Behav Res (2018) 50:1778–1792
DOI 10.3758/s13428-017-0946-y

BRIEF COMMUNICATION

The reliability of eyetracking to assess attentional bias to threatening words in healthy individuals

Ian W. Skinner[1,2] · Markus Hübscher[1,2] · G. Lorimer Moseley[1,3] · Hopin Lee[1,2,6,7] · Benedict M. Wand[4] · Adrian C. Traeger[1,2,8] · Sylvia M. Gustin[1,5] · James H. McAuley[1,2]

Published online: 15 August 2017

Abstract Eyetracking is commonly used to investigate attentional bias. Although some studies have investigated the internal consistency of eyetracking, data are scarce on the test–retest reliability and agreement of eyetracking to investigate attentional bias. This study reports the test–retest reliability, measurement error, and internal consistency of 12 commonly used outcome measures thought to reflect the different components of attentional bias: overall attention, early attention, and late attention. Healthy participants completed a preferential-looking eyetracking task that involved the presentation of threatening (sensory words, general threat words, and affective words) and nonthreatening words. We used intraclass correlation coefficients (ICCs) to measure test–retest reliability (ICC > .70 indicates adequate reliability). The ICCs(2, 1) ranged from –.31 to .71. Reliability varied according to the outcome measure and threat word category. Sensory words had a lower mean ICC (.08) than either affective words (.32) or general threat words (.29). A longer exposure time was associated with higher test–retest reliability. All of the outcome measures, except second-run dwell time, demonstrated low measurement error (<6%). Most of the outcome measures reported high internal consistency (α > .93). Recommendations are discussed for improving the reliability of eyetracking tasks in future research.

Keywords Eyetracking · Reliability · Attentional bias · Preferential looking · Threat

Electronic supplementary material The online version of this article (doi:10.3758/s13428-017-0946-y) contains a slide of the figure for download, which is available to authorised users.

✉ James H. McAuley
j.mcauley@neura.edu.au

Ian W. Skinner
i.skinner@neura.edu.au

[1] Neuroscience Research Australia, Margarete Ainsworth Building, Barker Street, Randwick 2031, New South Wales, Australia

[2] Prince of Wales Clinical School, University of New South Wales, Sydney, Australia

[3] Sansom Institute for Health Research, University of South Australia, Adelaide, Australia

[4] School of Physiotherapy, University of Notre Dame Australia, Fremantle, Australia

[5] School of Psychology, University of New South Wales, Sydney, Australia

[6] School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

[7] Centre for Rehabilitation Research, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[8] Sydney Medical School, University of Sydney, Sydney, Australia

Eyetracking is increasingly being used to investigate attentional bias (Armstrong & Olatunji, 2012; Liossi, Schoth, Godwin, & Liversedge, 2014; Mogg, Bradley, Field, & De Houwer, 2003). Compared to traditional reaction time based methods of attentional bias, such as the dot-probe task, eyetracking is proposed to provide a more direct, and therefore superior, measure of sensory processing (Armstrong & Olatunji, 2012; Toh, Rossell, & Castle, 2011). However, because few reports have been published on the reliability of eyetracking, it is unknown whether study results are valid. This is because the reliability of a measure can influence Type II error rates, effect sizes, and confidence intervals (Kopriva & Shaw, 1991; Loken & Gelman, 2017; Meyer, 2010).

Attentional bias

Attentional bias describes the preferential allocation of cognitive resources to the detection of salient stimuli (Crombez, Van Ryckeghem, Eccleston, & Van Damme, 2013). Attentional bias to threat stimuli has been identified in the development and maintenance of clinical conditions such as addiction, anxiety, depression and chronic pain (Sharpe, Haggman, Nicholas, Dear, & Refshauge, 2014; White, Suway, Pine, Bar-Haim, & Fox, 2011). Recently, attentional-bias modification training has been found to reduce symptoms of affective and pain disorders (Amir, Beard, Burns, & Bomyea, 2009; Amir, Weber, Beard, Bomyea, & Taylor, 2008; Sharpe et al., 2012).

Models of attentional bias, such as the "vigilance–avoidance" model (Mogg, Bradley, Miles, & Dixon, 2004) and "threat interpretation" model (Todd et al., 2015), consider attentional bias to be dynamic; attentional bias may shift toward or away from a stimulus during the stimulus exposure. For example, the vigilance-avoidance model posits that individuals may attend to a threat stimulus during initial exposure (vigilance) but then after detection, avoid the threat stimulus (avoidance) (Mogg et al., 2004). These models incorporate a temporal component of processing, broadly categorised into overall, early and late processing. To investigate such models, methods that can consistently distinguish between the temporal components of attentional processing are needed.

Eyetracking

Eyetracking continuously measures eye movements to stimuli presented on either a computer screen or mobile head centered video device. Prespecified spatial (e.g., displacement) and temporal (e.g., velocity and acceleration) eye movement parameters are used to derive "fixations" and "saccades." Fixation-based measures can be categorized according to the component of attention they are proposed to measure: overall, early, or late. Overall attention combines early and late stage processing and reflects the viewing pattern across the total stimulus duration. For example, if stimuli are presented for 4,000 ms, the total dwell time toward the salient stimulus is considered an indicator of overall attention. Early attention reflects the initial viewing pattern when stimuli are first presented and has been used to indicate initial vigilance, which may be important in threat detection (Armstrong & Olatunji, 2012). Examples include location of the first fixation, first fixation latency, and duration of the first fixation to a salient stimulus. Late attention reflects the viewing pattern that occurs after the initial viewing pattern and is thought to reflect rumination or maintenance, which are important in theories of depression (Donaldson, Lam, & Mathews, 2007). Examples of late attention outcome measures are second- or last-run dwell times and the dwell time for the second half of the stimulus duration.

Eyetracking has been used to investigate attentional bias in clinical conditions such as depression (Armstrong & Olatunji, 2012), anxiety (Armstrong & Olatunji, 2012), addictive disorders (Mogg et al., 2003), obesity (Gao et al., 2011), post-traumatic stress disorder (Felmingham, Rennie, Manor, & Bryant, 2011), and pain (Fashler & Katz, 2014; Yang, Jackson, & Chen, 2013). Distinguishing the temporal components allows researchers to more clearly define the role of attentional bias in these clinical conditions. For example, Duque (2015) found that participants with major depressive disorder had an attentional bias to sad faces for maintenance indices (late processing) such as total fixation time, but not for orientating (early processing) attention indices.

There is considerable variability in the tasks and procedural variables used in eyetracking research (Radach & Kennedy, 2004). For example, different tasks (e.g., preferential-looking task, visual-search tasks, dot-probe tasks), outcome measures (e.g., first-fixation latency, percentage of initial fixations, average visit duration), and stimuli (e.g., words, images, faces) are common (Fashler & Katz, 2014; Felmingham et al., 2011; Yang, Jackson, Gao, & Chen, 2012). However the reliability of attentional-bias tasks has been questioned (Rodebaugh et al., 2016), and good quality information on the reliability of procedural variables will help inform which tasks, outcome measures, and stimuli to use in future research studies.

Reliability

It is important that paradigms and procedural variables, such as those used to investigate attentional bias, produce measurements that are reliable. Poor reliability has statistical and conceptual implications. It has been demonstrated, for example, that effect sizes can vary depending on the sample size (Loken & Gelman, 2017), and statistical power is reduced as the reliability of a task decreases (Meyer, 2010). Conceptually, it is difficult to reproduce study findings if tasks and procedural variables are not reliable. Conclusions from experiments with poor reliability are therefore questionable (Loken & Gelman, 2017).

Because there is some variation in descriptions of reliability, we used the taxonomy described by Mokkink et al. (2010). Reliability comprises three measurement properties: test–retest reliability, measurement error and internal consistency (Mokkink et al., 2010). These three measurement properties reflect conceptually different aspects of reliability and should all be considered when investigating reliability (Mokkink et al., 2010; Scholtes, Terwee, & Poolman, 2011). A minimum of two testing sessions is required to assess test–retest reliability and agreement, whereas internal consistency can be evaluated using data from a single testing session.

