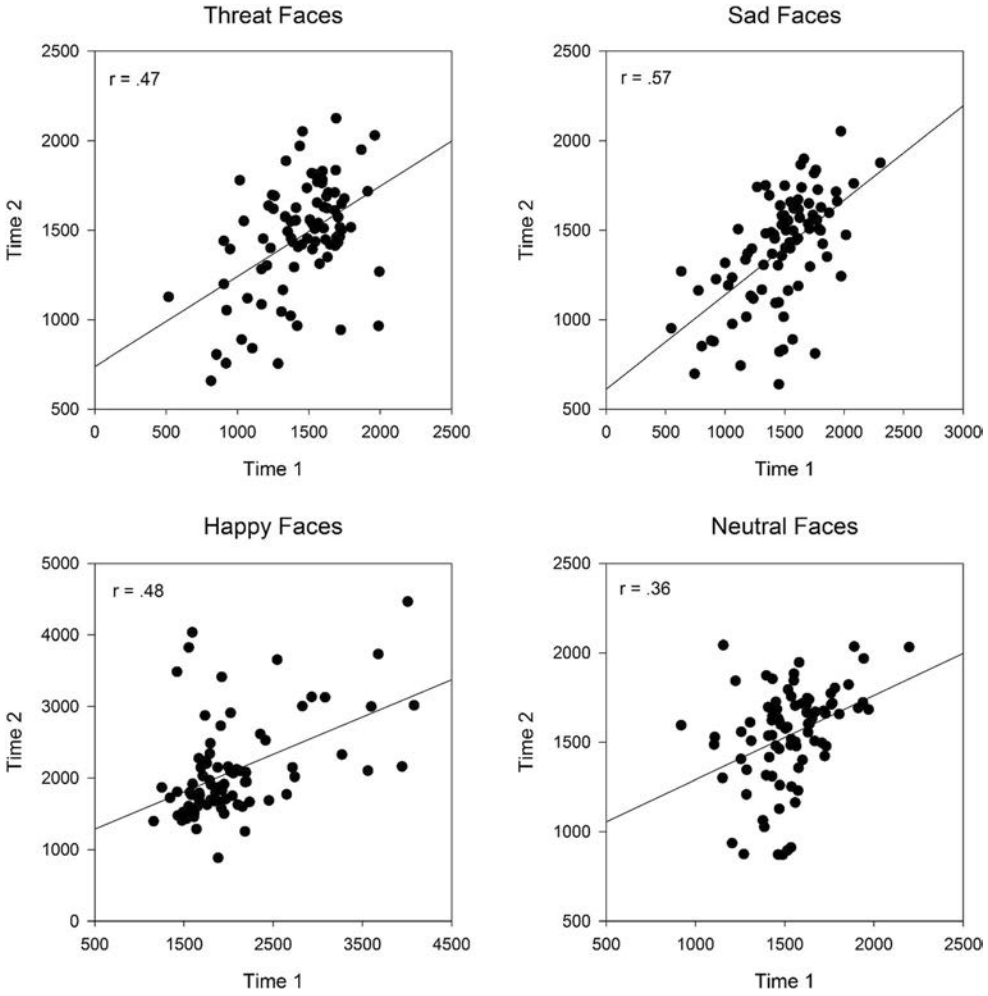

Fig. 2 Correlations for total dwell times for each type of face image at the first (Time 1) and second (Time 2) laboratory visits (i.e., test–retest correlations)

tracking paradigms will be necessary in order to clarify the different results for naturalistic and face images.

Strengths, limitations, and future directions

The present study has several strengths. The evaluation of two forms of reliability, internal consistency and test–retest reliability, allowed for the evaluation of both the consistency and stability of attentional bias indices from free-viewing eye-tracking paradigms. Furthermore, we examined the reliability of two commonly reported attentional bias indices (dwell time and number of fixations), for both naturalistic and face images. This design feature revealed both similarities and differences in the results for the two image sets that might inform stimulus choice in future research. Finally, the differences in our sample and methods as compared to those of prior studies (Lazarov et al., 2016; Skinner et al., 2018; Waechter et al., 2014) enhance the generalizability of our collective findings. Specifically, we used a mixed sample of community and student participants, selected to have minimal depression symptoms throughout the duration of the study, as well as a free-viewing paradigm that involved 8-s displays of four image types (threat, sad, positive, and neutral). Despite these methodological variations, our results were largely consistent with the published data, which increases confidence in the reliability of attentional bias indices based on eye-tracking paradigms.

