Cognitive Therapy and Research (2023) 47:772–787
https://doi.org/10.1007/s10608-023-10395-4

ORIGINAL ARTICLE

Beyond Face Value: Assessing the Factor Structure of an Eye-Tracking Based Attention Bias Task

Mary E. McNamara[1] · Kean J. Hsu[2,3] · Bryan A. McSpadden[1] · Semeon Risom[1] · Jason Shumake[1] · Christopher G. Beevers[1]

Accepted: 20 May 2023 / Published online: 15 June 2023

Abstract

Background Behavioral measurement of attention bias for emotional stimuli has traditionally ignored whether trial-level task data have a strong enough general factor to justify a unidimensional measurement model. This is surprising, as unidimensionality across trials is an important assumption for computing bias scores.

Methods In the present study, we assess the psychometric properties of a free-viewing, eye-tracking task measuring attention for emotional stimuli. Undergraduate students ($N = 130$) viewed two counterbalanced blocks of 4×4 matrices of sad/neutral and happy/neutral facial expressions for 10 seconds each across 60 trials. We applied a bifactor measurement model across ten attention bias metrics (e.g., total dwell time for neutral and emotional stimuli, ratio of emotional to total dwell time, difference in dwell time for emotional and neutral stimuli, a variable indicating whether dwell time on emotional stimuli exceeded dwell time on neutral stimuli) to assess whether trial-level data load on to a single, general factor. Unidimensionality was evaluated using omega hierarchical, explained common variance, and percentage of uncontaminated correlations.

Results Total dwell time had excellent internal consistency for sad ($\alpha = .95$, $\omega = .96$) and neutral stimuli ($\alpha = .95$, $\omega = .95$), and met criteria for unidimensionality, suggesting the trial-level data within each task reflect a single underlying construct. However, the remaining bias metrics fell short of the unidimensionality thresholds, suggesting not all metrics are good candidates for creating bias scores.

Conclusion Total dwell time by valence had the best psychometrics in terms of internal consistency and unidimensionality. This study demonstrates the importance of assessing whether trial-level data load onto a general factor, as not all metrics are equivalent, even when derived from the same task data.

Keywords Attention bias · Measurement · Reliability · Cognitive-behavioral tasks · Psychometrics

Attention that is biased towards mood-congruent information, typically dysphoric information in depression, is a central mechanism of Beck's cognitive model of depression (Beck, 2008) and has been implicated in both the etiology and maintenance of depressive episodes (Disner et al., 2011). More specifically, depressed adults tend to display sustained attention for mood-congruent, dysphoric information, particularly in the later stages of information processing. In addition to sustained attention for dysphoric information, they often show reduced attention towards positive stimuli compared to non-depressed adults (for meta-analyses, see Armstrong & Olatunji, 2012; Suslow et al., 2020). Although many studies have examined the role of attention in depression, there is now growing recognition that the measurement of attentional bias has traditionally been suboptimal.

Earlier work concluded that the dot-probe task, arguably the most common measurement approach for measuring attentional bias, was completely unreliable in terms of internal consistency and test–retest reliability in nonclinical samples (Schmukle, 2005), making it unsuitable for measuring between-person differences (Hedge et al., 2018). Others have corroborated that the psychometrics of

✉ Mary E. McNamara
molly.mcnamara@utexas.edu

[1] Department of Psychology and Institute for Mental Health Research, University of Texas at Austin, 108 E Dean Keeton St, Austin, TX 78712, USA

[2] Department of Psychology, National University of Singapore, Singapore, Singapore

[3] Department of Psychiatry, Georgetown University Medical Center, Washington, DC, USA

the traditional bias metric derived from the dot-probe task is quite low (Machulska et al., 2022; Staugaard, 2009), and have called for the field to develop better assessments of attention bias (Kappenman et al., 2014; Rodebaugh et al., 2016). (Although see Price et al., 2015 for an example of reliable metrics derived from the dot-probe, such as attention bias variability).

One possible remedy is to measure line of visual gaze with eye tracking methods, as eye movements are thought to be a strong proxy for overt attention (Hayhoe & Ballard, 2005). Prior work suggests eye movements can produce internally consistent measurements of attentional bias for emotional stimuli. For instance, one study (Sears et al., 2019) measured eye movements in counterbalanced tasks of emotional facial expressions and natural scene stimuli. Each free-viewing task was comprised of trials of 4 images per screen: sad, positive, threatening, and neutral. A number of attentional bias indices had good test–retest reliability over a 6 month period (*rs* ranged from 0.36 to 0.80) and good internal consistency (αs ranged from 0.59 to 0.92) within each assessment time point (Sears et al., 2019). Similarly, another study examined twelve different attentional bias metrics for threat stimuli derived from eye movement data and found that reliability varied quite a bit across the metrics (Skinner et al., 2018). Results suggested that metrics involving gaze over longer periods of time, such as total dwell time for affective stimuli, had stronger internal consistency ($\alpha = .94$) and better test–retest reliability ($ICC = .61$) than metrics measuring early attentional components, such as initial orienting of attention towards threat ($\alpha = .98$, $ICC = .13$; see also Chong & Meyer, 2021).

These promising data suggest that eye tracking may improve the measurement of attentional bias, but additional psychometric work remains to be completed. Specifically, bifactor measurement models are routinely applied to psychopathology and personality questionnaire data to determine whether items are associated with a general factor and can be combined to form a total score (Rodriguez et al., 2016b). The importance of unidimensionality has been discussed in the context of self-report questionnaires (Stochl et al., 2020), leading some to suggest that total scores should not be used with certain questionnaires because the items do not load on to a single unidimensional factor (Fried et al., 2016). Further, it is possible to have a large coefficient alpha (good internal consistency) even if there are multiple underlying dimensions. As noted by Tavakol and Dennick (2011), "Internal consistency is a necessary but not sufficient condition for measuring homogeneity or unidimensionality in a sample of test items." This issue has not yet been addressed in the context of behavioral tasks, which we think is an important oversight, given that task trial data are routinely used to form a single metric of attention bias-–analogous to creating a total score on a questionnaire.

Therefore, in addition to examining internal consistency, the present study applied a bifactor model to determine which attention bias metrics derived from free viewing tasks with emotional (dysphoric/positive) and neutral stimuli while eye movements were obtained have a strong enough general factor to be considered unidimensional. The task used in the current study included a 4 × 4 matrix of faces (16 faces total) and was modeled after prior work (Lazarov et al., 2018), except instead of interleaving trials with happy and sad stimuli, two separate task blocks were created: one with neutral and sad facial expression stimuli and a second with matched neutral and happy facial expression stimuli (also similar to Klawohn et al., 2020; Sears et al., 2019).

A number of different bias metrics can be derived from this free viewing eye tracking task. We examined the internal consistency and unidimensionality of five attention bias metrics for each task block (10 metrics total): (1) Total dwell time for emotional stimuli; (2) Total dwell time for neutral stimuli; (3) Ratio of dwell time for emotional stimuli to total dwell time; (4) Percentage of trials where dwell time for emotional stimuli exceeded dwell time for neutral stimuli; (5) Dwell time difference score for emotional stimuli minus neutral stimuli. In each case, larger bias scores indicated greater attention towards the emotional stimuli. Additionally, we assessed a number of fixation-related metrics, including latency and length of first fixation, total count of fixations per AOI, a difference score (emotional- neutral fixations), and a proportion ratio (number of emotional fixations/ number of total fixations). These results can be found in our supplementary materials at the Texas Data Repository (https://dataverse.tdl.org/dataverse/factor_structure_attention) as early attentional biases are not typically observed for depression-related stimuli and total fixation counts are often strongly correlated with dwell time (Armstrong & Olatunji, 2012).

Methods

Participants

Participants were $N = 130$ undergraduate college students who received course credit for their participation. We collected data from as many participants as possible over the course of a year. Initially, 138 participants completed the study but 8 were excluded for not completing one of the two task blocks or for having > 50% missing data in a single task block (due to a technical malfunction). We describe how we handled data from participants missing < 50% of their trials in the "Missing Data" section below.