Test–retest reliability indicates how well a task can distinguish between participants with reference to the consistency between measurements (de Vet, Terwee, Knol, & Bouter, 2006). Both the

consistency of results between measurements and the participant variance are used to calculate test–retest reliability—that is, did participants all score the same, or was there adequate variability in the results to distinguish participants from each other? The preferred method for assessing test–retest reliability is the intraclass correlation coefficient (ICC). ICCs vary between 0 and 1, although theoretically it is possible to report values lower than 0. Higher numbers reflect stronger evidence of test–retest reliability (Weir, 2005).

Measurement error reflects the consistency of results between measurements—that is, how similar are the results between testing sessions (de Vet et al., 2006)? Unlike test–retest reliability, the variance between participants is not considered when calculating measurement error (Kottner & Streiner, 2011). Measurement error is reported in the same unit as the task. Low measurement error is preferred.

Low measurement error (i.e., consistent results between testing sessions) and poor test–retest reliability (i.e., inability to distinguish participants) arise when there is too little variance between participants (their scores are too similar). For example, if newborn human babies were weighed twice on the same day using scales designed for newborn elephants all the human babies would have consistent scores between measurements (low measurement error), however the scores would not be able to distinguish between the human babies due to the low variance in scores (poor test–retest reliability). Test–retest reliability is therefore considered more beneficial for discriminative testing—that is, when aiming to differentiate participants on the basis of a set of scores from a certain task, as in cross-sectional studies. Measurement error is preferred for evaluative testing, when testing participants over time and measuring within-subjects change. Variance in participant scores is regarded as being less important for evaluative testing. When researchers investigate attentional-bias tasks, procedural variables are required that will be able to produce scores that can both accurately discriminate between participants (discriminative testing indicated by test–retest reliability) and accurately measure change over time for individual participants (evaluative testing indicated by agreement) (Guyatt, Walter, & Norman, 1987).

Internal consistency indicates how the subjects respond to individual items on a task—that is, the homogeneity of the items on a scale (interrelatedness) (Streiner, 2003c). For example, when investigating attentional bias to threat-related anxiety, participants would be expected to view each threat-related word in a similar manner. A high level of internal consistency provides confidence that the interpretation of the composite score is an accurate measure of the underlying construct being investigated. Cronbach's alpha is the preferred method for analyzing internal consistency, as it considers the mean of all possible splits. Split-half reliability tends to underestimate reliability because it splits a scale in half, so depending on how a scale is split, a different reliability may be returned (Streiner, 2003c).

Eyetracking tasks that measure attentional bias should be able to discriminate between people (high test–test reliability) and should have consistent scores on repeated testing (low measurement error) and high interrelatedness of the items (high internal consistency) (Kottner & Streiner, 2011; Streiner, 2003b).

Previous research

Waechter, Nelson, Wright, Hyatt, and Oakman (2014) examined the internal consistency of eyetracking within a dot-probe paradigm in university students with high and low social anxiety. Anger, disgust, and happy facial images were paired with calm or neutral facial images using a 5,000-ms exposure time. The reliability coefficients for early attention were low (e.g., proportion of first fixations to angry faces: $\alpha = -2.18$). Conversely, the eye movement indices using the full stimulus exposure (overall attention) had excellent reliability (e.g., proportion of viewing time to angry faces 0–5,000 ms: $\alpha = .94$). Waechter et al. concluded that more research was needed to establish reliable methods to assess attentional bias.

Price et al. (2015) reported the test–retest reliability of eyetracking within a dot-probe paradigm using fearful and neutral facial images and a 2,000-ms exposure time. Single and average ICC measures were reported for healthy children (aged 9–13 years) at five time points over a 14-week period. The ICC scores for all trials varied between –.03 to .55, depending on the data-filtering process (e.g., excluding reaction times <300 ms and >2,500 ms and ±3 *SD*s from individual's session means) and the reliability statistic used to interpret the results (i.e., single or average ICC). Importantly, none of the ICCs were above a standard threshold for acceptability (i.e., ICC > .70).

Lazarov, Abend, and Bar-Haim (2016) tested the internal consistency and test–retest reliability of eyetracking within a free-viewing task using 16 simultaneously presented facial images for 6,000 ms (half the faces displayed disgusted expressions, and half were neutral). The participants were 20 university students with high social anxiety and 20 university students with low social anxiety. Measures of early attentional bias (latency to first fixation, first-fixation location, first-fixation dwell time) and overall attention (total dwell time) were reported. Cronbach's alpha scores, representing internal consistency, for overall measures of attentional bias were high, ranging from .89 to .95. One week test–retest reliability for overall attentional bias using Pearson's correlation coefficients ranged from .62 to .68. Test–retest reliability was lower for early measures of attentional bias, ranging from .06 to .08, than for the measures of overall attention.

The data from these studies suggest that measures of early attention or measures that use less of the available stimulus presentation time may have lower internal consistency and

poorer test–retest reliability than measures that use more of the available stimulus duration.

There are no published data on the reliability of using words as stimuli in attentional-bias research using eyetracking. This is important as a systematic review found that words are the most common stimuli in attentional-bias tasks (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van IJzendoorn, 2007).

There is also a lack of published data on the agreement of eyetracking when it is used to investigate attentional bias. The evidence from one measurement property of reliability does not provide evidence for another measurement property (Guyatt et al., 1987). For example, test–retest reliability and internal consistency are not suitable measures of reliability for evaluative studies—that is, those comparing within-subjects measures over time. Instead, agreement is the measurement property that indicates whether a measurement tool is appropriate for determining longitudinal changes (de Vet et al., 2006; Guyatt et al., 1987). An understanding of all three measurement properties of reliability can allow researchers to decide for what purpose, between-subjects (discriminative) or within-subjects (evaluative) testing, a tool is appropriate.

Healthy participants are commonly used as a comparison group in attentional-bias studies. No studies have reported reliability data on healthy adult participants using eyetracking with words as stimuli. Reliability is known to be specific to the population in which it is being tested; therefore, it is possible that the measurement properties of eyetracking may vary between clinical and healthy participants (Lakes, 2013). As compared with healthy control participants, greater variation is often found in data obtained from clinical populations (Bartko, 1991). If measurement error is stable, then increased between-subjects variance will increase test–retest reliability, whereas decreased between-subjects variance may decrease test–retest reliability. For example, Farzin, Scaggs, Hervey, Berry-Kravis, and Hessl (2011) investigated the reliability of gaze aversion to different facial features, in participants with Fragile X syndrome (FXS) and healthy controls. The test–retest reliability for the proportion of time spent looking at the mouth was higher (ICC = .97) in the FXS cohort than in the healthy controls (ICC = .63). Farzin et al. noted that the reduced between-subjects variance in the healthy controls may explain the lower ICC values (test–retest reliability). Accurately investigating between-group differences requires adequate reliability in both clinical populations and healthy controls.

Present study

The primary aim of this study was to assess the reliability of eyetracking when it is used to investigate attentional bias to threat-related words in healthy participants. Reliability will be assessed using test–retest reliability, measured with ICC(2, 1); measurement error, measured with the standard error of measurement; and internal consistency, measured with Cronbach's alpha.

Method

Study design

We used an observational test–retest design. Healthy participants completed identical preferential-looking tasks on two occasions. A methodological protocol for the study was published to the completion-of-data collection (Open Science Framework Project MT3K8). Deviations from the protocol are noted in this article. Ethics approval was obtained from the University of New South Wales Human Research Ethics Committee (HC14240).

Participants

Healthy adult participants were recruited from the Sydney metropolitan area. Participants were included in the study if they were 18–75 years old, had good level of English proficiency, and had normal or corrected-to-normal vision.

English proficiency was assessed using three questions from the Language Experience and Proficiency Questionnaire. Participants were asked to select, on a scale from 0 to 10, their levels of proficiency in speaking, understanding, and reading English. A minimum score of 7, which is regarded as "good," was required for participants to be included (Marian, Blumenfeld, & Kaushanskaya, 2007). We excluded participants with poor English proficiency because fixations might be unrelated to the threat value of the word. Global measures of self-report proficiency are good indicators of actual performance on specific measures of language ability (Marian et al., 2007).

