JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY, 2016
VOL. 38, NO. 4, 416–433
http://dx.doi.org/10.1080/13803395.2015.1121969

Routledge
Taylor & Francis Group

OPEN ACCESS

Test–retest reliability and effects of repeated testing and satiety on performance of an Emotional Test Battery

Jason Michael Thomas[a], Suzanne Higgs[a] and Colin Trevor Dourish[b]

[a]School of Psychology, University of Birmingham, Birmingham, UK; [b]P1vital, Manor House, Howbery Park, Wallingford, UK

ABSTRACT

The P1vital® Oxford Emotional Test Battery (ETB) comprises five computerized tasks designed to assess cognition and emotional processing in human participants. It has been used in between-subjects experimental designs; however, it is unclear whether the battery can be used in crossover designs. This is of particular importance given the increasing use of ETB tasks for repeated assessment of depressed patients in clinical trials and clinical practice. In addition, although satiety state has been reported to affect performance on some cognitive and emotional tasks, it is not known whether it can influence performance on the ETB. Two studies explored these issues. In Experiment 1, 30 healthy women were tested on the ETB on 4 separate occasions (each a week apart) in a within-subjects design. In Experiment 2, another 30 healthy women were randomized to either a satiated or a hungry condition, where they were given an ad libitum lunch of cheese sandwiches, before (satiated) or after (hungry) they were asked to complete the ETB. Experiment 1 demonstrated good test–retest reliability for the ETB. One of the tasks was free from practice effects, whilst performance on the other four tasks stabilized after the first two sessions. In Experiment 2, eating to satiety only affected performance on a single ETB task. These results suggest that the ETB can be used in crossover designs after two initial training sessions. Further, as a robust satiety manipulation had only a limited effect on a single ETB task, it is unlikely that appetitive state will confound ETB performance.

ARTICLE HISTORY
Received 4 September 2015
Accepted 16 November 2015

KEYWORDS
Emotional Test Battery; ETB; practice effects; satiety; crossover design

Computerized test batteries have been used extensively to investigate the effects of behavioral and pharmacological interventions on cognitive function. For example, the P1vital® Oxford Emotional Test Battery (ETB, e.g., Murphy, Downham, Cowen, & Harmer, 2008) has been used to detect early effects of antidepressant drugs on cognitive–emotional functioning and has been validated over a number years (e.g., Harmer et al., 2003; Harmer et al., 2010; Harmer, Shelley, Cowen, & Goodwin, 2004; Horder, Cowen, Di Simplicio, Browning, & Harmer, 2009) in healthy volunteers (Harmer, Bhagwagar, Cowen, & Goodwin, 2002) and in patients with depression (Browning et al., 2015; Harmer et al., 2009; Post et al., 2014).

The ETB (see www.p1vital.com) comprises five validated cognitive tests that can be used to assess cognition and emotional processing (e.g., Murphy et al., 2008). The Facial Expression Recognition Task (FERT) displays faces that participants must categorize into one of six emotional categories based on their expression: happiness; fear; anger; disgust; sadness; surprise; and neutral (250 trials in total). The primary measure for this task is response bias, which measures the tendency to respond more or less to one stimulus than another by taking into account the number of false alarms (when participants incorrectly respond that a stimulus is present) and misses (when participants incorrectly respond that a stimulus is not present). Response accuracy and reaction times can also be calculated to examine potential speed–accuracy trade-off.

The Faces Dot Probe Task (FDOT) involves the presentation of two faces, which are replaced by a

CONTACT Jason Michael Thomas thomasjm@bham.ac.uk School of Psychology, University of Birmingham, Birmingham, B15 2TT, UK.

This research was conducted at School of Psychology, University of Birmingham.
Colin Dourish is an employee and shareholder of P1vital Limited, and Suzanne Higgs is a member of P1vital's Advisory Panel and Jason Michael Thomas is funded by the Steve Cooper P1vital–BBSRC PhD Studentship.

pair of dots (192 trials in total). On some trials, one of the faces has an emotional expression (happy versus fearful). Participants must report the orientation of the pair of dots (i.e., vertical versus horizontal) for each trial. For this task a vigilance score is calculated as the primary measure. This is a measure of sustained attention for a given stimulus and is derived by subtracting the reaction times from congruent trials (trials where the probe appears in the same location as the stimulus) from incongruent trials (trials where the probe appears in a different location from the stimulus). Accuracy and reaction times can also be calculated to examine potential speed–accuracy trade-off.

The Emotional Categorization Task (ECAT) displays 30 positive and 30 negative self-referent personality descriptors (e.g., “cheerful” versus “hostile,” respectively) that participants must respond to, indicating whether they would like or dislike to be referred to as such. Reaction time is the primary measure for this task; accuracy is also examined for speed–accuracy trade-off. In the Emotional Recall Task (EREC) participants are asked to recall as many words as they can remember from the ECAT (out of the total 60 words). This element is partly computerized: instructions given via computer, but words written down using pen and paper. The number of words correctly recalled during this task is the primary measure for the EREC, though recall of incorrect words can also be examined.