Participants were eligible for the study so long as they were (a) between the ages of 18–45 years old; (b) able to speak, read, and understand English fluently; and (c) willing

Table 1 Participant demographics

Characteristic	$N = 130$
Age in years, mean (SD)	19.4 (1.4)
Female gender (%)	73 (56.2%)
Hispanic ethnicity (%)	42 (32.3%)
Race (%)	
American Indian/Alaska Native	4 (3.1%)
Asian	38 (29.2%)
Black or African American	8 (6.2%)
White	63 (48.5%)
Multiracial	7 (5.4%)
Unknown or not reported	10 (7.7%)
Single (%)	129 (99.2%)
Years in school (SD)	13.9 (1.1)
Household income (%)	
\$0–\$24,999	16 (12.3%)
\$25,000–\$49,999	16 (12.3%)
\$50,000–\$74,999	12 (9.2%)
\$75,000–\$99,999	27 (20.8%)
\$100,000+	59 (45.4%)
BDI-II (20-items) (SD)	9.1 (7.9)
GAD-7 (SD)	4.8 (4.6)

and able to provide informed consent. Average age was 19.4 (SD = 1.4)[1] and the sample was majority female (56.2%). The most common race and ethnicity reported was white (48.5%) and non-Hispanic (67.7%). As we did not recruit for a clinical sample, average depression severity on the Beck Depression Inventory-II was low ($M = 9.1$, $SD = 7.9$), as was anxiety severity (Generalized Anxiety Disorder scale (GAD-7), $M = 4.8$, $SD = 4.6$). Full participant demographics can be found in Table 1. Ethical approval for the study was given by the University of Texas at Austin Institutional Review Board and written, informed consent was obtained from all participants. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Sample Size Justification

We conducted post-hoc simulations to confirm our sample was adequately powered for these analyses, the full results of which can be found at this link within the supplementary materials: https://doi.org/10.18738/T8/NHX2VA. In summary, these simulations suggest that a sample size of $n = 75$ to 200 is sufficient to obtain accurate estimates with acceptable levels of bias and good statistical power for the general factor loadings for a bifactor model with one general factor and four group factors (we detail our choice of four group factors in the "Data Analysis" section below). However, convergence rates for these models was below the acceptable threshold of 90%, which suggests that larger sample sizes are needed for a more reliable expectation that the bifactor models will converge. The sample sizes used in the current study should be able to accurately estimate variance explained for the general factor, but may experience sub-optimal model convergence. Larger sample sizes in the future should help overcome this limitation.

Materials

Depression severity was measured using the Beck Depression Inventory (BDI-II; Beck et al., 1996). The BDI-II is a widely-used 21-item questionnaire that measures the core symptoms of depression as defined by the Diagnostic and Statistical Manual (DSM-5; American Psychiatric Association, 2013), as well as other cognitive, motivational, and physical symptoms. Past research has documented decent test–retest reliability (*r* ranging between .73–.96) and internal consistency (alpha ranging between .83 and .96) for the BDI-II (Wang & Gorenstein, 2013). In the present study, we administered a 20-item version that excluded the suicidal ideation item. Internal consistency was strong ($\alpha = .91$, 95% CI [.87, .93], $\omega = .92$).

Apparatus

Eye position was measured using a video-based eye-tracker (EyeLink 1000 Plus Desktop Mount; SR Research, Osgoode, ON, Canada). Sampling was done at a rate of 250 Hz using the participant's dominant eye. Stimulus presentation was controlled by OpenSesame, a graphical experiment builder, with the back-end set to utilize PsychoPy (Mathôt et al., 2012). Data acquisition utilized Eyelink software. Stimuli were presented on a 23.6-inch CRT monitor (ViewPixx; VPixx Technologies, Quebec, Canada) at a screen resolution of 1920 × 1080 pixels (120 Hz refresh rate). Data was processed using Eyelink Data Viewer.

Eye-Tracking Task and Stimuli

Stimuli were chosen from the FACES dataset, which was developed to create a naturalistic dataset of facial expressions from people of varying ages (Ebner et al., 2010). The total dataset consists of 2052 photos taken of young, middle-aged, and older men and women. All models used for the image database were Caucasian and did not have any

[1] While 18–45 was the inclusion criteria for age of the sample, average age was 19.4 years (SD = 1.4), and maximum age was 28 years. Only 4 participants were over the age of 22 (23, 23, 24, and 28). We confirmed the inclusion of these participants in our sample had no meaningful impact on the results through supplementary analyses that can be found at https://dataverse.tdl.org/dataverse/factor_structure_attention.

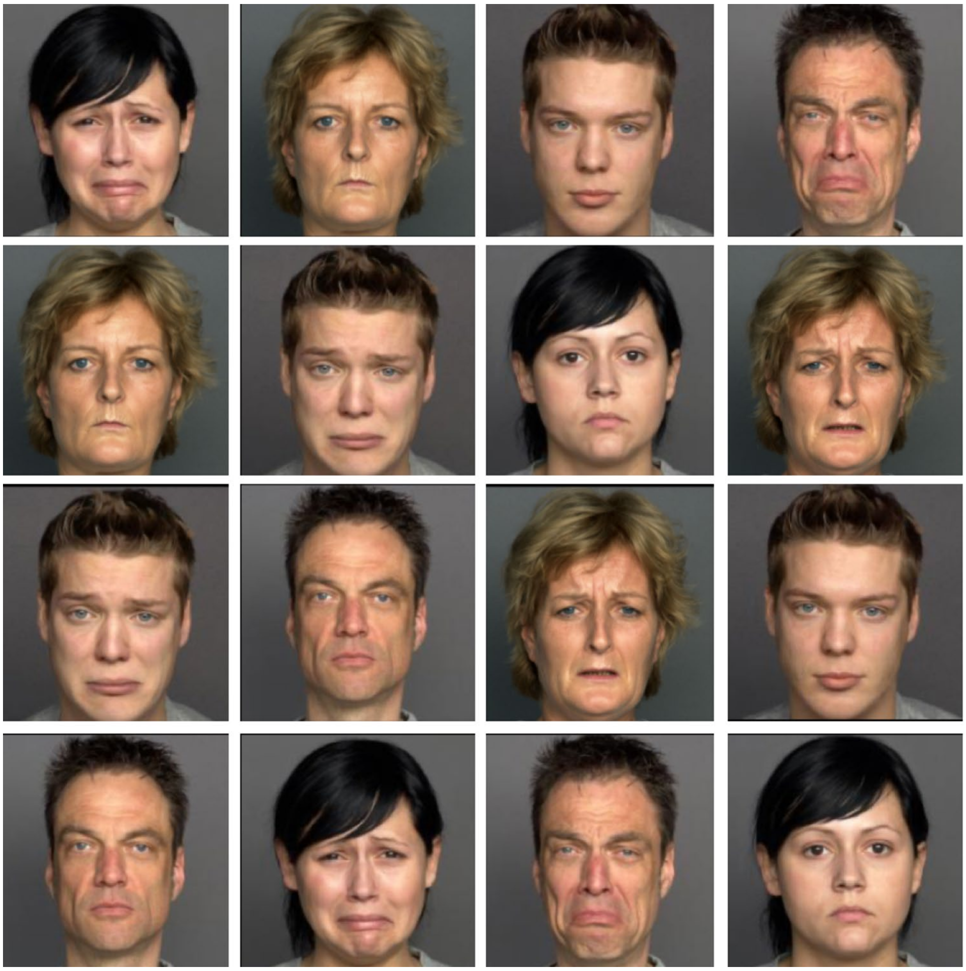

Fig. 1 Sample Trial: Matrix Stimulus Presentation. In our actual trials, we had the following constraints, consistent with Lazarov et al. (2016): **a** each actor could appear only once on the matrix, **b** there was an even split of genders in each matrix (8 male and 8 female), **c** there was an even split of valences in each matrix (e.g. 8 neutral and 8 sad), and **d** the four inner faces always contained two emotional and two neutral faces. A limited number of images from this stimuli set are approved for publication, so while criteria (**b**)–(**d**) are met, criterion (**a**) (each actor could only appear once on the matrix) is not fulfilled in this example (Ebner et al., 2010)

distinctive features (e.g. beards, piercings, etc.). In our task, we did not use the older actor photos and selected from the young and middle-aged actors evenly.

To help ensure facial stimuli were unambiguous (e.g., neutral expressions were unlikely to be mistaken for sad expressions), images were chosen based on previously documented accuracy ratings of the emotional faces (i.e., the percentage of raters who accurately identified the intended emotion, Ebner et al., 2010). The 32 top rated emotional and neutral pairs of stimuli from the same actor for both the happy and sad task blocks were selected, with a few exceptions: since we needed an equal balance of genders and ages (described below), once the quota for a particular demographic characteristic was filled, we skipped to the next available image of the desired group. For instance, if we already reached the needed number of female happy faces, we skipped over female faces to get to the next highest-ranked male face. Additionally, when we later cropped the images to be 200×200 pixels, if an actor's face was partially cropped, the image was replaced with the next best stimuli to ensure eye gaze was not drawn to the face because the image cropping created a visually distinct stimulus (e.g., face not centered in the image). A list of the images used can be found in the supplementary materials, along with additional details about how images were selected.