Participants were also excluded if they were currently reporting pain in any body region, reported a previous pain condition that lasted more than 6 months, or reported pain in any body region that had lasted more than 72 h at any time during the past 3 months.

Materials

Apparatus

An EyeLink 1000 eyetracker (Version 4.56; SR Research, Ontario, Canada) with remote camera upgrade, desktop mount, 16-mm lens, and target sticker was used to record monocular eye movements from the right eye at 500 Hz. Stimuli were displayed on a HP Compaq LA2205 wide LCD monitor with a 1,680 × 1,050 resolution, 32 bits per

pixel, and a refresh rate of 60 Hz. The preferential-looking task was programmed with Experiment Builder (Version 1.10.1241; SR Research, Ontario, Canada). A 5-point calibration procedure was used and accepted when the average calibration error was less than 1° of visual angle. We used a 5-point calibration, instead of the default 9-point calibration, as the stimuli did not extend to the corners of the display. This is in keeping with previous other eyetracking studies that use remote eyetrackers with no fixed headmount (Lazarov et al., 2016). All stimuli were presented in white on a black background.

Procedure

Testing took place at Neuroscience Research Australia. Each participant attended one 90-min session. Upon completion of the task, participants completed a demographic questionnaire along with the short-form version of the Depression Anxiety and Stress Scales (Lovibond & Lovibond, 1995) and the Pain Catastrophising Scale (Sullivan, Bishop, & Pivik, 1995). Participants were given 20 min to complete the questionnaires, followed by a compulsory 10-min washout period, during which time they were seated quietly, before the task was conducted a second time (retest).

Preferential-looking task

In preferential-looking tasks, two competing stimuli are displayed and participants are instructed to view stimuli as they wish. We used a preferential-looking task instead of the more traditional dot-probe task because previous research had suggested that the dot-probe task is not reliable (Rodebaugh et al., 2016; Schmukle, 2005). The test–retest reliability and agreement for words is unknown when using eyetracking.

The preferential-looking task consisted of eight practice trials and 48 active trials. Each trial consisted of three sequentially presented still screens. The first screen displayed a fixation cross (font: Times New Roman normal; size: 90; location: x = 840, y = 525 [center of screen]). Participants were instructed to fix their gaze on the middle of the cross. A researcher sitting in an adjacent room monitored the participants' gaze. After a stable fixation had been made on the cross for 2,000 ms, the researcher manually progressed the trial to the next screen. The researcher used a timer on the display screen that was automatically reset at the start of each trial. The second screen displayed two words (the stimuli), presented on the left and right sides of the screen for 4,000 ms (Tahoma normal font, size 30). One of the words was a "threat word" and the other a "neutral (control) word." Participants were instructed to read both words and to keep reading them until the words had disappeared. The third screen, a blank screen, was displayed automatically for 1,000 ms. Prior to each trial, a drift check was performed. If the calibration error was more than 1° of visual angle, a new calibration was performed.

To avoid participant fatigue, trials were arranged into three equal blocks of 16 trials. After each block, participants were given a self-timed break of 30 s or longer. The threat words in each block of trials came from one of three threat categories (1) "sensory pain," (2) "affective pain," or (3) "general threat." Each block only contained words from one threat category. The eight words from each threat category (target) were paired with a "neutral (control) word" matched for length and frequency in everyday language, using an English control word search engine (Table 1; Guasch, Boada, Ferré, & Sánchez-Casas, 2013). Word pairs were presented twice within each block, with each word being presented on both the left and the right. The word pairs were randomized within each block, and the same word pair was not presented during consecutive trials. The order of blocks were randomized.

Table 1 Threat (target) and matched neutral (control) words presented to participants

Sensory Pain		Affective Pain		General Threat	
Target	*Control*	*Target*	*Control*	*Target*	*Control*
sharp	minor	tiring	orient	crushing	footpath
ache	eats	unbearable	delicately	frightful	stonework
throbbing	visionary	punishing	polishing	terrifying	theatrical
cramping	allusive	exhausting	decisively	threat	sounds
burning	samples	annoying	marketed	scared	drives
dull	maps	troublesome	nutritional	danger	fields
shooting	entering	irritating	installing	harmful	drifted
pain	hill	nagging	planner	suffocating	interviewee

Threat word selection

The "sensory pain" words (Table 1) were selected from a study that had investigated the words that participants used to describe their back pain (Jensen, Johnson, Gertz, Galer, & Gammaitoni, 2013). The "affective pain" words (Table 1) were selected from a study that had investigated attentional bias in participants with acute low back pain (Sharpe et al., 2014). The general threat words (Table 1) had previously been used to investigate attentional bias to threat in chronic-pain patients (Dehghani, Sharpe, & Nicholas, 2003).

Statistical analysis

Outcome measures Twelve eyetracking outcome measures commonly used to assess attentional bias were calculated from the extracted data (Table 2; Kimble, Fleming, Bandy, Kim, & Zambetti, 2010; Liossi et al., 2014; Yang et al., 2013). These outcome measures, selected a priori, were chosen to reflect the different stages of attentional bias: overall attention, early attention, and late attention. Each outcome measure was calculated as a ratio of fixation time of the target word to the control word, and then converted to a percentage. A mean attentional-

Table 2 Outcome measures and associated equations used to assess the different stages of attentional bias

Stage of Attention	Outcome Measure and Equation
Overall attention	Total dwell time of threat words (0–4,000 ms): $\frac{\text{mean dwell time of target word 0–4,000ms}}{(\text{mean dwell time of target word 0–4,000ms}+\text{mean dwell time of control word 0–4,000ms})}$
	Total dwell time of threat words (0–500 ms): $\frac{\text{mean dwell time of target word 0–500ms}}{(\text{mean dwell time of target word 0–500ms}+\text{mean dwell time of control word 0–500ms})}$
	Total dwell time of threat words (0–2,000 ms): $\frac{\text{mean dwell time of target word 0–2,000ms}}{(\text{mean dwell time of target word 0–2,000ms}+\text{mean dwell time of control word 0–2,000ms})}$
	Total dwell time of threat words (0–3,000 ms): $\frac{\text{mean dwell time of target word 0–3,000ms}}{(\text{mean dwell time of target word 0–3,000ms}+\text{mean dwell time of control word 0–3,000ms})}$
Early attention	Probability of first fixation to target word: $\frac{\text{number of times the first fixation was to the target word}}{\text{total number of trials with a first fixation}} \times 100$
	Latency to first fixation of threat words: $\frac{\text{mean latency of first fixation to target word}}{(\text{mean latency of first fixation to target words}+\text{mean latency of first fixation to control word})}$
	First-run dwell time of threat words: $\frac{\text{mean first}-\text{run dwell time target word}}{(\text{mean first}-\text{run dwell time target word}+\text{mean first}-\text{run dwell time control word})}$
	First-fixation duration of threat words: $\frac{\text{mean first}-\text{fixation duration}}{\text{mean first}-\text{fixation duration target word}+\text{mean first}-\text{fixation duration control word}}$
Late attention	Second-run dwell time of threat words: $\frac{\text{mean second}-\text{run dwell time target word}}{(\text{mean second}-\text{run dwell time target word}+\text{mean second}-\text{run dwell time control word})}$
	Last-run dwell time of threat words: $\frac{\text{mean last}-\text{run dwell time target word}}{(\text{mean last}-\text{run dwell time target word}+\text{mean last}-\text{run dwell time control word})}$
	Total dwell time of threat words (500–4,000 ms): $\frac{\text{mean dwell time of target word 500–4,000ms}}{(\text{mean dwell time of target word 500–4,000ms}+\text{mean dwell time of control word 500–4,000ms})}$
	Total dwell time of threat words (1,000–4,000 ms): $\frac{\text{mean dwell time of target word 1,000–4,000ms}}{(\text{mean dwell time of target word 1,000–4,000ms}+\text{mean dwell time of control word 1,000–4,000ms})}$

bias score was calculated for each participant in each word category for the test and retest sessions.