Finally, in the Emotional Recognition Memory Task (EMEM) words are re-presented from the ECAT (60 old words), along with new distractor words (60 novel words), and participants are asked to report whether they have previously seen the word. For this task, response bias (see above) is calculated as the primary measure; accuracy and reaction times are also examined for speed–accuracy trade-off. Across all four sessions, for each task, the same fixed set of stimuli (faces and words) is used for each test session.

The majority of previous ETB studies have used a between-subjects design in which participants were tested in a single session only. A between-subjects design avoids issues with repeated exposure to stimuli such as practice effects or other factors that could result in changes in baseline levels of responding, such as variation in the test setting and motivation of the participants to engage with the tasks (Kane & Kay, 1992). However, in experimental settings there are advantages of using within-subjects designs to assess the effect of interventions because of their greater power to detect significant effects and the reduction in error variance associated with individual differences. In addition, computerized tests including some or all of component tasks of the ETB are increasingly being used in clinical settings to assess drug efficacy, and there often is a need to assess changes in performance over time in individual patients (Browning et al., 2015; Goldberg, Keefe, Goldman, Robinson, & Harvey, 2010; Post et al., 2014).

The use of multiple stimulus sets or alternate test forms across test sessions can overcome some of the issues associated with repeated testing because participants are unable to learn responses to specific stimuli, but this does not address changes in performance over time due to procedural learning (Roebuck-Spencer, Sun, Cernich, Farmer, & Bleiberg, 2007). Another useful approach to examine whether the rate of change in performance in an experimental group differs from that in a control or reference group is test–retest variability or measurement error (Jacobson & Truax, 1991). This can identify the variability over time that is expected by chance or due to other factors such as practice. Such approaches can also be used to compare the performance of individuals to that of a group, for example to assess whether a patient is responding to treatment (Chelune, 2002). However, an issue with this approach is that a reference group may not be well matched on individual difference variables that affect the degree of learning or practice on the tasks. In this case, an effect attributed to an intervention may be better explained by preexisting differences in the rate of change between groups (Wesnes & Pincock, 2002). One way of minimizing these issues is to assess normative change when performance has plateaued, and test–retest reliability is stable.

The test–retest reliability of specific tests has been evaluated, and a meta-analysis of practice effects for a range of neuropsychological tests revealed substantial practice effects for many tasks although the size of the effects were dependent on factors such as the age of the participants and the length of the retest interval (Calamia, Markon, & Tranel, 2013). Moreover, an examination of the reliability of the dot-probe attentional task suggested that performance was neither internally consistent nor stable in a nonclinical sample of participants (Schmukle, 2005). These data underscore the importance of assessing the reliability of specific cognitive tests (Heilbronner et al., 2010). To date there has been no examination of test–retest reliability or how many sessions are required for performance on the ETB tasks to stabilize, although previous work suggests that practice effects on other cognitive tasks are minimized after 2–3 sessions (Collie, Maruff, Darby, & McStephen, 2003). It has been recommended that four prestudy training sessions in psychopharmacology should be adopted as a standard procedure (McClelland, 1987). Hence, the aim of Experiment 1 was to assess the test–retest reliability and stability of performance on ETB measures over four test sessions. Such information is needed if learning effects are to be precluded from clinical studies where accurate baseline measures of cognitive performance are required. In addition, such data add to the body of knowledge on practice effects for cognitive tasks assessing different domains of function.

Another methodological issue that arises when testing the effects of an intervention on cognitive function is the extent to which hunger and satiety should be controlled for prior to test. It known that ingestion of specific macronutrients can affect performance on some cognitive tasks (Dye, Lluch, & Blundell, 2000) and that consumption or omission of a meal immediately prior to test can also affect cognitive performance (Gibson & Green 2002). For example, negative effects on cognition, particularly attention, have been reported after consumption of a large lunch (Smith, Ralph, & McNeill, 1991). Consuming breakfast is reported to improve cognitive performance on memory tasks under some circumstances (Benton & Parker, 1998) but not others (Smith, Kendrick, Maben, & Salmon, 1994). The extent to which performance on the ETB is affected by hunger is also unknown. Investigating this issue in relation to specific cognitive test batteries is important because it provides researchers with information on whether performance may be affected by recent food consumption. Hence the aim of Experiment 2 was to investigate the effect of consuming a standard lunch to satiety on ETB measures.