In designing the task, we followed similar parameters previously used by others (Klawohn et al., 2020; Lazarov et al., 2016, 2018). For each task block, we chose 64 photos of 16 male and 16 female actors, each contributing a neutral and emotional expression. These images were separated into four pools, so that each trial would contain 16 randomly selected images. The pools were generated with the following constraints: (a) each actor could appear only once on the matrix, (b) there was an even split of genders in each matrix (8 male and 8 female), (c) there was an even split of valences in each matrix (e.g., 8 neutral and 8 sad), and (d) the four inner faces always contained two emotional and two neutral faces. Fifteen trials were then generated from each of the four pools for a total of 60 trials per task block. The 60 unique 4×4 matrices were randomized so that each participant viewed the same stimuli matrices but in a different, random, order. An example of the stimuli configurations is presented in Fig. 1. Note that the sample trial shown in Fig. 1 is limited to images that are approved for publication,

Table 2 Paired samples t test and descriptive statistics

Task	Metric	Mean (SD)	t test	p value	95% CI	Cohen's d
Sad-neutral						
	Dwell time (sad stimuli)	3.39 (0.65)	t(129) = 1.44	.15	– 0.02, .15	0.099
	Dwell time (neutral stimuli)	3.33 (0.61)				
Happy-neutral						
	Dwell time (happy stimuli)	3.51 (0.58)	t(129) = 5.30	4.772e–07*	.15, .32	0.44
	Dwell time (neutral stimuli)	3.28 (0.47)				

*Statistical significance using Bonferroni-corrected alpha level is p = .004167. Correction was done accounting for the total number of tests (12), including those in the supplementary materials

Table 3 One-sample t-test and descriptive statistics

Task	Metric	Mean (SD)	t test	p value	95% CI	Cohen's d
Sad-neutral						
	Dwell ratio (sad stimuli/total stimuli)	0.50 (0.034)	t(129) = 1.42[a]	0.16	.50, .51	0.12
	Sad dwell time > neutral dwell time	0.52 (0.098)	t(129) = 2.11[a]	0.037	.50, .54	0.19
	Sad dwell time—neutral dwell time	0.063 (0.50)	t(129) = 1.44[b]	0.15	– 0.02, .15	0.13
Happy-neutral						
	Dwell ratio (happy stimuli/total stimuli)	0.52 (0.035)	t(129) = 5.57[a]	1.406e–07*	.51, .52	0.49
	Happy dwell time > neutral dwell time	0.54 (0.10)	t(129) = 4.92[a]	2.569e–06*	.53, .56	0.43
	Happy dwell time—neutral dwell time	0.24 (0.51)	t(129) = 5.30[b]	4.772e–07*	.15, .32	0.47

*Statistical significance using Bonferroni-corrected alpha level is p = .004167. Correction was done accounting for the total number of tests (12), including those in the supplementary materials

[a]Difference from 0.5

[b]Difference from 0

so while criteria (b)—(d) are met, criterion (a) (each actor could only appear once on the matrix) could not be fulfilled in this example (Ebner et al., 2010).

Attention Bias Measures

Within each 4 × 4 matrix, sixteen areas of interest (AOIs) were generated (i.e., one AOI for each individual stimulus in each matrix). For the purpose of these analyses, these AOIs were collapsed into two categories: neutral and emotional AOIs. We used fixation time for each AOI category on each trial to derive the following 5 metrics for each task block: (1) Total dwell time for emotion stimuli (Beevers et al., 2011; Bodenschatz et al., 2019; Duque & Vázquez, 2015; Klawohn et al., 2020; Lazarov et al., 2018; Wells et al., 2014) was calculated by summing total fixation time for the AOI across all trials and then dividing by the total number of trials to obtain average dwell time. Total dwell time could range from 0 to 10 s, the length of a trial. (2) Total dwell time for neutral stimuli was calculated in the same manner. (3) A ratio of dwell time for emotional stimuli was calculated by dividing the dwell time metric described above by the total amount of time spent viewing any stimuli (Kellough et al., 2008; Lanza et al., 2018; Owens & Gibb, 2017; Sanchez et al., 2013). This metric ranges from 0 to 1, with 0 indicating all dwell time was spent on neutral AOIs, 0.5 indicating dwell time was evenly split between emotional and neutral AOIs, and 1.0 indicating all dwell time was spent on emotional AOIs. (4) A variable indicating the percentage of trials where dwell time for emotional stimuli exceeded dwell time for neutral stimuli. This metric also ranges from 0–1, where 0 indicates dwell time for neutral stimuli exceeded dwell time for emotional stimuli on every trial, 0.5 indicates dwell time for emotional stimuli exceeded dwell time for neutral stimuli on 50% of the trials, and 1.0 indicates dwell time for emotional stimuli exceeded dwell time for neutral stimuli on all trials. Prior work suggests this metric has good psychometric properties (Beevers et al., 2019; Hsu et al., 2021). (5) Consistent with traditional bias scores from reaction time tasks, a trial-level difference score between dwell time for emotional stimuli and dwell time for neutral stimuli that was then averaged across all trials (Liu et al., 2017). A score of 0 indicates no difference in dwell time for neutral and emotional stimuli, whereas positive scores indicate a bias towards emotional stimuli and negative scores indicate a bias towards neutral stimuli.

Descriptive statistics can be found in Tables 2 and 3. Additionally, we assessed latency and length of first fixation

(Duque & Vázquez, 2015; Lazarov et al., 2018), total count of fixations per AOI (Kellough et al., 2008), a difference score (emotional- neutral fixations, Price et al., 2016), and a proportion ratio (number of emotional fixations/ number of total fixations, Soltani et al., 2015), the results of which can be found in the supplementary materials on the Texas Data Repository: https://dataverse.tdl.org/dataverse/factor_structure_attention.

Procedure

Participants were told they were taking part in a study attempting to better understand how people interact with facial stimuli. The experiment consisted of two separate task blocks which utilized sad and neutral faces in one block or happy and neutral faces in the other block. Each participant completed both task blocks in a counterbalanced order. Participants sat in an illuminated room (12.0 cd/m^2) at a distance of 60 cm from the screen. Each subject's dominant eye was determined using a modified version of the near-far alignment task (Miles, 1930). Prior to beginning the task, a thirteen-point calibration routine was used to map the subject's gaze onto the screen coordinates. We did not restrict head position using a headrest, but instead allowed for natural head movement and utilized a head-based tracker to provide consistent eye tracking.

Both task blocks consisted of 60 trials with a break every 30 trials. A fixation dot was presented for 1000 ms, followed by the matrix stimulus presentation for 10,000 ms, contingent on the participant's fixation on the dot for 1000 ms. Each free-viewing task took approximately 11 min to complete. Both task blocks began with a practice trial. Participants were first presented with a fixation dot and told to fixate on it when it appeared. Subjects were then given the following instructions: "Before each matrix, a fixation dot will appear on the screen. Make sure to fixate on the dot when it appears. When the matrix appears, look at the images freely and naturally. Do you have any questions?" After completing 30 trials, participants were encouraged to take a break to rest their eyes.

Missing Data

Consistent with prior work (Hsu et al., 2020; McNamara et al., 2021), we maximized the number of observations for the analyses for each of the tasks. Our factor analysis utilized trials as the items in the analyses, so it was essential that participants have some data for almost every trial. Prior to conducting factor analysis on each metric, we filtered out participants that were missing data for more than 2 trials due to not fixating on the screen (different participants were filtered out by this standard for the different metrics). That is, while we had a total sample of $N = 130$, for the factor analysis of individual metrics, the analyzed sample ranged from $Ns = 121–126$. For instance, for dwell time for happy stimuli, 4 individuals were excluded for missing more than 2 trials. For dwell time for sad stimuli, 5 individuals were excluded, but none of these individuals were the same as the 4 excluded from the happy data task. Therefore, we retained a total sample of $N = 130$, but the samples for the individual metrics vary slightly, depending on missing data.

Data Analysis

Data analysis was conducted in R (version 4.2.3) and made extensive use of the *tidyverse* packages (Wickham et al., 2019), as well as an in-house package *itrak* developed for processing eye-tracking data (https://github.com/jashu). Exploratory factor analysis was completed using the omega function in the *psych* package (Revelle, 2019), as well as the *GPArotation* package (Bernaards & Jennrich, 2005). Other packages used included *parallel* (R Core Team, 2020)*, splithalf* (Parsons, 2020)*, readxl* (Wickham & Bryan, 2019)*, scales* (Wickham & Seidel, 2020)*, boot* (Canty & Ripley, 2019)*, gt* (Iannone et al., 2020)*, broom* (Robinson et al., 2020)*, knitr* (Xie, 2014)*, EFAtools* (Steiner & Grieder, 2020), and *Hmisc* (Harrell et al., 2020). All analysis code can be found in a supplementary document titled, "Matrix Main Analyses.pdf" located within our dataverse on the Texas Data Repository at https://doi.org/10.18738/T8/JCUWXJ.