Data reduction The raw gaze data were automatically parsed into sequences of saccades and fixations and loaded into the SR Research EyeLink Dataviewer (Version 2.3.22, Ontario, Canada). The standard cognitive configuration was used to define fixations (i.e., recording parse type: gaze saccade; velocity threshold: 30 ms; saccade acceleration threshold: 8,000 ms/s; saccade motion threshold: 0.1 ms/s^2). A 100-pixel area of interest, dependent on the word length, was set around each word (i.e., the area was set relative to the start and end of each word). No other filters were applied to the data—for example, no merge of fixations, no minimum fixation duration, and no blink correction. An interest period was created for each respective outcome measure and an interest area report was extracted. The subsequent data filtering and reliability analyses were completed in STATA (version 13.1; Stata Corp., Texas, USA).

We excluded trials during which the eyetracker lost and did not regain view of the eye (e.g., trials in which a blink occurred were still included if the eyetracker regained view of the eye after the blink), or when the participant did not adhere to the instructions (i.e., participants were instructed to look directly at the middle of the fixation cross until it disappeared and then to read both words and keep reading them until the words disappeared). Three criteria were used to exclude invalid trials:

1. A fixation was not made to both interest areas. No detection of a fixation to both interest areas implies the eyetracker may have lost view of the eye and not regained the view of the eye, or the participant did not read both words (Mogg et al., 2003). Since participants were instructed to read both words if a fixation was not captured on both words, this was considered an invalid trial.
2. The first-fixation latency to either interest area was less than 30 ms. Any fixations that occurred less than 30 ms after word presentation were likely not due to the content of the words.
3. Less than 3,000 ms (75%) of fixations were captured during the interest period (e.g., 0–4,000 ms) anywhere on the screen. That is, trials were still included if more than 75% of the fixations were captured at any location of the screen, not just within the interest areas. If less than 75% of fixations were captured during the interest period, the eyetracker might have lost tracking of the eye and not regained view of it, or the participant might have looked away from the screen after viewing both words (Fashler & Katz, 2014).

Participants were instructed to keep reading the words while they were being presented. After applying these criteria, if more than 25% of a participant's trials were excluded, then all of the participant's data were also excluded (Vervoort, Trost, Prkachin, & Mueller, 2013).

Reliability analysis An intraclass correlation coefficient (ICC) was calculated to assess test–retest reliability. ICCs are able to detect systematic differences between testing sessions and are preferred over other correlation coefficients such as Pearson's *r*, which in contrast does not consider systematic differences between testing sessions (Weir, 2005).

We used a two-way random-effects model with absolute agreement (ICC 2,1) (Shrout & Fleiss, 1979) as our primary outcome measure of test–retest reliability. A random-effects model is preferred because it considers systematic differences between testing sessions. The single measure was used, since this reflects how eyetracking is normally done in experimental research; that is, participants are normally tested on one occasion. A two-way random-effects model using an average measure (ICC 2,2) was also calculated. This average measure was included to indicate whether testing people twice and using the mean score is more reliable than using the results from one testing session (see the supplementary material, Table S1). As per our protocol, we also calculated a two-way fixed-effects model for consistency of agreement (ICC 3,1), to investigate the consistency of the scores (supplementary material, Table S1) (Streiner, 2003b). A two-way fixed-effects model does not consider systematic difference between testing sessions (de Vet et al., 2006).

The standard error of measurement (*SEM*) was calculated as an indicator of measurement error. We deviated from our protocol (Open Science Framework MT3K8) by using the variance scores, $SEM_{\text{agreement}} = \sqrt{\sigma^2_{\text{retest}} + \sigma^2_{\text{residual}}}$ (de Vet et al., 2006) instead of the standard deviation, $SEM = SD \times \left(\sqrt{1 - ICC_{2,1}}\right)$, to calculate the *SEM*. We did this because variance scores consider systematic differences between measurements (de Vet et al., 2006). With each outcome measure entered as the dependent variable, the participants and the test–retest sessions were considered random factors in a mixed model in order to estimate the variance for the participants (σ^2_{p}), the test–retest variance (σ^2_{retest}), and the residual variance ($\sigma^2_{\text{residual}}$). These variances are reported in the supplementary material (Table S2).

Internal consistency, reflecting "the interrelatedness of items on a test," was calculated using Cronbach's alpha for each set of words and each outcome measure, using

the scores from the first testing session (Cronbach, 1951; Streiner, 2003c).

Sample size We followed the recommendations from de Vet, Terwee, Mokkink, and Knol (2011) to calculate the required sample size. Using the simulated power calculations in Giraudeau and Mary (2001), we estimated that 50 participants would be required, using two repeated measurements, to detect an ICC of .8 with a confidence interval of ±0.1 and an alpha of .05.

Results

Participants

We recruited and screened 50 participants from the community. Informed consent was obtained from all individual participants included in the study. After the preplanned data filtering, 49 participants were included in the final analysis (see below). The mean participant age was 27.5 years (*SD* = 10.0, range = 18–73), and 26 (52%) of the participants were female. Education details, psychological scales, and language information are provided in Table 3. The mean scores for depression, anxiety, stress, and catastrophizing were in the normal range (Lovibond & Lovibond, 1995; Sullivan et al., 1995).

Data reduction

We excluded 315 (6.56%) of trials in accordance with our preplanned data filtering procedure. Seventy nine trials were excluded because no fixation was detected in both interest areas (13 trials had no fixations to either interest area, 66 trials had a fixation in only one interest area). A total of 37 trials were excluded when a fixation was detected less than 30 ms after the words were displayed. Another 135 trials were excluded when less than 75% (3,000 ms) of fixations were detected anywhere on the screen. The data reduction process resulted in one participant with less than 75% of their trials remaining (i.e., <36 trials). For this participant, in addition to the previously removed trials, all of the remaining trials were excluded (64 trials across both testing sessions). In all, 49 participants (4,485 trials) were included in the final analysis.

Table 3 Education, psychological scales, and language data for participants included in the final analysis

Participant Characteristics	*N*	Score	*SD*	Range
Highest Level of Education				
High School	16 (33%)			
Diploma TAFE	3 (6%)			
Bachelor degree or higher	30 (61%)			
Psychological Scales				
DASS-21: Depression (0–21)		1.2	1.8	0–6
DASS-21: Anxiety (0–21		1.2	1.7	0–7
DASS-21: Stress (0–21)		2.6	2.4	0–9
Pain Catastrophising Scale (0–52)		9.7	10.4	0–33
Self-Rated Proficiency in English (LEAP-Q)				
Speaking (0–10)		8.7	1.1	7–10
Understanding (0–10)		8.9	1.0	7–10
Reading (0–10)		9.0	1.0	7–10

DASS-21 Depression Anxiety and Stress Scales; *LEAP-Q* Language Experience and Proficiency Questionnaire

Test–retest reliability

Test–retest reliability data are presented in Table 4. Point estimates ranged from ICC(2, 1) = –.31 to .71. The sensory words had a lower mean ICC (.08) than the affective words (.32), and the general threat words (.29). Considering only the affective words and general threat words the total dwell time (0–4,000 ms) demonstrated the highest reliability (affective words: ICC = .61, general threat words: ICC = .71). The reliability coefficients for the affective and general threat words were also higher for the total dwell time (500–4,000 ms) and total dwell time (1,000–4,000 ms) (Table 4).

Measurement error

The *SEM* results are also presented in Table 4; lower *SEM*s represent more stable outcome measures. Point estimates for the *SEM* ranged between 3.02% and 14.59% across all word groups in all outcome measures. All word groups demonstrated a similar pattern of *SEM*s. The mean *SEM*s were 5.59% for the sensory words, 4.82% for the affective words, and 4.98% for the general threat words. The first-fixation duration recorded the lowest *SEM* scores (affective words *SEM* = 3.03%, general threat words *SEM* = 3.11%, sensory words *SEM* = 3.40%). The second-run dwell time demonstrated the highest *SEM*s, indicating less stable scores between testing sessions (affective words *SEM* = 13.43%, general threat words *SEM* = 11.21%, sensory words *SEM* = 14.49%).