Experiment 1

Method

Participants

Thirty healthy women student volunteers (mean age = 18.9 years; mean body mass index, BMI = 21.5; mean National Adult Reading Test, NART, score = 111) were recruited for the study from the University of Birmingham. Informed consent was obtained, and participants were given either £20 cash or course credits upon completion. The study was approved by the University of Birmingham Research Ethics Committee and was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. Participants were excluded from the study if they were under 18 or over 65 years of age and if they were not fluent English speakers. Using a screening questionnaire, participants were excluded if they: had previously taken part in an ETB study; were dyslexic; were smokers; were taking medication; had consumed a high amount of caffeine (>750 mg; Winston, Hardwick, & Jaberi, 2005) or alcohol (>3 units; National Institute for Health and Care Excellence, NICE, 2010) in the last 24 hours; or had current or past depression, determined by using the questions for assessing depression only, from the Structured Clinical Interview for DSM–IV Axis I Disorders (SCID; Spitzer, Williams, Gibbon, & First, 2004; DSM–IV - *Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition*; American Psychiatric Association, 1994).

Design

A within-subjects design was used, with a single factor of session composed of four levels: Session 1, Session 2, Session 3, and Session 4. Each session was run at the same time of day, one week apart, and participants completed the ETB during all four sessions. The order of completing questionnaires and the ETB during sessions was counterbalanced across participants; half of the participants always completed the questionnaires followed by the ETB, while the other half were tested in the reverse order each time.

Procedure

Participants completed a consent form before completing the screening measures. They had their height and weight measured for BMI calculation then completed: the NART (Nelson, 1982) as an estimate of verbal IQ; the SCID (questions relating to depression only); a lifestyle questionnaire (including questions about age, gender, medical conditions, smoker status, etc.); and an alcohol and caffeine questionnaire (documenting intake during the last 24 hours). Participants were then given visual analogue scales (VAS) with the following mood and appetite items to rate on a scale from 0–100 mm (0 mm anchor = not at all, 100 mm anchor = extremely): "alertness"; "disgust"; "drowsiness"; "light-headed"; "anxiety"; "happiness"; "nausea"; "sadness"; "withdrawn"; "faint"; "hungry"; "full"; "desire to eat"; and "thirst." After this, participants completed the ETB (which took approximately 60 minutes) and then the Three Factor Eating Questionnaire (TFEQ; Stunkard & Messick, 1985) and the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) in a counterbalanced order. Finally, participants completed another VAS questionnaire.

Participants returned for three further sessions, which were seven days apart from one another, and always at the same time of day. The procedure above was repeated for each session with the exception of consent, BMI measurement, NART, SCID, and the lifestyle questionnaire. On completing their last session, participants were debriefed, thanked for their time, and compensated with either £20 cash or course credits.

Data analysis

General. Within-subjects analysis of variance (ANOVA) was used to analyze the data. Bonferroni correction was used for all post hoc *t* tests, and violations of sphericity were addressed using the Greenhouse–Geisser correction.

VAS. To establish a factor structure for the VAS, a principal components analysis (PCA) was run with varimax rotation. Analysis of the 14 items provided four factors with eigenvalues >1, accounting for 66.64% of the variance. Items that loaded >0.5 onto a factor were included, resulting in four factors of three or more items: Appetite (desire to eat, hungry, fullness, and thirst); Negative Physical Effects (faint, lightheaded, and nausea); Arousal (alertness, happiness, and drowsiness); Negative Mood (anxiety, sadness, and disgust). Withdrawn did not load >0.5 onto any of the factors and was analyzed separately. Scores for each of the factors were calculated by summing the scores for all items in that factor and then dividing by the number of items. Items with a negative scale were inverted to match the other items.

ETB data. Effects of session are reported first, followed by task-specific effects that were relevant to the task but not to the experimental manipulation. These are presented to confirm the ability to detect effects of emotion and or valence. Main effects and interactions (Session × Valence/Emotion) were followed with *t* tests to further analyze the data. For sessions, comparisons consisted of Sessions 1 versus 2, 2 versus 3, and 3 versus 4.

Intraclass correlation coefficients. To examine test–retest reliability for ETB task measures, intraclass correlation coefficients (ICCs) were calculated using a two-way mixed-effects model for absolute level of agreement. ICCs were calculated between Sessions 1 to 2, 2 to 3, and 3 to 4 for the primary measures of interest for the ETB tasks (split by emotion): FERT response bias; ECAT reaction times; EREC correct word recall; and EMEM response bias. ICCs were not conducted on FDOT vigilance scores as healthy participants do not show an emotional bias on this task, hence it would not be expected that this measure would be reliable over time. Instead, accuracy and reaction times were examined for reliability. Across measures, an ICC less than .40 was considered poor test–retest reliability, .40–.75 adequate, and .75 or greater was considered good to very good (Weintraub et al., 2014).