We used paired-samples t test to compare dwell times for emotional and neutral stimuli in both the sad and happy face versions of the task, along with their Bonferroni-corrected p-values. We report both coefficient α and ω using the omega function in the *psych* package, as previous work has advocated for a shift from solely relying on Cronbach's α as a measure of internal consistency (Dunn et al., 2014; Flora, 2020).

Internal Consistency

To calculate split-half reliability, we computed Pearson's *r* with the Spearman-Brown correction. Trials were randomly permuted and then split into two groups and the correlation between the two computed. The Spearman-Brown correction takes the form of $r = (2*r)/(r + 1)$. Confidence intervals for the reliability statistic were obtained by bootstrap resamples of the split. Consistent with recommendations (Machulska et al., 2022), we report both the uncorrected and Spearman-Brown corrected split-half reliabilities.

Bifactor Analyses

The factor structures of the metrics were also derived using the omega function and the *GPArotation* package. Although we hypothesized a unidimensional model, we also tested an exploratory four-factor bifactor model solution to assess the possibility that factor groups contributed variance above and beyond the general factor (Rodriguez et al., 2016a). We specified four for the number of factor groups since we had four image pools to create the 60 stimuli presentations. Items were considered to have loaded onto a general factor if they had a Schmid Leiman Factor loading greater than 0.2.

We evaluated the models using several indices. First, we examined root mean squared error of approximation (RMSEA) of both models for each metric, where values < .06 indicate a relatively good model fit for the data (Hu & Bentler, 1999). However, since RMSEA will tend to decrease (i.e. improve) with greater model complexity, we only use this index to assess an adequate fit for the data (Shi et al., 2019).

Parameters to Assess Unidimensionality

Next, we gauged the appropriateness of treating the data as a unidimensional model by consulting omega hierarchical, explained common variance, and percent uncontaminated correlations, each of which are described in more detail in the subsequent paragraph (Reise et al., 2013; Rodriguez et al., 2016a, 2016b). If the use of the unidimensional model was still justifiable, we then used the Bayesian information criterion (BIC) index to ascertain the best model fit.

Perfect unidimensionality rarely exists, and some level of multidimensionality is usually present. When assessing bifactor models, we need to assess whether the presence of some multidimensionality in the data is minimal enough to fit a unidimensional model to the data (particularly when we are most interested in a total score as represented by a single general factor, as is our case). Omega hierarchical (ω_h), explained common variance (ECV), and percent uncontaminated correlations (PUC) are important indices to determine whether a unidimensional model is defensible. Importantly, it is considered acceptable to interpret these indices for exploratory bifactor models (Rodriguez et al., 2016b).

The ω_h indexes the amount of variance in total scores produced by a general factor, versus omega which takes into account all sources of common variance. A larger ω_h indicates that the general factor is the primary source of variance, despite the presence of multidimensionality in the data due to the group factors. When multidimensionality in the data is present but the general latent factor is the factor of interest (as in our case), ECV can be a useful guide for determining whether the data is "unidimensional enough" (Rodriguez et al., 2016a). ECV is calculated by taking the proportion of variance explained by a general factor and dividing it by the total variance explained by both the general and group factors. Thus, a larger ECV value is driven by a larger numerator (e.g. greater variance explained by a general factor), signaling a strong general factor. Importantly, while ω_h is sensitive to the number of items used, ECV is not.

Finally, PUC is important to help contextualize ECV values in regard to the overall data structure (Rodriguez et al., 2016a). To understand PUC, imagine you are looking at the correlations between different items within the bifactor model. PUC is the count of correlations between items across group factors divided by the total count of unique correlations. This can also be understood as 1—(the count of correlations within group factors divided by the total number of unique correlations, Rodriguez et al., 2016b).

Parameter Thresholds to Determine Unidimensionality

A larger PUC value suggests that more information in the correlation matrix relates to a general factor, as well as a lower risk of bias (Rodriguez et al., 2016b). At high levels of PUC, the magnitude of ECV is less important for assessing the appropriateness of fitting a unidimensional model to the data. Reise et al. (2013) determined that when PUC values exceed .80, the other indices are less critical for assessing multidimensionality. Beneath a PUC of .80, they suggest that ECV values > .60 and $\omega_h > .70$ may be acceptable benchmarks to consider (Reise et al., 2013). Rodriguez et al. give a similar suggestion, proposing that when both ECV and PUC > .70, the data can be considered "essentially unidimensional" (Rodriguez et al., 2016a). These benchmarks are meant to be used as guides for evaluating models, rather than strict rules of thumb. Nevertheless, we compare each of our metrics to these standards to evaluate the appropriateness of treating the metric as unidimensional.

It is important to keep in mind when reviewing these psychometric results that each item is a 4 × 4 matrix of emotional and neutral faces. For instance, split-half reliability is calculated at the level of trial, not image, and the item analysis uses each matrix of stimuli, not each individual actor's photograph.

Results

Descriptive Statistics

Descriptive statistics of each of the attention bias metrics are presented in Tables 2 and 3. There were no differences in total dwell time on sad faces versus neutral faces in the sad face version of the task, t(129) = 1.44, p = .15, d = .099. Dwell time for sad faces and neutral faces was significantly correlated, however, *r* = .69, p < .001. There was a difference in total dwell time between neutral and happy faces in the happy version of the task, t(129) = 5.30, p < .001, d = 0.44. Participants spent

Table 4 Internal consistency indices

Bias metric	Alpha	Omega	Split-half correlation (r)	Spearman-Brown corrected r (95% CI)	Proportion of trials with a positive Schmid Leiman factor loading greater than 0.2
Sad-neutral face stimuli					
Dwell time for sad faces	.95	.96	.90 (.86, .92)	.95 (0.93–0.96)	60/60 (100%)
Dwell time for neutral faces	.95	.95	.89 (.86, .92)	.94 (.92–.96)	59/60 (98.3%)
Dwell time ratio (sad dwell time: total dwell time on stimuli)	.68	.73	.49 (.36, .60)	.66 (0.52–0.75)	19/60 (31.7%)
Sad dwell time > neutral dwell time	.58	.62	.39 (.28, .50)	.57 (.44–.67)	2/60 (3.3%)
Difference score: sad-neutral dwell time	.76	.80	.59 (.49, .68)	.74 (.66–.81)	10/60 (16.7%)
Happy-neutral face stimuli					
Dwell time for happy faces	.93	.94	.88 (.84 -. 90)	.93 (.91 -.95)	59/60 (98.3%)
Dwell for neutral faces	.89	.91	.82 (.77, .86)	.90 (.87–.93)	46/60 (76.7%)
Dwell time ratio (emotional dwell time: total dwell time on stimuli)	.71	.75	.52 (.42, .62)	.69 (.59–.77)	14/60 (23.3%)
Happy dwell time > neutral dwell time	.64	.67	.46 (.36, .55)	.63 (.53–.71)	2/60 (3.3%)
Difference score: happy-neutral dwell time	.76	.79	.60 (.51, .69)	.75 (.67–.81)	13/60 (21.7%)

significantly more time (in seconds) viewing happy faces ($M = 3.51$, $SD = .58$) during each trial, compared to neutral faces ($M = 3.28$, $SD = .47$). Total dwell time for happy faces and neutral faces was also strongly correlated, r = .55, p < .001.

The dwell time ratio for sad stimuli was not significantly different from 0.5 (t(129) = 1.42, p = .16, d = .12). Proportion of trials where sad dwell time exceeded neutral was significantly different from 0.5 (t(129) = 2.11, p = .037, d = .19). The trial-level difference score for sad—neutral dwell time was also not significantly different from 0, indicating little support for a preference for sad stimuli, t(129) = 1.44, p = .15, d = .13. Taken together, this suggests little to no differences in attention to sad versus neutral stimuli.

However, all of the happy/neutral task counterparts were statistically significant. The ratio for dwell time for happy stimuli relative to total dwell time was significant, suggesting that participants spent more time viewing happy than neutral faces, t(129) = 5.57, p < .001, d = .49. The metric indicating the percentage of trials where happy dwell time exceeded neutral dwell time was significantly different from 0.5, t(129) = 4.92, p < .001, d = .43, again suggesting more time spent viewing happy faces. And the difference score for happy-neutral stimuli was significant, also indicating a preference for happy faces, t(129) = 5.30, p < .001, d = .47.

Internal Consistency and Assessment of Factor Structure

The internal consistency statistics and proportion of items with positive loadings onto a general factor can be found in Table 4. Model fit statistics can be found in Table 5, and unidimensionality indices in Table 6.