Internal consistency

Finally, Cronbach's alpha scores for the first testing session are presented in Table 4. Point estimates ranged from .57 to .99 (mean = .89). Most outcome measures reported high internal consistency (e.g., total dwell time: affective words α = .94, general threat words α = .93, sensory words α = .94). The lowest Cronbach's alpha scores were recorded for the first-fixation duration (affective words α = .57, general threat words α = .67, sensory words α = .70) and second-run dwell

Table 4 Mean results from the two testing sessions, internal consistency as measured with Cronbach's alpha, test–retest reliability as measured with ICC(2, 1), and measurement error as measured with the standard error of measurement (*SEM*)

Stage of Attention	Outcome Measure	Word Category	Mean Test 1	*SD* Test 1	Mean Test 2	*SD* Test 2	Cronbach's α	ICC(2, 1)	95% CI		*SEM*
Overall	Total dwell time (0–4,000 ms)	Affective	48.86%	5.62%	48.81%	5.03%	.94	.61	.39	.76	3.34%
		General threat	48.06%	5.96%	48.76%	7.48%	.93	.71	.55	.83	3.61%
		Sensory	49.42%	3.47%	48.55%	6.18%	.94	.20	–.08	.46	4.47%
Overall	Total dwell time (0–500 ms)	Affective	48.89%	4.62%	50.60%	5.18%	.99	–.01	–.28	.26	5.00%
		General threat	50.23%	4.26%	49.41%	5.24%	.98	–.31	–.54	–.02	4.76%
		Sensory	50.70%	4.13%	49.98%	5.34%	.99	.12	–.16	.39	4.46%
Overall	Total dwell time (0–2,000 ms)	Affective	48.76%	4.12%	50.02%	3.68%	.85	.43	.17	.63	3.02%
		General threat	46.85%	4.75%	48.18%	5.68%	.88	.44	.19	.64	3.95%
		Sensory	49.15%	3.92%	48.85%	4.65%	.90	.21	–.08	.47	3.79%
Overall	Total dwell time (0–3,000 ms)	Affective	49.11%	4.66%	49.25%	4.25%	.91	.51	.27	.69	3.11%
		General threat	47.86%	5.66%	48.72%	6.21%	.92	.62	.42	.77	3.65%
		Sensory	49.50%	3.76%	48.89%	5.35%	.95	.17	–.11	.43	4.18%
Early	Probability of first fixation to target word	Affective	50.83%	6.22%	49.93%	6.12%	.97	–.15	–.42	.14	6.16%
		General threat	49.09%	4.80%	50.48%	6.35%	.96	–.18	–.43	.11	5.65%
		Sensory	49.07%	6.84%	49.58%	6.22%	.96	.06	–.23	.33	6.30%
Early	First-fixation latency	Affective	50.61%	2.94%	49.89%	3.50%	.98	.13	–.15	.39	3.03%
		General threat	49.97%	2.59%	50.46%	3.57%	.98	–.30	–.54	–.02	3.11%
		Sensory	49.75%	3.28%	49.98%	3.66%	.98	.03	–.26	.31	3.40%
Early	First-run dwell time	Affective	49.21%	4.22%	50.02%	3.32%	.93	.32	.05	.54	3.16%
		General threat	47.87%	4.24%	49.17%	4.63%	.91	.22	–.06	.46	3.98%
		Sensory	49.48%	4.38%	49.53%	3.77%	.94	.10	–.19	.37	3.84%
Early	First-fixation duration	Affective	50.26%	4.90%	48.68%	5.40%	.57	.30	.03	.53	4.40%
		General threat	49.21%	5.46%	49.53%	5.19%	.67	.01	–.28	.29	5.25%
		Sensory	48.95%	5.14%	48.37%	4.81%	.70	–.15	–.42	.14	4.96%
Late	Second-run dwell time	Affective	49.84%	13.64%	46.24%	15.70%	.72	.16	–.12	.42	13.43%
		General threat	52.40%	14.96%	49.79%	16.52%	.70	.50	.25	.69	11.21%
		Sensory	50.22%	11.45%	44.67%	16.49%	.72	–.02	–.28	.25	14.59%
Late	Last-run dwell time	Affective	50.80%	6.06%	48.64%	5.86%	.82	.49	.25	.68	4.36%
		General threat	47.95%	6.42%	47.88%	8.44%	.84	.36	.08	.58	5.98%
		Sensory	49.95%	4.99%	48.73%	7.79%	.83	–.17	–.43	.12	6.54%
Late	Total dwell time (500–4,000 ms)	Affective	48.78%	6.17%	48.67%	5.72%	.96	.57	.34	.73	3.89%
		General threat	47.77%	6.61%	48.77%	7.97%	.95	.70	.52	.82	4.03%
		Sensory	49.35%	4.02%	48.45%	6.71%	.96	.24	–.05	.48	4.83%
Late	Total dwell time (1,000–4,000 ms)	Affective	48.90%	7.68%	48.63%	6.90%	.96	.54	.31	.72	4.91%
		General threat	48.07%	7.36%	49.04%	9.10%	.96	.70	.52	.82	4.56%
		Sensory	49.57%	4.31%	48.53%	8.01%	.96	.21	–.08	.46	5.72%

time (affective words α = .72, general threat words α = .70, sensory words α = .72).

Discussion

We assessed the reliability of a preferential-looking eyetracking task to investigate attentional bias to threat-related words in healthy participants. Test–retest reliability varied according to the threat word category (sensory pain words, general threat words, and affective pain words) and outcome measure. Low ICCs were found for most outcome measures (e.g., first-fixation latency), indicating that they may not be appropriate for discriminative testing (when comparing participant groups). The results for the measurement error (*SEM*) suggest stable outcome measures between sessions and for internal consistency (Cronbach's α) a high level of interrelatedness between the word stimuli within each threat word category.