The limitations of the present study point to several directions for future research. It would have been valuable to include a clinical sample (e.g., currently depressed participants), in addition to a nondepressed sample, to test for differences in reliability between the samples. Of relevance, Waechter et al. (2014) reported that they did not find systematic differences in the reliability of attentional bias indices between high and low socially anxious participants. Second, our use of a single test–retest period limits our conclusions about the test–retest reliability of attentional bias indices to this retest period (approximately 6 months). Future studies should use test–retest periods of varying durations in order to clarify the stability of attentional bias indices over time. Third, our decision to

examine the internal consistency of attentional bias indices in 2-s intervals was based on the use of 2-s intervals in previous research on the time course of attentional biases (e.g., Arndt et al., 2014; Soltani et al., 2015). The reliability of attentional bias indices over time, however, may depend on the time interval used in the analysis. Because reliability pertains to the specific way a variable is measured or defined, we recommend that researchers test and report the reliability of the attentional bias indices they use, which may include dwell times over different time intervals (e.g., 500 ms, 1,000 ms). Finally, it should also be noted that the generalizability of the present results is limited by the use of a sample consisting only of women.

Conclusions

The present study provides support for the reliability of attentional bias indices from a free-viewing eye-tracking paradigm. The total dwell times for threat, sad, and positive naturalistic and face images showed good to excellent reliability, as did the total numbers of fixations. When the 8-s display was divided into shorter intervals, the dwell times during the 0- to 2-s interval had lower reliability, especially for the face images. These results suggest that attentional bias indices obtained during early time intervals in eye-tracking tasks should be interpreted with caution. Our evaluation of test–retest reliability indicated adequate to good stability of attentional bias indices for the naturalistic image set. Although there are not clear standards for test–retest reliability, especially for behavioral paradigms, the level of stability of attentional bias indices that we observed for the naturalistic images is compelling, given that the test and retest administrations were approximately 6 months apart. These results thus provide support for the trait-like conceptualization of attentional biases. Test–retest reliability estimates for the face images were lower, and further research will be required in order to determine the cause of this discrepancy. We echo the call of other researchers (Skinner et al., 2018; Waechter et al., 2014) for additional investigations of the reliability of attentional bias indices, so as to improve understanding of the conditions and methodological variations that may influence the reliability of these indices. Nevertheless, converging evidence suggests that eye tracking is a reliable method for measuring attentional biases, supporting its use as a paradigm for studying selective attention to emotional stimuli in clinical and nonclinical populations.

Author note This research was supported by a grant from the Natural Sciences and Engineering Research Council (RGPIN 203664-2013) to C.S. and by a grant from the Canadian Institutes of Health Research (MOP-136988-2014) to K.D. and C.S. We thank two anonymous reviewers for their excellent feedback and suggestions.

References

Allen, M. J., & Yen, W M. (1979). *Introduction to measurement theory*. Monterrey, CA: Brooks/Cole.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed. [DSM-5]). Washington, DC: American Psychiatric Association.

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review*, *38*, 704–723. https://doi.org/10.1016/j.cpr.2012.09.004

Arndt, J. E., Newman, K. R., & Sears, C. R. (2014). An eye tracking study of the time course of attention to positive and negative images in dysphoric and nondysphoric individuals. *Journal of Experimental Psychopathology*, *5*, 399–413. https://doi.org/10.5127/jep.035813

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

Beck, A. T., & Clark, D. A. (1997). An information processing model of anxiety: Automatic and strategic processes. *Behaviour Research and Therapy*, *35*, 49–58. https://doi.org/10.1016/S0005-7967(96)00069-1

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

Bradley, B. P., Mogg, K., White, J., Groom, C., & De Bono, J. (1999). Attentional bias for emotional faces in generalized anxiety disorder. *British Journal of Clinical Psychology*, *38*, 267–278. https://doi.org/10.1348/014466599162845

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart & Winston.