Sad/Neutral Task Data

Sad Dwell Time

Internal consistency was strong across four indices for the sad dwell time metric ($\alpha = .95$, ω total = .96, split-half correlation = .90, Spearman-Brown corrected r = .95), and every item (trial) mapped onto a general factor with a factor loading of > .2 (see Fig. 2). RMSEA for both the general factor (.046, 90% CI [.041, .053]) and bifactor models (.036, 90% CI [.029, .044]) were beneath the .06 threshold, suggesting both models are a good fit for the data. $\omega_h = .70$, ECV > .60, and PUC > .70, providing strong support that the data can be interpreted as unidimensional. A smaller BIC of − 6076.94 for the general factor model compared to the bifactor model (− 5620.39) further supports a unidimensional measurement model.

Neutral Dwell Time

Internal consistency was similarly strong for dwell time on neutral faces within the sad-neutral task ($\alpha = .95$, $\omega = .95$, $r = .89$, Spearman-Brown corrected r = .94). Model fit was adequate for the general factor model, RMSEA = .045, 90% CI [.040, .052], and the bifactor model, RMSEA = .031, 90% CI [.023, .040]. Nearly all (98.3%) of items mapped onto a general factor. All of the unidimensionality bias

Table 5 Model fit statistics

Metric	Model	RMSEA and 90% CI	BIC
Dwell time			
Dwell time for sad stimuli (sad/neutral task)	Unidimensional	.046 [.041, .053]	− 6076.94
	Bifactor	.036 [.029, .044]	− 5620.39
Dwell time for neutral stimuli (sad/neutral task)	Unidimensional	.045 [.040, .052]	− 6100.11
	Bifactor	.031 [.023, .040]	− 5684.89
Dwell time for happy stimuli (happy/neutral task)	Unidimensional	.053 [.048, .059]	− 5951.88
	Bifactor	.038 [.032, .046]	− 5596.21
Dwell time for neutral stimuli (happy/neutral task)	Unidimensional	.052 [.047, .058]	− 5971.77
	Bifactor	.036 [.030, .044]	− 5622.56
Dwell time ratio			
Dwell time ratio for sad stimuli (sad/neutral task)	Unidimensional	.064 [.059, .070]	− 5663.23
	Bifactor	.051 [.046, .058]	− 5367.89
Dwell time ratio for happy stimuli (happy/neutral task)	Unidimensional	.058 [.054, .064]	− 5817.08
	Bifactor	.045 [.039, .052]	− 5494.19
Emotional dwell time > neutral dwell time			
Sad dwell time > neutral dwell time (sad/neutral task)	Unidimensional	.013 [0, .027]	− 6497.78
	Bifactor	0.0 [0, 0]	− 5983.78
Happy dwell time > neutral dwell time (happy/neutral task)	Unidimensional	.024 [.014, .034]	− 6418.28
	Bifactor	.008 [0, .025]	− 5868.52
Difference score			
Dwell time for sad stimuli—dwell time for neutral stimuli (sad/neutral task)	Unidimensional	.055 [.051, .062]	− 5875.15
	Bifactor	.043 [.036, .050]	− 5520.20
Dwell time for happy stimuli—dwell time for neutral stimuli (happy/neutral task)	Unidimensional	.062 [.058, .068]	− 5721.91
	Bifactor	.045 [.039, .052]	− 5484.51

Table 6 Unidimensionality indices

Bias metric	Omega hierarchical (ω_h)	Explained common variance (ECV)	Percent uncontaminated correlations (PUC)
Sad-neutral face stimuli			
Dwell time for sad faces	0.70	0.65	0.73
Dwell time for neutral faces	0.71	0.65	0.74
Dwell time ratio (sad dwell time: total dwell time on stimuli)	0.14	0.38	0.67
Sad dwell time > neutral dwell time	0.08	0.08	0.76
Difference score: sad-neutral dwell time	0.28	0.20	0.76
Happy-neutral face stimuli			
Dwell time for happy faces	0.58	0.49	0.70
Dwell for neutral faces	0.44	0.34	0.73
Dwell time ratio (emotional dwell time: total dwell time on stimuli)	0.29	0.32	0.67
Happy dwell time > neutral dwell time	0.14	0.12	0.68
Difference score: happy-neutral dwell time	0.15	0.26	0.66

indices exceeded the specified thresholds ($\omega_h > .70$, ECV > .60, and PUC > .70). A BIC of − 6100.11 for the general factor model versus BIC = − 5684.89 for the bifactor model also supports the use of the unidimensional measurement model.

Dwell Time Ratio

The ratio of dwell time for sad faces to dwell for all stimuli had mixed internal consistency statistics (α = .68, ω = .73, *r* = .49, Spearman-Brown corrected r = .66). RMSEA for

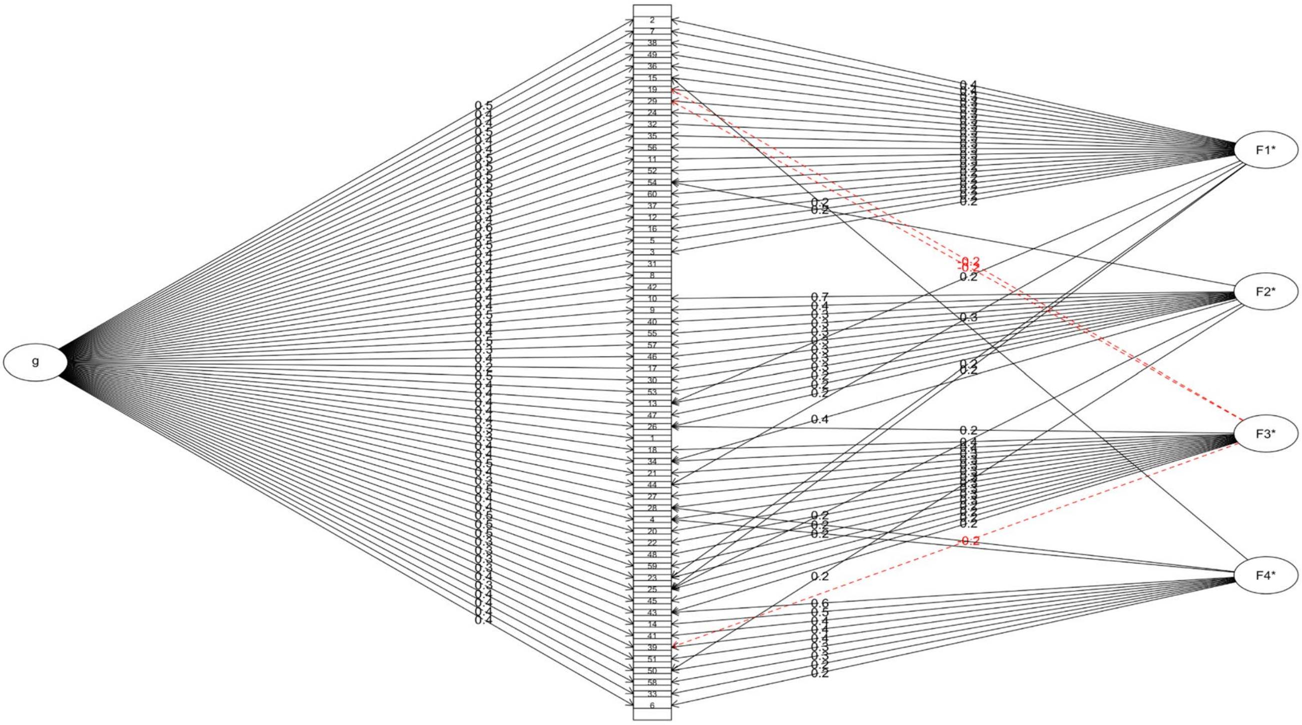

Fig. 2 Factor structure of the total dwell time for sad faces metric. Trials for this metric load well onto a single, general factor, which means it is appropriate to combine these trials into a single outcome (e.g., averaging dwell time across trials)

the general factor model suggested a less than adequate fit (.064, 90% CI [.059, .070]), though the bifactor model met the criteria (.051, 90% CI [.046, .058]). However, only approximately one third of the items had positive loadings onto the general factor. The unidimensionality indices further suggested a high risk of bias when trying to fit a unidimensional model to this data, thus we did not consider that model further.

Dwell Time for Sad Faces Exceeds Dwell Time for Neutral Faces (Sad Dwell Time > Neutral Dwell Time)

The percentage of trials where sad dwell time > neutral dwell time had acceptable internal consistency (α = .58, ω = .62, *r* = .39, Spearman-Brown corrected r = .57). However, only 2 of the 60 items mapped onto a general factor. RMSEA values suggest an adequate fit for both the general factor model (.013, 90% CI [0, .027]) and bifactor model (0.0, 90% CI [0, 0]), yet none of the unidimensionality indices suggested this metric should be treated as unidimensional (ω_h = .08, ECV = .08, PUC = .76).