Results

Questionnaire data

BDI scores were in the low range (mean = 6.8, *SE* = 1.2), alcohol consumption prior to testing was low (mean = 0.04 units, *SE* = 0.02), and caffeine consumption was well within the defined study limit (mean = 187.2 mg, *SE* = 20.5). ANOVA comparing these measures across the four test sessions did not show any significant differences (all $p > .05$). For the TFEQ measures, cognitive restraint, disinhibition, and hunger scores were all in the normal range (mean = 7.2, *SE* = 1.2; mean = 6.5, *SE* = 0.6; mean = 7.4, *SE* = 0.7) and did not differ significantly between sessions (all $p > .05$). Analysis of VAS ratings revealed that there were no effects of session, time, or interaction between these factors for the following (all $p > .05$); Appetite (mean = 44.8, *SE* = 1.6); Negative Physical Effects (mean = 5.9, *SE* = 1.6); Negative Mood (mean = 8.1, *SE* = 1.6); Withdrawn (mean = 7.9, *SE* = 2.0); however, for arousal there was a main effect of session, $F(3, 87) = 3.12$, $p < .05$. Bonferroni corrected *t* tests comparing sessions were not significant, though the closest to significance was the decrease in arousal from Session 1 to Session 3, $t(29) = 2.70$, $p = .07$ (Session 1 mean = 64.1, $SE = 2.7$; Session 2 mean = 57.6, $SE = 2.8$; Session 3 mean = 57.0, $SE = 2.9$; Session 4 mean = 59.3, $SE = 3.1$). There was no effect of time or a significant interaction for this measure (both $p > .05$).

ETB data

For reaction time measures, only data for correct responses were used. All data were examined for outliers (±3 *SD*s from the mean), resulting in the removal of 1.1% of the total ETB data set.

Intraclass correlation coefficients (ICCs)

Average ICC scores across all four sessions ranged from .4–.8 for 16 out of the 17 measures (94%), indicating adequate test–retest reliability for the majority of measures (Table 1). The only exception was the FDOT accuracy score for positive words, which displayed an average ICC of .3, indicating poor test–retest reliability.

Facial expression recognition task (FERT). Repeated measures ANOVA with session (4 levels: 1, 2, 3, and 4) and emotion (7 levels: anger, disgust, fear, happy, neutral, sad, and surprise) as factors revealed that for response bias there was no effect of session, $F(3, 72) = 1.25, p > .05$ (Figure 1), but there was an effect of emotion, $F(4, 86) = 105.06, p < .001$, and an interaction approaching significance, $F(5, 114) = 2.28$, $p = .05$ (Figure 1). Breaking down the interaction by emotion, there was a main effect of session for anger, neutral, and surprise (all $p < .05$), but not for disgust, fear, happy, and sad (all $p > .05$). Examining the effect of session for anger, neutral, and surprise, Bonferroni corrected *t* tests showed a significant increase in response bias to anger expressions from Session 1 to Session 2 (.63 versus .71); $t(29) = 2.905, p < .05$ (Figure 1). There were no other significant effects for any other emotions.

For accuracy, there were main effects of session, $F(3, 78) = 5.65$, $p < .01$ (Figure 2) and emotion, $F(3, 79) = 16.85$, $p < .01$, but no significant interaction ($p > .05$) (Figure 2). Bonferroni corrected *t* tests on the effect of session revealed that accuracy increased from Session 1 to Session 2 (55.7% versus 58.2%); $t(27) = -2.86, p < .05$, but did not differ significantly between Sessions 2 to 3 and 3 to 4 (both $p > .05$). Following up the effect of emotion, accuracy in categorizing anger (45.4%), disgust (53.8%), fear (51.6%), sadness (53.7%), and surprise (59.9%) was lower than that for neutral faces (70.8%; all $p < .01$), while accuracy for happy faces (69.2%) was not significantly different from accuracy for neutral faces ($p > .05$).

Table 1. Intraclass correlation coefficients for ETB tasks split by emotion over sessions.

	ICC			
Task and measure	Session 1–Session 2	Session 2–Session 3	Session 3–Session 4	Average ICC
FERT response bias				
Anger	.6***	.7***	.8***	.7
Disgust	.6***	.7***	.8***	.7
Fear	.4**	.8***	.7***	.6
Happy	.4*	.4*	.5**	.4
Neutral	.5**	.6***	.6***	.6
Sad	.8***	.7***	.8***	.8
Surprise	.7***	.8***	.8***	.8
FDOT accuracy				
Positive[a]	.4*	.3	.5**	.4
Negative[a]	.6***	.3	.1	.3
FDOT reaction times				
Positive	.5***	.7***	.8***	.7
Negative	.6***	.6***	.8***	.7
ECAT reaction times				
Positive	.7***	.8***	.7***	.7
Negative	.6***	.7***	.8***	.7
EREC correct words				
Positive[a]	.2*	.7***	.7***	.5
Negative	.5***	.5***	.5**	.5
EMEM response bias				
Positive	.5**	.5**	.6***	.6
Negative[a]	.4**	.2	.4*	.4

Note. ICC = intraclass correlation coefficient; ETB = Emotional Test Battery; FERT = Facial Expression Recognition Task; FDOT = Faces Dot Probe Task; ECAT = Emotional Categorization Task; EREC = Emotional Recall Task; EMEM = Emotional Recognition Memory Task.