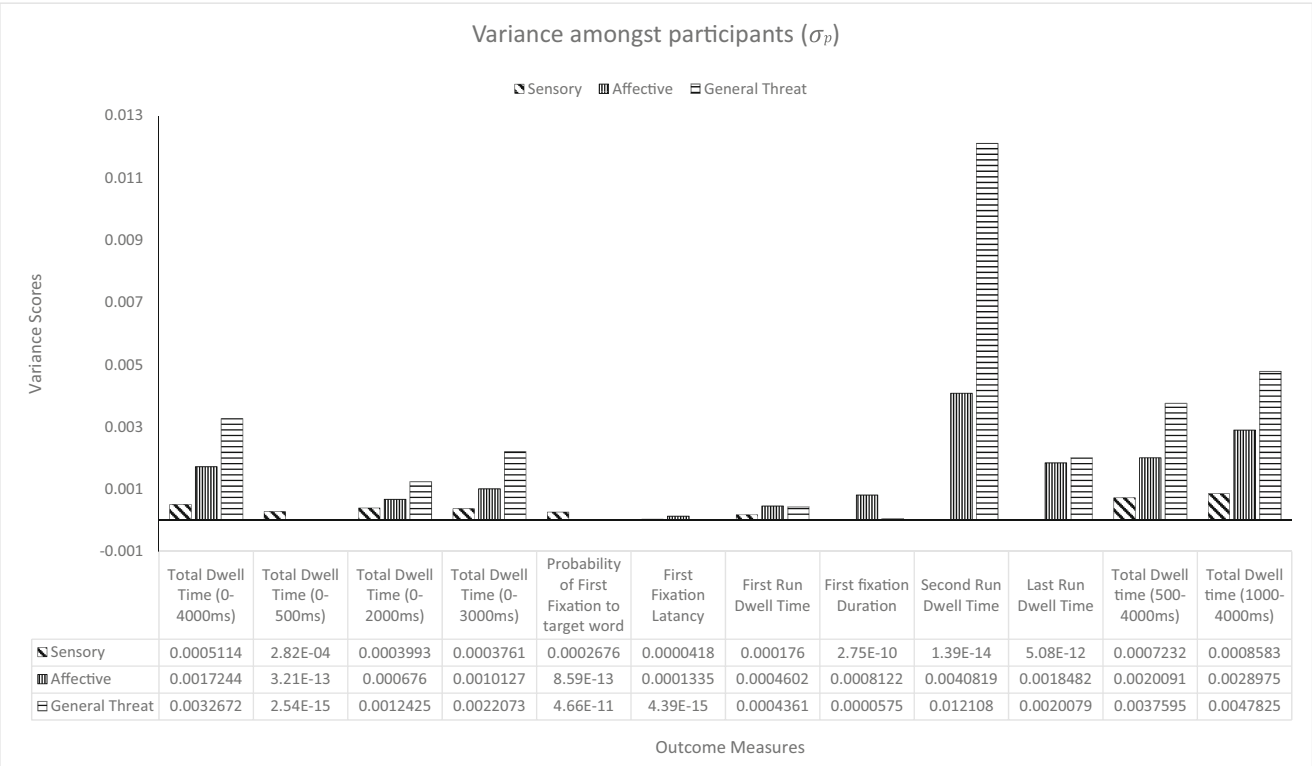
Test–retest reliability

Test–retest reliability varied according to the threat word category. Sensory pain words demonstrated the lowest test–retest reliability. Test–retest reliability considers the variance between a subject's repeated measurements relative to the overall group variance (de Vet et al., 2006). Decreased participant variance, relative to measurement error, decreases the test–retest reliability. When we examined the variance between participants (σ_p^2) across all word groups, there was less variance between participants for the sensory pain words (Fig. 1). It is not clear why the sensory pain words had less variance than the affective pain words and general threat words.

The second-run dwell time demonstrated high participant variance (Fig. 1) but still recorded low ICCs. The high variance between participants was not enough to overcome the relatively high measurement error between testing sessions.

Considering the different outcome measures available for researchers, our study showed that more reliable results are likely when one uses outcome measures that utilize more of the trial duration. Outcome measures that incorporated more of the 4,000-ms trial duration, such as the *total dwell time on threat words* (*0–4,000 ms*), demonstrated higher test–retest reliability than outcome measures that used less of the trial duration, such as the *total dwell time on threat words* (*0–500 ms*).

Furthermore, the outcome measures selected to reflect early attention (probability of first fixation to target word, first-fixation latency, first-run dwell time, and first-fixation duration) had lower test–retest reliability



	Total Dwell Time (0-4000ms)	Total Dwell Time (0-500ms)	Total Dwell Time (0-2000ms)	Total Dwell Time (0-3000ms)	Probability of First Fixation to target word	First Fixation Latancy	First Run Dwell Time	First fixation Duration	Second Run Dwell Time	Last Run Dwell Time	Total Dwell time (500-4000ms)	Total Dwell time (1000-4000ms)
Sensory	0.0005114	2.82E-04	0.0003993	0.0003761	0.0002676	0.0000418	0.000176	2.75E-10	1.39E-14	5.08E-12	0.0007232	0.0008583
Affective	0.0017244	3.21E-13	0.000676	0.0010127	8.59E-13	0.0001335	0.0004602	0.0008122	0.0040819	0.0018482	0.0020091	0.0028975
General Threat	0.0032672	2.54E-15	0.0012425	0.0022073	4.66E-11	4.39E-15	0.0004361	0.0000575	0.012108	0.0020079	0.0037595	0.0047825

Fig. 1 Participant variances (σ_p^2) for each outcome measure

than those selected to measure late attention [second-run dwell time, last-run dwell time, total dwell time (500–4,000 ms), and total dwell time (1,000–4,000 ms)]. Early attention outcome measures used less of the available viewing time and demonstrated less variance between participants than the late attention outcome measures (Fig. 1). This demonstrates that both the threat word group selected and the proportion of viewing time incorporated in the outcome measure are important procedural variables for the test–retest reliability of eyetracking measures.

We found higher test–retest reliability than Price et al. (2015). In their study, using a pediatric sample, facial stimuli were presented for 2,000 ms, whereas in our study the stimuli were presented for 4,000 ms (Price et al., 2015). It may be that increased stimulus exposure time allows greater variation, thereby increasing the ICC value. In support of this, Lazarov et al. (2016) presented stimuli for 6,000 ms and reported test–retest reliabilities of more than .62, using outcome measures that made use of longer stimulus exposure times—for example, total dwell time on threat faces. However, it may be that improved reliability for longer stimulus durations has a ceiling. The optimal stimulus duration to optimize test–retest reliability is likely related to the number and type of stimuli presented; for example, more stimuli may require longer exposure times, and pictures may require a longer presentation time than words. As was noted by Waechter et al. (2014), reliability is task and population specific.

Measurement error

The consistent and relatively low *SEM* values indicated stable measurements between sessions. The second-run dwell time was an exception, demonstrating higher *SEM* values (affective words = 13.4%, general threat words = 11.2%, sensory words = 14.6%) than the other outcome measures (all less than 6.6%). This was explained by the test–retest variance and reflected in the standard deviations of the mean scores (Table 4). The large standard deviations of the second-run dwell time suggest that there was considerable variability in the viewing patterns between test sessions. Because the *SEM* second-run dwell time values were higher than all other outcome measures, we would caution against using this outcome measure for discriminative or evaluative purposes when other, more reliable outcome measures are available. The results suggest the remaining outcome measures are appropriate for evaluative testing.

We are not aware of any other studies that have reported measurement error for eyetracking tasks that investigated attentional bias. We would encourage future research to report measurement error alongside other indicators of reliability. Because interest is growing in using the outcomes from eyetracking in interventional studies (Todd, Sharpe, & Colagiuri, 2016; Vazquez, Blanco, Sanchez, & McNally, 2016), it is important to know whether participant change scores are greater than the measurement error of the task.

Internal consistency

Our internal consistency results suggest that fewer test items could be used to achieve the same scores. Internal consistency measures the interrelatedness among items, and as such, high Cronbach's alpha scores suggests that using fewer stimuli may achieve the same scores for participants. Waechter et al. (2014) reported similar internal consistency results in a preferential-looking eyetracking task, measuring attentional bias using 72 trials. They reported Cronbach's alpha scores of .94, .94, and .96 for the total viewing time over 5,000 ms for angry, disgust, and happy images, respectively. This further suggests that when using the more stable and reliable outcome measures (longer proportion of viewing time) fewer items could potentially be used, thus reducing time involved for testing (Scholtes et al., 2011).

Individual variation

Researchers are commonly interested in testing for differences between groups. Test–retest reliability is the most informative reliability construct for that purpose. The nuance of test–retest reliability is that too little variance between participants will result in low reliability, (unable to distinguish participants). However, if measurements are not stable between sessions, this will also produce low reliability (too much variability between measurements). These effects are highlighted when examining measures of early attention. Location of first fixation and first-run dwell time both have low test–retest reliability; however, this is likely true for different reasons.

The poor reliability for location of the first fixation is most likely due to low variance between participants. Waechter et al. (2014) suggested low reliability may be due to the "look up" bias, in which participants will consistently look up first if stimuli are presented vertically or look left first if stimuli are presented horizontally. Viewing the word on the left first is consistent with the normal left to right reading pattern observed in English readers (Liversedge & Findlay, 2000; Rayner, 1989). Decreased variability between participants, due to normal reading patterns, is likely to reduce the test–retest reliability for the location of first fixation.

The low reliability coefficients reported for first-fixation duration to threat words is likely due to poor stability of measurements between sessions. In this context other factors that influence viewing patterns such as global speed of processing may be at play. This hypothesis also extends to first-run dwell

time and second-run dwell time outcome measures. Individual viewing patterns influenced the between-participant variation.

It may be that outcome measures that use more of the available viewing time strike a balance in having sufficient between participant variance but similar enough scores between testing sessions. In this study outcome measures that used more of the stimulus duration (e.g., 0–4,000 ms) were stable between measurements, and also not confounded by other individual viewing patterns such as global speed of processing.

It must be emphasized that reliability is specific to the population and the task for which it has been evaluated. The results of our study using healthy participants, words as stimuli, and a presentation time of 4,000 ms cannot be assumed to generalize to other populations (e.g., anxiety patients) or to other stimuli (e.g., images) or presentation times (e.g., 500 ms).