Cronbach, L. J. (1946). A case study of the split-half reliability coefficient. *Journal of Educational Psychology*, *37*, 473–480.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Dozois, D. J., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, *10*, 83–89. https://doi.org/10.1037/1040-3590.10.2.83

Duque, A., & Vázquez, C. (2015). Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 107–114. https://doi.org/10.1016/j.jbtep.2014.09.005

Eide, P., Kemp, A., Silberstein, R., Nathan, P., & Stough, C. (2002). Test–retest reliability of the emotional Stroop task: Examining the paradox of measurement change. *Journal of Psychology*, *136*, 514–520. https://doi.org/10.1080/00223980209605547

Eizenman, M., Yu, L. H., Grupp, L., Eizenman, E., Ellenbogen, M., Gemar, M., & Levitan, R. D. (2003). A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry Research*, *118*, 117–128. https://doi.org/10.1016/S0165-1781(03)00068-4

Epp, A. M., Dobson, K. S., Dozois, D. J. A., & Frewen, P. A. (2012). A systematic meta-analysis of the Stroop task in depression, *32*, 316–328. https://doi.org/10.1016/j.cpr.2012.02.005

Frayn, M., Sears, C. R., & von Ranson, K. M. (2016). A sad mood increases attention to unhealthy food images in women with food addiction. *Appetite*, *100*, 55–63. https://doi.org/10.1016/j.appet.2016.02.008

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology*, *6*, 285–312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Ingram, R. E., Steidtmann, D. K., & Bistricky, S. L. (2008). Information processing: Attention and memory. In K. S. Dobson, & D. J. A. Dozois (Eds.), *Risk factors in depression* (pp. 145–169). San Diego, CA: Elsevier. https://doi.org/10.1016/B978-0-08-045078-0.00007-1

Joormann, J., & Gotlib, I. H. (2007). Selective attention to emotional faces following recovery from depression. *Journal of Abnormal Psychology*, *116*, 80–85. 2007-01891-007

Kelley, T. L. (1927). *Interpretation of educational measurements*. Oxford, UK: World Book Co.

Kellough, J. L., Beevers, C. G., Ellis, A. J., & Wells, T. T. (2008). Time course of selective attention in clinically depressed young adults: An eye tracking study. *Behaviour Research and Therapy*, *46*, 1238–1243. https://doi.org/10.1016/j.brat.2008.07.004

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). *International affective picture system (IAPS): Technical manual and affective ratings* (Technical Report No. A-1). Gainesville, FL: University of Florida, NIMH Center for the Study of Emotion and Attention.

Lazarov, A., Abend, R., & Bar-Haim, Y. (2016). Social anxiety is related to increased dwell time on socially threatening faces. *Journal of Affective Disorders*, *193*, 282–288. https://doi.org/10.1016/j.jad.2016.01.007

Leyman, L., De Raedt, R., Vaeyens, R., & Philippaerts, R. M. (2011). Attention for emotional facial expressions in dysphoria: An eye-movement registration study. *Cognition and Emotion*, *25*, 111–120. https://doi.org/10.1080/02699931003593827

Lord, F. M., & Novick, M. R. (1968), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mathews, A., & MacLeod, C. (1986). Discrimination of threat cues without awareness in anxiety states. *Journal of Abnormal Psychology*, *95*(2), 131–138.

McCabe, S. B., Gotlib, I. H., & Martin, R. A. (2000). Cognitive vulnerability for depression: Deployment of attention as a function of history of depression and current mood state. *Cognitive Therapy and Research*, *24*, 427–444. https://doi.org/10.1023/A:1005579719849

Nelson, A., Quigley, L., Carriere, J., Purdon, C., & Smilek, D. (2010). *The role of state and trait-anxiety on the time-course of selective attention towards emotional images*. Poster presented at the 6th World Congress of Behavioral and Cognitive Therapies, Boston, MA.