Difference Score: Sad Dwell Time—Neutral Dwell Time

The final metric, the difference score of sad—neutral dwell time, had mostly good internal consistency (α = .76, ω = .80, *r* = .59, Spearman-Brown corrected *r* = .74). Both the general factor and bifactor models met the adequacy threshold, although the upper limit of the 90% confidence interval for the general factor model exceeded .060. Only 16.7% of trials mapped onto the general factor. The unidimensionality indices further suggest that the amount of multidimensionality in this metric preclude interpreting the data as unidimensional.

Happy/Neutral Task Data

Happy Dwell Time

Dwell time for happy faces had excellent internal consistency across the four metrics (α = .93, ω = .94, r = .88, Spearman-Brown corrected r = .93). RMSEA values indicated an adequate fit for both the general factor and bifactor model. The vast majority of items mapped onto a single, general factor; however, while PUC met the cutoff at = .70, the other metrics signal significant multidimensionality in the data (ECV = .49 and ω_h = .58).

Neutral Dwell Time

Dwell time for neutral faces also displayed strong internal consistency (α = .89, ω = .91, *r* = .82, and Spearman-Brown corrected *r* = .90). Both the general factor and bifactor models met the RMSEA < .06 criteria for adequate fit, and roughly three-fourths of the items mapped onto the general factor. However, the risk of bias indices for assessing unidimensionality failed to meet the specified thresholds ($\omega_h = .44$, ECV = .34, PUC = .73).

Dwell Time Ratio

The ratio of dwell time for happy faces to total dwell time on stimuli had moderate internal consistency. The RMSEA for the general factor model met the .060 cutoff for adequate fit; however, the upper limit of the 90% confidence interval did not. Less than a quarter of the items mapped onto the general factor. Additionally, values of $\omega_h = .29$, ECV = .32, and PUC = .67 suggest fitting a unidimensional model to this data would generate significant bias.

Dwell Time for Happy Faces Exceeds Dwell Time for Neutral Faces (Happy Dwell Time > Neutral Dwell Time)

Internal consistency metrics for the happy dwell time > neutral dwell time ranged from .46 to .67. RMSEA values for both models met the adequate fit criteria. The unidimensionality indices did not meet the thresholds, however ($\omega_h = .14$, ECV = .12, PUC = .68). Two items mapped onto the general factor, a mere 3.3%.

Difference Score: Happy Dwell Time—Neutral Dwell Time

The trial level difference score for happy dwell time—neutral dwell time displayed good measures of internal consistency ($\alpha = .76$, $\omega = .79$, $r = .60$, Spearman-Brown corrected $r = .75$). However, the general factor model's RMSEA exceeded the .060 threshold, indicating the unidimensional measurement model is an inappropriate fit for this data. About 20% of the items mapped onto the general factor. The unidimensionality indices also support the idea that fitting a unidimensional measurement model to this metric's data would be biased.

Brief Summary of Supplementary Analyses

An alternative way attention bias metrics are often calculated uses the number of fixations, rather than summing the duration of fixations to calculate dwell time as we did here. The interested reader can find psychometric analyses of the number of fixations per emotional AOI, number of fixations per neutral AOI, a difference score (number emotional fixations—neutral fixations), and a proportion score (number of emotional fixations/number of total fixations) for both the sad/neutral and happy/neutral task data within our supplementary materials: https://dataverse.tdl.org/dataverse/factor_structure_attention. The results are consistent with what we present here regarding total dwell time in each AOI: the number of fixations per AOI metrics are generally quite strong in terms of internal consistency and unidimensionality. The bias scores derived from them, however, are not. The supplementary materials also include analyses of the latency and length of first fixation metrics. The internal consistency of these tended to be lower, which is consistent with what others have found when testing these metrics using dysphoric stimuli (Lazarov et al., 2018), and the unidimensionality was very poor in this sample.

Discussion

In the present study, we evaluated the internal consistency and factor structure of stimuli presentations for two versions of a free-viewing, eye tracking task designed to assess attention bias for negative and positive stimuli (i.e., sad-neutral and happy-neutral facial expressions). In addition, we examined the appropriateness of fitting a general factor versus a four-factor bifactor model using five different attention bias metrics for each task. This study was particularly novel in its evaluation of using bifactor models to assess the unidimensionality of the task stimuli, which has not been applied to behavioral tasks despite its common use and utility when evaluating self-report measures.

Consistent with prior work (Lazarov et al., 2018), total dwell time had the best psychometric properties in these free viewing tasks. Dwell time for sad and neutral stimuli both demonstrated strong internal consistency measured by α, ω, split-half correlation, and Spearman-Brown corrected r (all values > .89). Dwell time for happy stimuli and neutral stimuli from the happy/neutral task also had conventionally strong internal consistency metrics (all values > .82). Internal consistency indices for the remaining metrics across both tasks showed low to moderate internal consistency.

However, out of the ten metrics evaluated, *only* the dwell time for sad and neutral stimuli metrics from the sad/neutral task met criteria for unidimensionality.[2] Thus, although several indices had adequate internal consistency, the task trials appear to be drawn from a single underlying construct only when considering total dwell time for sad/neutral stimuli. Given that the other metrics likely reflect a multidimensional construct, this would suggest that using these metrics to create a single bias score may be problematic. We posit that this multidimensionality of task trials for these other metrics could be an important, and until now unmeasured, reason for why many attention bias tasks may be highly inconsistent across the literature. This issue has been raised for the

[2] As a reminder, the stimuli selection processes for each task were independent, so different pools of neutral facial expressions are used in the happy version and sad version (some of the same neutral faces may be in both tasks).

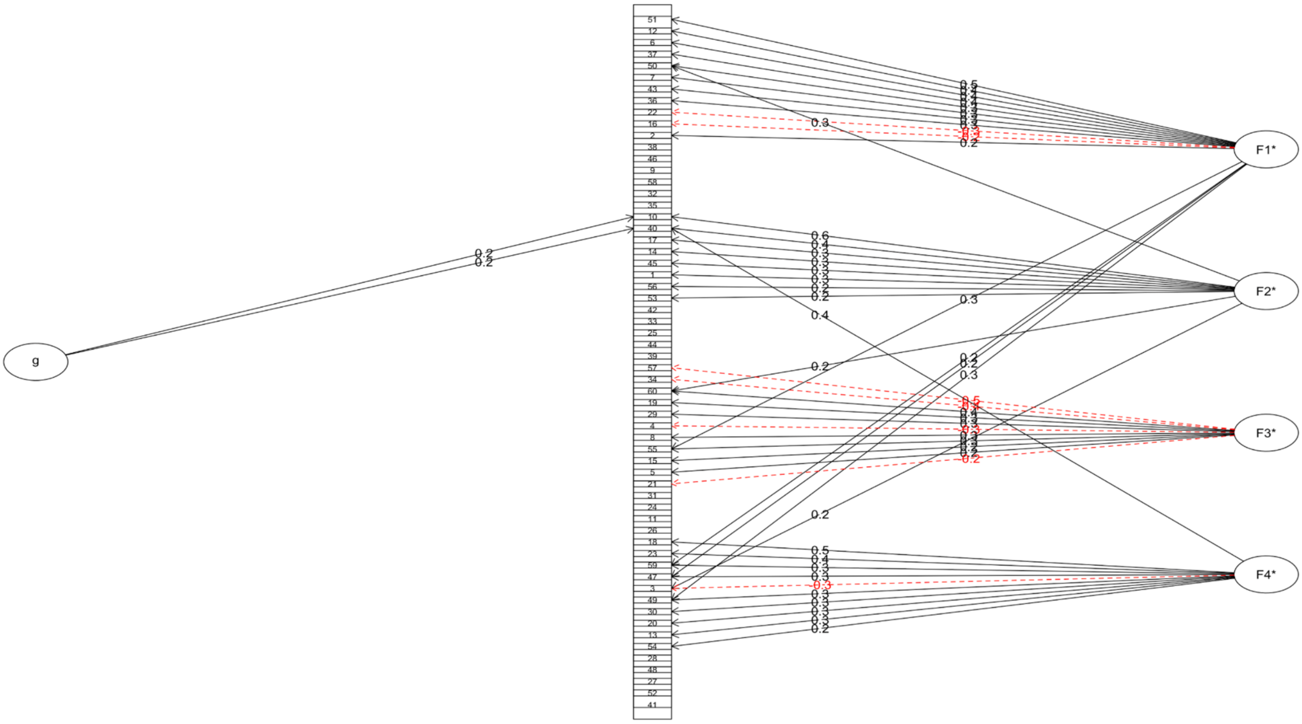

Fig. 3 Factor structure of the sad dwell time exceeds neutral dwell time metric. Only two of the 60 trials load onto a general factor, suggesting it is not appropriate to combine trial level data into a single outcome

use of sum scores with depression assessments for similar reasons (Fried et al., 2016). Indeed, recent work has begun to emphasize the importance of checking the psychometric properties of behavioral task data (Parsons et al., 2019).