[a]Measures with ICC < .4.

*$p < .05$. **$p < .01$. ***$p < .001$.

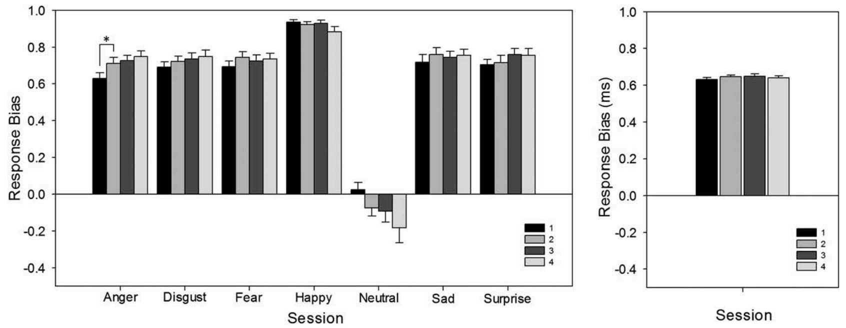

Figure 1. Facial Expression Recognition Task (FERT): response bias, split by emotion and test session (left), and split by session only (right). To the presentation of anger expressions only, response bias increased from Session 1 to Session 2. Error bars represent standard error of the mean. *$p < .05$.

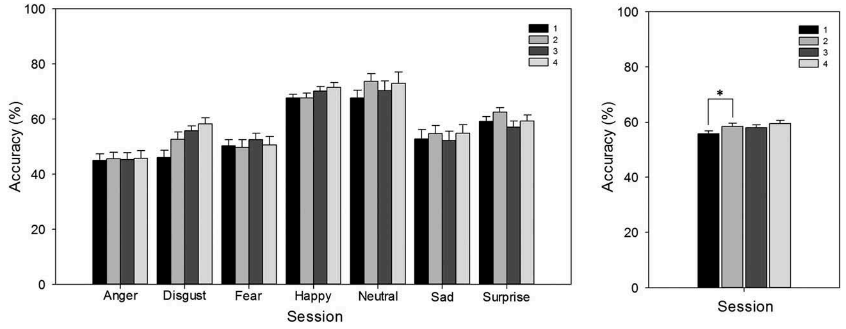

Figure 2. Facial Expression Recognition Task (FERT): accuracy, split by emotion and test session (left), and split by session only (right). There was an overall effect of session, whereby accuracy increased from Session 1 to Session 2. Error bars represent standard error of the mean. *$p < .05$.

For reaction time there were main effects of session, $F(3, 69) = 28.53$, $p < .001$ (Figure 3) and emotion, $F(3, 80) = 27.91$, $p < .001$, but no significant interaction ($p > .05$) (Figure 3). Reaction times significantly decreased between Sessions 1 and 2 (1331.6 ms versus 1239.3 ms), $t(25) = 3.63$, $p < .01$, and 2 and 3 (1242.9 ms versus 1164.1 ms), $t(27) = 3.46$, $p < .01$, but not between Sessions 3 and 4 ($p > .05$). For the effect of emotion, reaction times to expressions of anger (1322.1 ms), disgust (1205.6 ms), fear (1452.2 ms), sadness (1184.2 ms), and surprise (1241.3 ms) were significantly slower than those to neutral faces (1049.7 ms; all $p < .01$), while reaction times to happy faces (1055.0 ms) and neutral faces did not differ ($p > .05$).

Faces dot probe task (FDOT). Repeated measures ANOVA with session (4 levels: 1, 2, 3, and 4), emotion (2 levels: fear and happy), and masking (2 levels: masked and unmasked) as factors revealed

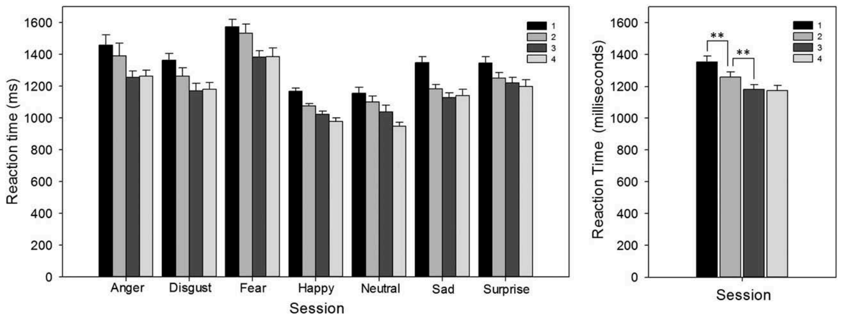

Figure 3. Facial Expression Recognition Task (FERT): reaction times, split by emotion and test session (left), and split by session only (right). There was an overall effect of session, whereby reaction time decreased from Session 1 to Session 2 and from Session 2 to Session 3. Error bars represent standard error of the mean. **$p < .01$.