What is an acceptable level of reliability?

There is no definitive benchmark regarding an acceptable level of reliability (Charter & Feldt, 2001). The sample size, setting (i.e., clinical or research), and purpose (e.g., clinical diagnosis of life-threatening illness) will all contribute to the subjective assessment of what is acceptable in a specific situation. Although reliability benchmarks have not been well justified, some guidance is necessary (Streiner, 2003b). Nunnally (1994) suggested a value of .70 may equate to modest reliability when used to compare groups, and Cicchetti (1994) suggested a tiered approach for determining acceptability (i.e., <.40 = poor, .40–.59 = fair, .60–.74 = good, .75–1.00 = excellent). We would caution against using eyetracking measures with reliability coefficients less than .60, for research purposes. Outcome measures with higher reliability may be required when investigating between group differences with a small sample size (e.g., less than 20 participants). Our results suggest that most outcome measures are not reliable enough to differentiate participants when assessing attentional bias in healthy participants using threat words. Some of the outcome measures, such as the total dwell time of threat words (0–4,000 ms), may be appropriate depending on the stimulus (i.e., general threat words are appropriate but not sensory words).

Limitations

Although it is important that reliability be established for a healthy sample, our results may not generalize to nonhealthy samples. Reliability estimates are only valid for the sample being tested, and to the stimuli and outcome measures used in an experiment. The reliability of attentional bias using eyetracking has been investigated in a sample of participant with high and low social anxiety (Lazarov et al., 2016; Waechter et al., 2014). However, since these studies used facial images in nonclinical populations (participants were university students screened as having high or low social anxiety), it is unknown whether these results will generalize to other clinical samples. Further studies will be required that investigate reliability in clinical samples using a variety of stimuli (words/pictures/faces) and outcome measures, across all three measurement properties of reliability (internal consistency, agreement, and test–retest reliability).

Researcher degrees of freedom (RDoF) denote the decisions researchers make when collecting and analyzing data (Simmons, Nelson, & Simonsohn, 2011). There are many RDoF during eyetracking data filtering—for example, what constitutes a valid trial, and which fixations to retain for analysis. Minimizing RDoF, by specifying in advance how data will be collected and analyzed, decreases the risk of false positive results and may increase the reproducibility of findings (Simmons et al., 2011). We used a preplanned data-filtering process (Skinner et al., 2016). Stating in advance how and why one plans to remove trials avoids biased and subjective influences on the fixation locations (i.e., individual trials were not manipulated by the investigator). There is, however, the potential for removing trials unnecessarily, and thereby decreasing the power of statistical analysis. We argue that potentially removing unnecessary trials is an appropriate compromise for increased transparency in data analysis, decreasing RDoF, minimizing false positive results and potentially increasing reproducibility.

Our data-filtering method excluded trials with a first-fixation latency less than 30 ms, resulting in the exclusion of 46 trials. Previous research has used a more conservative cutoff (e.g., 80–100 ms); if we used a more traditional 80-ms cutoff, we would have included an additional nine trials. Rather than include the additional nine trials, we chose to preserve our a priori published data reduction plan. The first-fixation latency cutoff is another RDoF, which highlights the many decisions that researchers must make.

Recommendations

Our results suggest that for discriminative testing, outcome measures with a short exposure time or that use sensory words may be unreliable (low test–retest reliability). However, for evaluative testing, all of our outcome measures except second-run dwell time may be appropriate (low measurement error). Given that we found high internal consistency yet low test–retest reliability, Cronbach's alpha should not be used to justify the reliability of a task (Gliner, Morgan, & Harmon, 2001; Streiner, 2003a).

Our findings suggest that the outcome measures that investigate early stages of attentional bias are unreliable. One of the proposed advantages of eyetracking is the ability to distinguish early and late stages of attention. Our results suggest

that the current outcome measures used to assess early stages of attention do not have adequate test–retest reliability and are therefore unable to distinguish reliably the different stages of attentional bias.

Comparing our results to those of other studies suggests that the test–retest reliability of eyetracking is superior to that of the dot-probe task in healthy participants. Dear, Sharpe, Nicholas, and Refshauge (2011) reported bivariate reliability coefficients of –.06, –.14, and .01 for the dot-probe task using words on two occasions in healthy participants. Schmukle (2005) reported similarly poor test–retest reliability coefficients using a word-based dot-probe task. Evidence therefore suggests that when investigating attentional bias, eyetracking may provide higher test–retest reliability than the dot-probe task. This, however, needs confirmation across different populations and with different stimuli. Any potential benefit gained from eyetracking, such as increased reliability, will need to be considered against the increased cost of eyetracking equipment and the more complex data analysis techniques.

The challenge moving forward is to use outcome measures and stimuli that are relevant to both the population and the underlying mechanism being investigated, while still providing reliable data. We suggest reporting reliability statistics for test–retest, measurement error, and internal consistency for all tasks and outcome measures used to investigate attentional bias. With rapid advances in technology and the emerging prospect of virtual reality to assess attentional bias, it is critical that reliability be reported.

Conclusion

The outcome measure and threat word category used in eyetracking experiments influence test–retest reliability. Outcome measures with longer exposure times have increased test–retest reliability. Measurement error in eyetracking appears to be low. These results require replication in clinical populations and with different stimuli.

Author note

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Funding

I.W.S. is supported by an NHMRC Postgraduate scholarship (APP1093794); G.L.M. is supported by a Principal Research Fellowship from the NHMRC (ID 1061279); H.L. is supported by an NHMRC Postgraduate scholarship (APP133828); A.C.T. is supported by an NHMRC Postgraduate scholarship (APP1075670); S.M.G. is supported by an NHMRC Project Grant (ID 1084240) and Al & Val Rosenstraus Rebecca L. Cooper Medical Resesarch funding; and J.H.M. is supported by NHMRC Project Grants (ID 1008003 and 1043621).

Conflicts of interest

G.L.M. has received support from Pfizer, Australian Institute of Sport, Grunenthal, Kaiser Permanente California, Return to Work SA, Agile Physiotherapy, and Results Physiotherapy; grants from National Health and Medical Research Council of Australia; speaker fees for lectures on pain and rehabilitation; royalties from *Explain Pain*, *Painful Yarns*, *Graded Motor Imagery Handbook*, and *The Explain Pain Supercharged Handbook*: *Protectometer*, Noigroup Publications. All other authors declare they have no conflicts of interest.

References

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

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

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review, 32,* 704–723. doi:10.1016/j.cpr.2012.09.004

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

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin, 17,* 483–489. doi:10.1093/schbul/17.3.483

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology, 23,* 530–537. doi:10.1076/jcen.23.4.530.1227

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. doi:10.1037/1040-3590.6.4.284

Crombez, G., Van Ryckeghem, D. M., Eccleston, C., & Van Damme, S. (2013). Attentional bias to pain-related information: A meta-analysis. *Pain, 154,* 497–510. doi:10.1016/j.pain.2012.11.013

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

de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59,* 1033–1039. doi:10.1016/j.jclinepi.2005.10.015

de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide* (1st ed.). Cambridge: Cambridge University Press.