Nelson, A. L., Purdon, C., Quigley, L., Carriere, J., & Smilek, D. (2015). Distinguishing the roles of trait and state anxiety on the nature of anxiety-related attentional biases to threat using a free viewing eye movement paradigm. *Cognition and Emotion*, *29*, 504–526. https://doi.org/10.1080/02699931.2014.922460

Newman, K. R., & Sears, C. R. (2015). Eye gaze tracking reveals different effects of a sad mood induction on the attention of previously depressed and never depressed women. *Cognitive Therapy and Research*, *39*, 292–306. https://doi.org/10.1007/s10608-014-9669-x

Peckham, A. D., McHugh, R. K., & Otto, M. W. (2010). A meta-analysis of the magnitude of biased attention in depression. *Depression and Anxiety*, *27*, 1135–1142. https://doi.org/10.1002/da.20755

Quigley, L., Nelson, A. L., Carriere, J., Smilek, D., & Purdon, C. (2012). The effects of trait and state anxiety on attention to emotional images: An eye-tracking study. *Cognition and Emotion*, *26*, 1390–1411. https://doi.org/10.1080/02699931.2012.662892

Rinck, M., Becker, E. S., Kellermann, J., & Roth, W. T. (2003). Selective attention in anxiety: Distraction and enhancement in visual search. *Depression and Anxiety*, *18*, 18–28. https://doi.org/10.1002/da.10105

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

Schofield, C. A., Johnson, A. L., Inhoff, A. W., & Coles, M. E. (2012). Social anxiety and difficulty disengaging threat: Evidence from eye-tracking. *Cognition and Emotion*, *26*, 300–311. https://doi.org/10.1080/02699931.2011.602050

Sears, C. R., Newman, K. R., Ference, J. D., & Thomas, C. L. (2011). Attention to emotional images in previously depressed individuals: An eye-tracking study. *Cognitive Therapy and Research*, *35*, 517–528. https://doi.org/10.1007/10608-011-9396-5

Sears, C. R., Thomas, C. L., LeHuquet, J. M., & Johnson, J. C. (2010). Attentional biases in dysphoria: An eye-tracking study of the allocation and disengagement of attention. *Cognition and Emotion*, *24*, 1349–1368. https://doi.org/10.1080/02699930903399319

Skinner, I. W., Hübscher, M., Moseley, G. L., Lee, H., Wand, B. M., Traeger, A. C., ... McAuley, J. H. (2018). The reliability of eyetracking to assess attentional bias to threatening words in healthy individuals. *Behavior Research Methods, 50,* 1778–1792. https://doi.org/10.3758/s13428-017-0946-y

Soltani, S., Newman, K., Quigley, L., Fernandez, A., Dobson, K., & Sears, C. R. (2015). Temporal changes in attention to sad and happy faces distinguish currently and remitted depressed individuals from never depressed individuals. *Psychiatry Research*, *230*, 454–463. https://doi.org/10.1016/j.psychres.2015.09.036

Speirs, C., Belchev, Z., Fernandez, A., Korol, S., & Sears, C. (2018). Are there age differences in attention to emotional images following a sad mood induction? Evidence from a free-viewing eye-tracking paradigm. *Aging, Neuropsychology, and Cognition*, *25*, 928–957. https://doi.org/10.1080/13825585.2017.1391168

Spitzer, R. L., Kroenke, K., Williams, J. B., & the Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA*, *282*, 1737–1744. https://doi.org/10.1001/jama.282.18.1737

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

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

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... Nelson C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *68*, 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

Waechter, S., Nelson, A. L., Wright, C., Hyatt, A., & Oakman, J. (2014). Measuring attentional bias to threat: Reliability of dot probe and eye movement indices. *Cognitive Therapy and Research*, *38*, 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*, 441–458. https://doi.org/10.1007/s10608-015-9670-z

Wenzlaff, R. M., Rude, S. S., Taylor, C. J., Stultz, C. H., & Sweatt, R. A. (2001). Beneath the veil of thought suppression: Attentional bias and depression risk. *Cognition and Emotion*, *15*, 435–452. https://doi.org/10.1080/02699930125871

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, *120*, 3–24. https://doi.org/10.1037/0033-2909.120.1.3

Wright, R. D., & Ward, L. M. (2008). *Orienting of attention*. Oxford, UK: Oxford University Press.

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