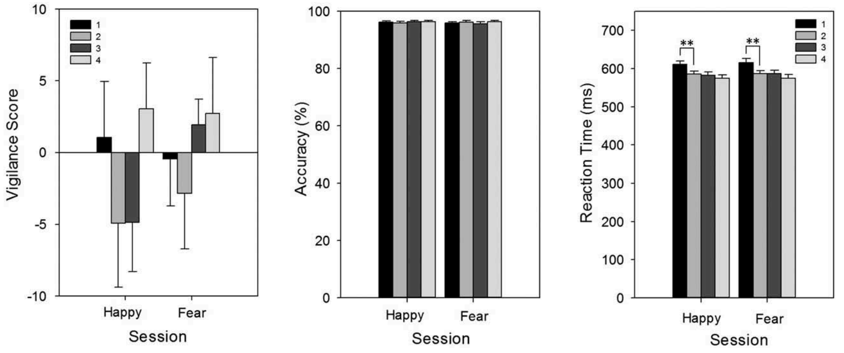

Figure 4. Faces Dot Probe Task (FDOT): vigilance score (left), accuracy (center), and reaction times (right) to happy and fearful expressions for the four test sessions. Reaction times to both happy and fearful faces decreased significantly from Session 1 to Session 2. Error bars represent standard error of the mean. $^{**}p < .01$.

that for vigilance scores, there was no main effect of session, $F(3, 78) = 1.13$, $p > .05$ (see Figure 4), emotion, $F(1, 26) = 0.74$, $p > .05$, or mask, $F(1, 26) = 0.05$, $p > .05$, nor any significant interactions (all $p > .05$). The same repeated measures ANOVA was used for accuracy and reaction times; however, the factor of congruence was added (2 levels: congruent and incongruent). For accuracy, there was a main effect of masking on accuracy (masked faces = 96.7% versus unmasked faces = 96.1%); $F(1, 25) = 4.31$, $p < .05$, but no effect of session (see Figure 4), emotion (fear versus happy), or congruence (congruent versus incongruent probe location), nor any interactions (all $p > .05$). For reaction time, there was a main effect of session, $F(2, 56) = 10.86$, $p < .001$, an interaction between emotion and session, $F(3, 75) = 3.95$, $p < .05$, and a four-way interaction between masking, emotion, congruence, and session, $F(3, 75) = 2.76$, $p < .05$. Breaking down the four-way interaction by emotion, there were main effects of session for reaction times to both fearful and happy expressions, $F(3, 78) = 10.62$, $p < .001$; $F(2, 61) = 10.52$, $p < .001$, but no other main effects or significant interactions (all $p > .05$). Bonferroni corrected paired t tests showed that response times reduced from Sessions 1 to 2 for both emotions (happy, Session 1 = 610.6 ms vs. Session 2 = 581.0 ms, $p < .01$; fear, Session 1 = 614.7 ms vs. Session 2 = 587.0 ms, $p < .01$; see Figure 4). There was also a trend for reaction times to fearful faces to decrease between Sessions 3 and 4 (583.6 vs. 571.5, $p = .06$).

Emotional categorization task (ECAT). Repeated measures ANOVA with session (4 levels: 1, 2, 3, and 4) and valence (2 levels: positive and negative) as factors revealed that for reaction times there was no effect of session (Figure 5), valence, or an

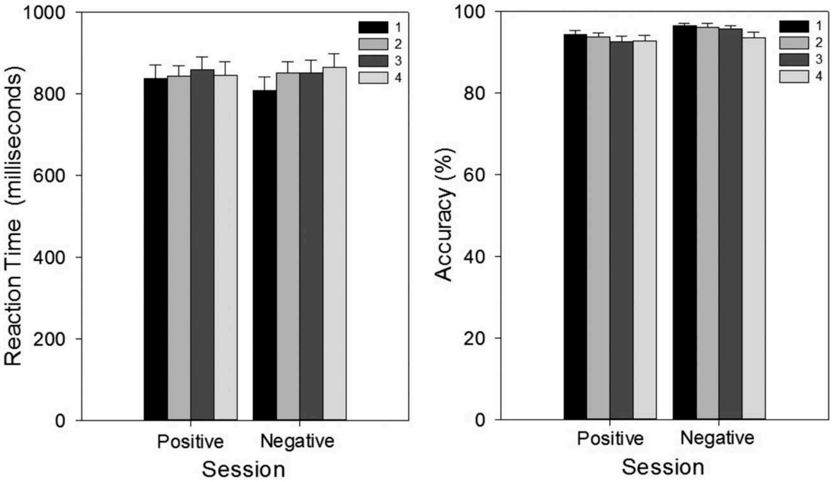

Figure 5. Emotional Categorization Task (ECAT): reaction times (left) and accuracy (right) to positive and negative words for the four test sessions. Error bars represent standard error of the mean.