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. *Journal of Pain, 12,* 1247–1254. doi:10.1016/j.jpain.2011.07.003

Dehghani, M., Sharpe, L., & Nicholas, M. K. (2003). Selective attention to pain-related information in chronic musculoskeletal pain patients. *Pain, 105,* 37–46. doi:10.1016/s0304-3959(03)00224-0

Donaldson, C., Lam, D., & Mathews, A. (2007). Rumination and attention in major depression. *Behaviour Research and Therapy, 45,* 2664–2678. doi:10.1016/j.brat.2007.07.002

Duque, A. (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. doi:10.1016/j.jbtep.2014.09.005

Farzin, F., Scaggs, F., Hervey, C., Berry-Kravis, E., & Hessl, D. (2011). Reliability of eye tracking and pupillometry measures in individuals with fragile X syndrome. *Journal of Autism and Developmental Disorders, 41,* 1515–1522. doi:10.1007/s10803-011-1176-2

Fashler, S. R., & Katz, J. (2014). More than meets the eye: Visual attention biases in individuals reporting chronic pain. *Journal of Pain Research, 7,* 557–570. doi:10.2147/JPR.S67431

Felmingham, K. L., Rennie, C., Manor, B., & Bryant, R. A. (2011). Eye tracking and physiological reactivity to threatening stimuli in posttraumatic stress disorder. *Journal of Anxiety Disorders, 25,* 668–673. doi:10.1016/j.janxdis.2011.02.010

Gao, X., Wang, Q., Jackson, T., Zhao, G., Liang, Y., & Chen, H. (2011). Biases in orienting and maintenance of attention among weight dissatisfied women: An eye-movement study. *Behaviour Research and Therapy, 49,* 252–259. doi:10.1016/j.brat.2011.01.009

Giraudeau, B., & Mary, J. Y. (2001). Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistical Medicine, 20,* 3205–3214. doi:10.1002/sim.935

Gliner, J. A., Morgan, G. A., & Harmon, R. J. (2001). Measurement reliability. *Journal of the American Academy of Child and Adolescent Psychiatry, 40,* 486–488. doi:10.1097/00004583-200104000-00019

Guasch, M., Boada, R., Ferré, P., & Sánchez-Casas, R. (2013). NIM: A Web-based Swiss army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods, 45,* 765–771. doi:10.3758/s13428-012-0296-8

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40,* 171–178. doi:10.1016/0021-9681(87)90069-5

Jensen, M. P., Johnson, L. E., Gertz, K. J., Galer, B. S., & Gammaitoni, A. R. (2013). The words patients use to describe chronic pain: Implications for measuring pain quality. *Pain, 154,* 2722–2728. doi:10.1016/j.pain.2013.08.003

Kimble, M. O., Fleming, K., Bandy, C., Kim, J., & Zambetti, A. (2010). Eye tracking and visual attention to threating stimuli in veterans of the Iraq war. *Journal of Anxiety Disorders, 24,* 293–299. doi:10.1016/j.janxdis.2009.12.006

Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor ANOVAs. *Educational and Psychological Measurement, 51,* 585–595. doi:10.1177/0013164491513006

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology, 64,* 701–702. doi:10.1016/j.jclinepi.2010.12.001

Lakes, K. D. (2013). Restricted sample variance reduces generalizability. *Psychological Assessment, 25,* 643–650. doi:10.1037/a0030912

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. doi:10.1016/j.jad.2016.01.007

Liossi, C., Schoth, D. E., Godwin, H. J., & Liversedge, S. P. (2014). Using eye movements to investigate selective attention in chronic daily headache. *Pain, 155,* 503–510. doi:10.1016/j.pain.2013.11.014

Liversedge, & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4,* 6–14. doi:10.1016/S1364-6613(99)01418-7

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355,* 584–585. doi:10.1126/science.aal3618

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

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50,* 940–967. doi:10.1044/1092-4388(2007/067)

Meyer, J. P. (2010). *Reliability.* New York: Oxford University Press.

Mogg, K., Bradley, B. P., Field, M., & De Houwer, J. (2003). Eye movements to smoking-related pictures in smokers: Relationship between attentional biases and implicit and explicit measures of stimulus valence. *Addiction, 98,* 825–836. doi:10.1046/j.1360-0443.2003.00392.x

Mogg, K., Bradley, B., Miles, F., & Dixon, R. (2004). Time course of attentional bias for threat scenes: Testing the vigilance–avoidance hypothesis. *Cognition and Emotion, 18,* 689–700. doi:10.1080/02699930341000158

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63,* 737–745. doi:10.1016/j.jclinepi.2010.02.006

Nunnally, J. C. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., … & Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment, 27,* 365–376. doi:10.1037/pas0000036

Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European Journal of Cognitive Psychology, 16,* 3–26. doi:10.1080/09541440340000295

Rayner, K. (1989). In A. Pollatsek (Ed.), *Psychology of reading.* New Jersey: Prentice Hall.

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

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

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable? *Injury, 42,* 236–240. doi:10.1016/j.injury.2010.11.042

Sharpe, L., Haggman, S., Nicholas, M., Dear, B. F., & Refshauge, K. (2014). Avoidance of affective pain stimuli predicts chronicity in patients with acute low back pain. *Pain, 155,* 45–52. doi:10.1016/j.pain.2013.09.004

Sharpe, L., Ianiello, M., Dear, B. F., Nicholson Perry, K., Refshauge, K., & Nicholas, M. K. (2012). Is there a potential role for attention bias modification in pain patients? Results of 2 randomised, controlled trials. *Pain, 153,* 722–731. doi:10.1016/j.pain.2011.12.014

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. doi:10.1037/0033-2909.86.2.420

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Open Science Framework. (MT3K8). Skinner, I. W., Hübscher, M., Moseley, G. L., Lee, H., Traeger, A., Wand, B. M., Gustin, S. M., & McAuley, J. (Eds.) (2016) The test- retest reliability of eye tracking to measure attentional bias. Retrieved from http://www.osf.io/mt3k8 doi:10.17605/OSF.IO/MT3K8

Streiner, D. L. (2003a). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80,* 217–222. doi:10.1207/S15327752JPA8003_01

Streiner, D. L. (2003b). *Health measurement scales: a practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.

Streiner, D. L. (2003c). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99–103. doi:10.1207/S15327752JPA8001_18

Sullivan, M., Bishop, S., & Pivik, J. (1995). The Pain Catastrophizing Scale: Development and validation. *Psychological Assessment, 7,* 524–532. doi:10.1037/1040-3590.7.4.524

Todd, J., Sharpe, L., Johnson, A., Nicholson Perry, K., Colagiuri, B., & Dear, B. F. (2015). Towards a new model of attentional biases in the development, maintenance, and management of pain. *Pain, 156,* 1589–1600. doi:10.1097/j.pain.0000000000000214

Todd, J., Sharpe, L., & Colagiuri, B. (2016). Attentional bias modification and pain: The role of sensory and affective stimuli. *Behaviour Research and Therapy, 83,* 53–61. doi:10.1016/j.brat.2016.06.002

Toh, W. L., Rossell, S. L., & Castle, D. J. (2011). Current visual scanpath research: A review of investigations into the psychotic, anxiety, and mood disorders. *Comprehensive Psychiatry, 52,* 567–579. doi:10.1016/j.comppsych.2010.12.005

Vazquez, C., Blanco, I., Sanchez, A., & McNally, R. J. (2016). Attentional bias modification in depression through gaze contingencies and regulatory control using a new eye-tracking intervention paradigm: Study protocol for a placebo-controlled trial. *BMC Psychiatry, 16,* 439. doi:10.1186/s12888-016-1150-9

Vervoort, T., Trost, Z., Prkachin, K. M., & Mueller, S. C. (2013). Attentional processing of other's facial display of pain: An eye tracking study. *Pain, 154,* 836–844. doi:10.1016/j.pain.2013.02.017

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. doi:10.1007/s10608-013-9588-2

Weir, J. P. (2005). Quantifying test–retest reliability using the intraclass correlation coefficient and the *SE*M. *Journal of Strength and Conditioning Research, 19,* 231–240. doi:10.1519/15184.1

White, L. K., Suway, J. G., Pine, D. S., Bar-Haim, Y., & Fox, N. A. (2011). Cascading effects: The influence of attention bias to threat on the interpretation of ambiguous information. *Behaviour Research and Therapy, 49,* 244–251. doi:10.1016/j.brat.2011.01.004

Yang, Z., Jackson, T., Gao, X., & Chen, H. (2012). Identifying selective visual attention biases related to fear of pain by tracking eye movements within a dot-probe paradigm. *Pain, 153,* 1742–1748. doi:10.1016/j.pain.2012.05.011

Yang, Z., Jackson, T., & Chen, H. (2013). Effects of chronic pain and pain-related fear on orienting and maintenance of attention: an eye movement study. *Journal of Pain, 14,* 1148–1157. doi:10.1016/j.jpain.2013.04.017