An especially important insight from these analyses is the necessity of considering other psychometrics beyond internal consistency coefficients. For instance, when considering the trial-level difference score in the sad/neutral task, even with acceptable α and ω total values (.76 and .80 respectively), only roughly 15% of the trials loaded onto a general factor. For the percentage of trials where sad dwell time > neutral dwell time metric, α and ω total indicate internal consistency values that some researchers might be willing to tolerate (alpha = .58, omega total = .62). However, learning that only two trials loaded onto a general factor would strongly suggest that this metric is unsuitable for this task (see Fig. 3 for a visual depiction).

The mismatch between the internal consistency metrics and indicators of unidimensionality is, at first glance, puzzling. This is an important reminder that while internal consistency metrics give estimations of how well correlated different halves of the data are with each other, they do not give us any indication of how well individual items (in this case, trials) are mapping onto a general factor. Indices like coefficient α are actually better described as measures of internal consistency reliability, and thus are sensitive to longer test length (Tang et al., 2014). Therefore, metrics with many items (e.g. trials) might have better internal consistency reliability, but poor unidimensionality or homogeneity. This same principle highlights the need to look across multiple indices, as ω_h and PUC can theoretically be influenced by a larger number of items (Rodriguez et al., 2016b). Therefore, researchers should consider evaluating the factor structure prior to examining internal consistency metrics (Green & Yang, 2015).

In light of these results, we would caution anyone programming stimuli-based tasks from assuming all trials are interchangeable. In particular, we discourage researchers from assuming trial uniformity and programming their task to select trials randomly, with replacement. As we have illustrated, not all stimuli presentations may load onto a general factor, and if a non-loading (or negative-loading) item is repeatedly selected for presentation, the psychometrics of that task data will be driven down even further. Another consequence of this design is that participants would receive different versions of the task from each other. If they complete multiple testing sessions across the course of the study, the version of the task will also likely differ. This makes it impossible to test the factor structure at a trial level, and consequently, there is no assurance that responses to the stimuli are consistent across the various versions of the task. We typically would not randomly select items from a self-report scale to be given to an individual; we would want to administer the same items across participants to ensure we are eliciting the same information from individuals. Fortunately, we can test the factor structure of the task stimuli for a given metric, but this cannot be done when the stimuli configurations across trials differ across participants (unless it is done intentionally and a random effect of stimuli is incorporated into the statistical models, which is also rarely done).

As a result, we propose several recommendations for future task development. First, this study highlights the importance of estimating and reporting the psychometric properties of each metric one hopes to use (we hope our open code will help to facilitate this: https://dataverse.tdl.org/dataverse/factor_structure_attention). After all, psychometricians have long stressed that reliability is not inherent to a measure or task but rather depends on the specific context in which a measure is used, as well as the specific approaches to deriving a metric (Armstrong et al., 2021). Multiple metrics can be derived from the same data from a given task (in this case, of attentional bias), but not all metrics are equal; simply because one metric shows good psychometric properties does not guarantee that all will. Researchers should be judicious about the metrics they select and ensure these elements have adequate psychometric properties before attempting to investigate individual differences, treatment effects, etc.

Second, trials and stimuli configurations are not all interchangeable, and evaluation of the factor structure of task metrics is still needed to confirm whether the items load onto a general factor. We suggest conducting an initial pilot study with many items included in the behavioral task being employed. After assessing the factor structure of the metrics of interest, the pool can be edited to use only those trials that hang together and produce a streamlined version of the task. These trials can be repeated (but not with replacement) to build out the ideal number of trials for the task. This reflects a data-driven and iterative process toward task development, which has not been widely applied, yet is generally standard practice for questionnaire development.

There are several important limitations to the present study. First, the average age of our participants was 19.4 years old; future research will also need to examine whether similar results are achieved in a sample with a more diverse age range. Second, given that this study was focused on the psychometrics of the task, we did not specifically recruit a clinical sample and used a sample of undergraduate college students. In future work, we hope to collect additional data in a clinical sample that would be better positioned to test the cognitive theory of depression. Nevertheless, there is ample evidence that depression exists along a continuum (Gibb et al., 2004; Hankin et al., 2005) and that non-clinical samples can provide useful initial evidence, particularly in the early stages within domains of research. Given that this is the first attempt to examine the unidimensionality of a cognitive bias task, we believe it is important to complete this proof-of-principal test in a convenient sample before moving on to the time and expense of collecting data from clinical samples. We believe this work also provides a nice foundation with which to compare future work on the psychometrics of this task in the clinical samples.

Additionally, the stimuli database we used only included faces of Caucasian actors, which is not uncommon for affective stimuli sets. While stimuli sets with racial diversity exist, there are very few that also include varying emotional expressions, which is necessary for affective attention bias tasks. Indeed, the only example of a stimuli set that included both racial diversity and varying emotional expressions was the NimStim stimuli database (Tottenham et al., 2009), but there were not enough actor photographs to generate the number of 4×4 matrices we needed while meeting the other specifications (e.g. each actor appears only once during the matrix, even split of genders, etc.). Certainly, if we are to improve our measurement of attention bias (and further, try to improve attention bias modification as a treatment), representative stimuli sets will need to be incorporated into our behavioral tasks. Finally, we only tested five metrics in each version of the task, and only one had excellent psychometric properties. Countless metrics could be extracted from this time series data, and each one's psychometrics will need to be evaluated before progressing to an evaluation of its utility as a marker of negative attention bias.

In addition, these analyses employed exploratory factor analytic techniques to identify factor structure. Given the study's relatively small sample size, we were unable to conduct a confirmatory factor analysis in a hold out (or separate) sample to formally evaluate the fit of the proposed factor structures for each metric. Nonetheless, we believe this preliminary work lays an important blueprint for examining the psychometrics of behavioral tasks in a more rigorous manner. We also recognize the application of these standards sets a high bar for evaluating behavioral tasks, especially given that some of these indices of unidimensionality are not yet routinely implemented even for questionnaires. Still, we believe a high bar is necessary if we are to overcome the plague of psychometric– and replication– issues afflicting the field.

In conclusion, the measurement of attentional bias for positive and negative stimuli and its association with psychopathology has been fraught with inconsistency, in part, we believe, because many of the tasks have poor or unknown psychometrics. We rigorously studied a free-viewing eye tracking task by carefully selecting emotion stimuli, checking the internal consistency of several attention metrics, and determining whether these metrics were drawn from the same underlying construct (i.e., a unidimensional measurement model), which is necessary for computing a bias score across all items. Total dwell time (measured separately for happy vs. neutral and sad vs. neutral facial expressions) in this free-viewing task appears to be a psychometrically sound way to assess attention bias in studies of psychopathology. The methods described here provide a new and important approach for examining the measurement properties of behavioral tasks that measure cognitive processes central to the maintenance of psychopathology. Many commonly used bias score metrics sum or average across trials with the assumption that they are interchangeable, which requires unidimensional data. As internal consistency reliability does not equate to unidimensionality, unidimensionality should also be examined for bias score metrics.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability The data that support the findings of this study are available in our Supplementary Materials within the Texas Data Repository at https://dataverse.tdl.org/dataverse/factor_structure_attention.

Declarations

Conflict of Interest Mary McNamara, Kean Hsu, Bryan McSpadden, and Semeon Risom report no conflicts of interest. Christopher Beevers has received grant funding from the National Institutes of Health, Brain and Behavior Foundation, and other not-for profit foundations. He has received income from the Association for Psychological Science for his editorial work and from Orexo, Inc. for serving on a Scientific Advisory Board related to digital therapeutics. Dr. Beevers' financial disclosures have been reviewed and approved by the University of Texas at Austin in accordance with its conflict-of-interest policies. Jason Shumake has received grant funding from the National Institutes of Health as well as salary and stock options from Aiberry, Inc. Drs. Beevers' and Shumake's financial disclosures have been reviewed and approved by the University of Texas at Austin in accordance with its conflict of interest policies.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review, 32*(8), 704–723.

Armstrong, T., Wilbanks, D., Leong, D., & Hsu, K. (2021). Beyond vernacular: Measurement solutions to the lexical fallacy in disgust research. *Journal of Anxiety Disorders, 82*, 102408.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *The American Journal of Psychiatry, 165*(8), 969–977.