interaction between session and valence (all $p > .05$). For accuracy there was an effect of session, $F(2, 57) = 3.53$, $p < .05$; however, Bonferroni corrected paired t tests comparing sessions (1 versus 2; 2 versus 3; and 3 versus 4) were not significant (all $p > .05$; see Figure 5). The nearest to significance was the comparison between Sessions 3 and 4 (94.3% versus 93.1%, respectively, $p = .7$). There was also an effect of valence on accuracy, whereby negative words were categorized more accurately than positive words (mean = 95.6%, $SE = 0.7$ vs. mean = 93.5%, $SE = 1.1$); $F(1, 25) = 6.76$, $p = .07$. There was no significant interaction between valence and session ($p > .05$).

Emotional recall task (EREC). Repeated measures ANOVA with session (4 levels: 1, 2, 3, and 4) and valence (2 levels: positive and negative) as factors revealed a main effect of session on the number of words correctly recalled, $F(3, 84) = 46.12$, $p < .001$. Bonferroni corrected t tests showed that accuracy increased from Session 1 to Session 2 and Session 2 to Session 3 (both $p < .001$; Figure 6), but did not change between Sessions 3 and 4 ($p > .05$). There was also a main effect of valence for the number of words correctly recalled (negative words = 8.1 versus positive words = 9.8), $F(1, 28) = 15.70$, $p < .001$, but no significant interaction between valence and session, $F(3, 84) = 1.88$, $p > .05$.

For the number of incorrectly recalled words, there was a main effect of session, $F(3, 81) = 8.59$, $p < .001$, a main effect of valence, $F(1, 27) = 13.62$, $p < .01$, and an interaction between session and valence, $F(3, 81) = 6.59$, $p < .001$. Breaking down the interaction by valence, there was no effect of session for incorrectly recalled negative words, $F(3, 84) = 0.56$, $p > .05$, but there was an effect of session for incorrectly recalled positive words, $F(3, 84) = 13.13$, $p < .001$. Bonferroni corrected t tests showed significant decreases in positive words falsely recalled from Session 1 to Session 2, $t(29) = 2.71$, $p < .05$, and Session 2 to Session 3, $t(28) = 2.64$, $p < .05$, but no difference between Sessions 3 and 4, $t(28) = 1.22$, $p > .05$ (see Figure 6).

Emotional recognition memory task (EMEM). Repeated measures ANOVA with session (4 levels: 1, 2, 3, and 4) and valence (2 levels: positive and negative) as factors revealed that for response bias there was no effect of session, $F(3, 84) = 1.24$, $p = .3$ (Figure 7), but there was a main effect of valence whereby participants showed a greater response bias to negative words than to positive (.37 versus −.14), $F(1, 28) = 140.99$, $p < .001$. There was no interaction between valence and session ($p > .05$). For accuracy there was no effect of session, $F(3, 84) = 0.22$, $p > .05$ (Figure 7), but there was a main effect of valence whereby positive words were recalled more accurately than negative (mean = 83.8%, $SE = 1.5$ vs. mean = 68.7%, $SE = 2.1$), $F(1, 28) = 79.45$, $p < .001$. There was no interaction between valence and session ($p > .05$). For reaction time, there was a main effect of session, $F(2, 59) = 4.51$, $p < .05$. Follow-up t tests (Bonferroni corrected) showed that reaction times significantly decreased between Sessions 1 and 2, $t(27) = 3.75$, $p < .01$ (Figure 7); however, there were no significant differences between Sessions 2 and 3, or 3 and

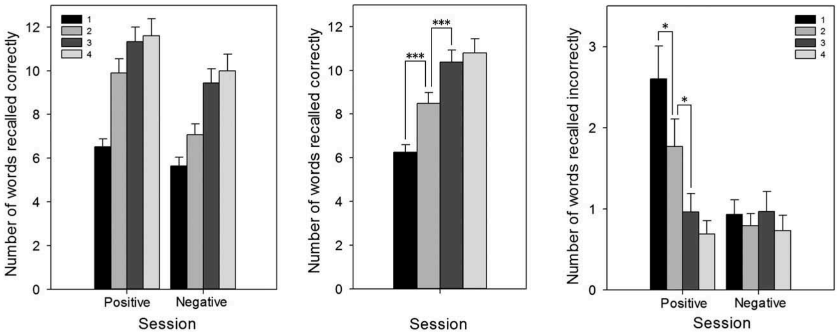

Figure 6. Emotional Recall Task (EREC): Correctly recalled words split by valence and session (left), and split by session only (center), and incorrectly recalled words split by valence and session (right). Number of words correctly recalled increased from Sessions 1 to 2 and 2 to 3, but not 3 to 4. For positive words incorrectly recalled, there was a significant decrease from Sessions 1 to 2 and 2 to 3, but again, no change between Sessions 3 and 4. Error bars represent standard error of the mean. *$p < .05$. ***$p < .01$.