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression Inventory–II. *Psychological Assessment*. https://doi.org/10.1037/t00742-000

Beevers, C. G., Lee, H.-J., Wells, T. T., Ellis, A. J., & Telch, M. J. (2011). Association of predeployment gaze bias for emotion stimuli with later symptoms of PTSD and depression in soldiers deployed in Iraq. *The American Journal of Psychiatry, 168*(7), 735–741.

Beevers, C. G., Mullarkey, M. C., Dainer-Best, J., Stewart, R. A., Labrada, J., Allen, J. J. B., McGeary, J. E., & Shumake, J. (2019). Association between negative cognitive bias and depression: A symptom-level approach. *Journal of Abnormal Psychology, 128*(3), 212–227.

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. In *Educational and Psychological Measurement* (Vol. 65, pp. 676–696).

Bodenschatz, C. M., Kersting, A., & Suslow, T. (2019). Effects of briefly presented masked emotional facial expressions on gaze behavior: An eye-tracking study. *Psychological Reports, 122*(4), 1432–1448.

Canty, A., & Ripley, B. D. (2019). *boot: Bootstrap R (S-Plus) Functions*.

Chong, L. J., & Meyer, A. (2021). Psychometric properties of threat-related attentional bias in young children using eye-tracking. *Developmental Psychobiology, 63*(5), 1120–1131.

Disner, S. G., Beevers, C. G., Haigh, E. A. P., & Beck, A. T. (2011). Neural mechanisms of the cognitive model of depression. *Nature Reviews. Neuroscience, 12*(8), 467–477.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412.

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.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods, 42*(1), 351–362.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using r to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time … Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28*(11), 1354–1367.

Gibb, B. E., Alloy, L. B., Abramson, L. Y., Beevers, C. G., & Miller, I. W. (2004). Cognitive vulnerability to depression: A taxometric analysis. *Journal of Abnormal Psychology, 113*(1), 81–89.

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement Issues and Practice, 34*(4), 14–20.

Hankin, B. L., Fraley, R. C., Lahey, B. B., & Waldman, I. D. (2005). Is depression best viewed as a continuum or discrete category? A taxometric analysis of childhood and adolescent depression in a population-based sample. *Journal of Abnormal Psychology, 114*(1), 96–110.

Harrell, F. E., Jr, from Charles Dupont, W. C., & others., M. (2020). *Hmisc: Harrell Miscellaneous*. Retrieved from https://CRAN.R-project.org/package=Hmisc

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186.

Hsu, K. J., McNamara, M. E., Shumake, J., Stewart, R. A., Labrada, J., Alario, A., Gonzalez, G. D. S., Schnyer, D. M., & Beevers, C. G. (2020). Neurocognitive predictors of self-reported reward responsivity and approach motivation in depression: A data-driven approach. *Depression and Anxiety, 37*(7), 682–697.

Hsu, K. J., Shumake, J., Caffey, K., Risom, S., Labrada, J., Smits, J. A. J., Schnyer, D. M., & Beevers, C. G. (2021). Efficacy of attention bias modification training for depressed adults: A randomized clinical trial. *Psychological Medicine*, 1–9.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Iannone, R., Cheng, J., & Schloerke, B. (2020). *gt: Easily Create Presentation-Ready Display Tables*. Retrieved from https://CRAN.R-project.org/package=gt

Kappenman, E. S., Farrens, J. L., Luck, S. J., & Proudfit, G. H. (2014). Behavioral and ERP measures of attentional bias to threat in the dot-probe task: Poor reliability and lack of correlation with anxiety. *Frontiers in Psychology, 5*, 1368.

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*(11), 1238–1243.

Klawohn, J., Bruchnak, A., Burani, K., Meyer, A., Lazarov, A., Bar-Haim, Y., & Hajcak, G. (2020). Aberrant attentional bias to sad faces in depression and the role of stressful life events: Evidence from an eye-tracking paradigm. *Behaviour Research and Therapy, 135*, 103762.

Lanza, C., Müller, C., & Riepe, M. W. (2018). Positive mood on negative self-statements: Paradoxical intervention in geriatric patients with major depressive disorder. *Aging & Mental Health, 22*(6), 748–754.

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.

Lazarov, A., Ben-Zion, Z., Shamai, D., Pine, D. S., & Bar-Haim, Y. (2018). Free viewing of sad and happy faces in depression: A potential target for attention bias modification. *Journal of Affective Disorders, 238*, 94–100.

Liu, Y., Ding, Y., Lu, L., & Chen, X. (2017). Attention bias of avoidant individuals to attachment emotion pictures. *Scientific Reports, 7*, 41631.

Machulska, A., Kleinke, K., & Klucken, T. (2022). Same same, but different: A psychometric examination of three frequently used experimental tasks for cognitive bias assessment in a sample of healthy young adults. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01804-9

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324.

McNamara, M. E., Shumake, J., Stewart, R. A., Labrada, J., Alario, A., Allen, J. J. B., Palmer, R., Schnyer, D. M., McGeary, J. E., & Beevers, C. G. (2021). Multifactorial prediction of depression diagnosis and symptom dimensions. *Psychiatry Research, 298*, 113805.

Miles, W. R. (1930). Ocular dominance in human adults. *The Journal of General Psychology, 3*(3), 412–430.

Owens, M., & Gibb, B. E. (2017). Brooding rumination and attentional biases in currently non-depressed individuals: An eye-tracking study. *Cognition & Emotion, 31*(5), 1062–1069.

Parsons, S. (2020). *splithalf; robust estimates of split half reliability*. Retrieved from https://doi.org/10.6084/m9.figshare.5559175.v5

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395.

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*(2), 365.

Price, R. B., Rosen, D., Siegle, G. J., Ladouceur, C. D., Tang, K., Allen, K. B., Ryan, N. D., Dahl, R. E., Forbes, E. E., & Silk, J. S. (2016). From anxious youth to depressed adolescents: Prospective prediction of 2-year depression symptoms via attentional bias measures. *Journal of Abnormal Psychology, 125*(2), 267–278.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5–26.

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Robinson, D., Hayes, A., & Couch, S. (2020). *broom: Convert Statistical Objects into Tidy Tibbles*. Retrieved from https://CRAN.R-project.org/package=broom

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

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150.

Sanchez, A., Vazquez, C., Marker, C., LeMoult, J., & Joormann, J. (2013). Attentional disengagement predicts stress recovery in depression: An eye-tracking study. *Journal of Abnormal Psychology, 122*(2), 303–313.

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality, 19*(7), 595–605.

Sears, C. R., Quigley, L., Fernandez, A., Newman, K., & Dobson, K. (2019). The reliability of attentional biases for emotional images measured using a free-viewing eye-tracking paradigm. *Behavior Research Methods, 51*(6), 2748–2760.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the Model Size Effect on SEM Fit Indices. *Educational and Psychological Measurement, 79*(2), 310–334.

Skinner, I. W., Hübscher, M., Moseley, G. L., Lee, H., Wand, B. M., Traeger, A. C., Gustin, S. M., & McAuley, J. H. (2018). The reliability of eyetracking to assess attentional bias to threatening words in healthy individuals. *Behavior Research Methods, 50*(5), 1778–1792.

Soltani, S., Newman, K., Quigley, L., Fernandez, A., Dobson, K., & Sears, C. (2015). Temporal changes in attention to sad and happy faces distinguish currently and remitted depressed individuals from never depressed individuals. *Psychiatry Research, 230*(2), 454–463.

Staugaard, S. R. (2009). *Reliability of two versions of the dot-probe task using photographic faces*. Citeseer. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.6498&rep=rep1&type=pdf

Steiner, M. D., & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software, 5*(53), 2521. https://doi.org/10.21105/joss.02521

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., Jones, P. B., & Perez, J. (2020). On Dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment*, 1073191120976863.

Suslow, T., Hußlack, A., Kersting, A., & Bodenschatz, C. M. (2020). Attentional biases to emotional information in clinical depression:

A systematic and meta-analytic review of eye tracking findings. *Journal of Affective Disorders, 274*, 632–642.

Tang, W., Cui, Y., & Babenko, O. (2014). Internal consistency: Do we really know what it is and how to assess it. *Journal of Psychology & Clinical Psychiatry., 2*(2), 205–220.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

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

Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. *Revista Brasileira de Psiquiatria (Sao Paulo, Brazil : 1999), 35*(4), 416–431.

Wells, T. T., Clerkin, E. M., Ellis, A. J., & Beevers, C. G. (2014). Effect of antidepressant medication use on emotional information processing in major depression. *The American Journal of Psychiatry, 171*(2), 195–200.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2019). *readxl: Read Excel Files*. Retrieved from https://CRAN.R-project.org/package=readxl

Wickham, H., & Seidel, D. (2020). *scales: Scale Functions for Visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Xie, Y. (2014). knitr: a comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research.* Chapman and Hall/CRC.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.