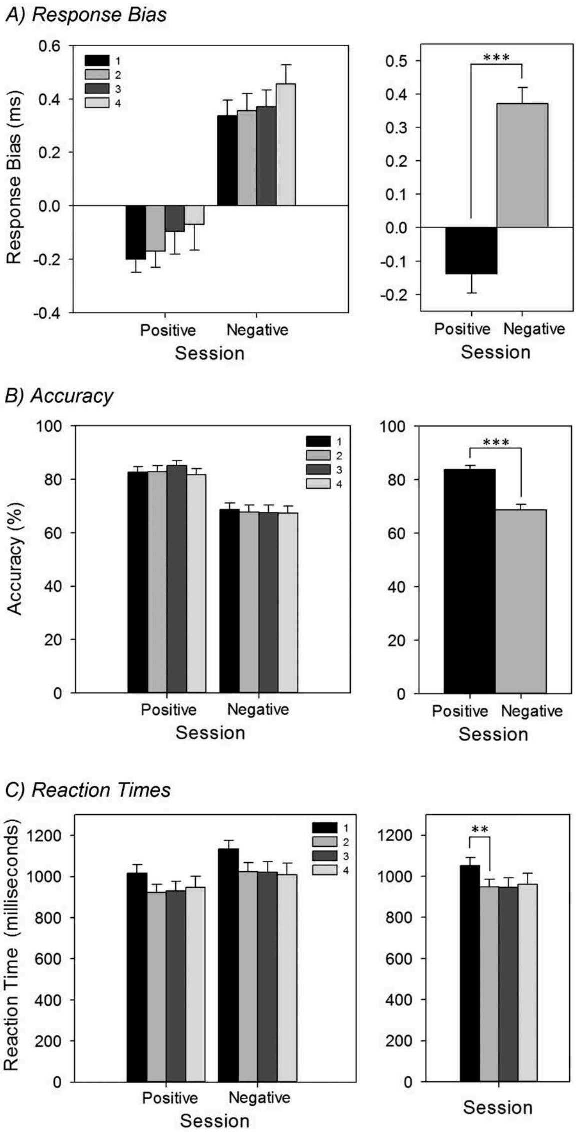

Figure 7. Emotional Recognition Memory Task (EMEM): (A) Response bias split by valence and session (top left) and valence only (top right); (B) accuracy split by valence and session (middle left) and valence only (middle right); (C) reaction times split by valence and session (bottom left) and session only (bottom right). There was a significant response bias towards negative words compared to positive words (but no main effect of session, $p = .3$); positive words were recognized with greater accuracy than negative words; and reaction times significantly decreased between the first and second sessions. Error bars represent standard error of the mean. $^{**}p < .01$. $^{***}p < .001$.

4 (both $p > .05$). An effect of valence was also noted for reaction time whereby responses were quicker to positive words than to negative words (mean = 929.3 ms, $SE = 39.9$ vs. mean = 1022.1 ms, $SE = 43.0$), $F(1, 26) = 52.89$, $p < .001$. There was no interaction between valence and session ($p > .05$).

Discussion

We report the investigation of the effects of test–retest reliability and repeated testing on performance for each of the ETB tasks. The majority of ETB measures demonstrate adequate test–retest reliability, and performance stabilizes after two test sessions, suggesting that the ETB can be used for repeated testing after a run in of two practice sessions.

The validity of using the ETB in repeated measures designs rests on the assumption of reliable test–retest results over sessions. Here we confirm that test–retest reliability scores for the majority of the ETB measures were adequate, with many tasks yielding ICCs of .7 or .8. These data are comparable with the results of a recent meta-analysis reporting the mean test–retest reliability of a range of cognitive tasks to be around .7 or higher (Calamia et al., 2013). Of the four measures showing poor test–retest reliability, FDOT accuracy scores (positive and negative) were particularly unreliable; however, this is comparable to previous work reporting a lack of internal consistency and stability in nonclinical samples with this task (Schmukle, 2005). Reliability for the other two measures (EREC correct positive words and EMEM negative response bias) reached adequate reliability for the final two sessions (.4 and .7, respectively), hence with the exception of the FDOT, all measures exhibit reasonable reliability after the first two sessions.

For the primary measures of interest we also assessed practice effects. For the FERT task, response bias to disgust, fear, happy, sad, surprise, and neutral emotions did not change over time. However, response bias to angry expressions increased from the first session to the second session, which is consistent with evidence of a sensitization to angry facial expressions with repeated exposure (Strauss et al., 2005). However, there were no further changes between Sessions 2, 3, and 4, suggesting that these practice effects are limited to the first session only. FDOT vigilance scores did not change significantly over time; however, there was no emotional bias on this task in the healthy volunteers tested in this study. Without a bias towards one emotion over the other, vigilance scores would not be expected to be consistent over time, but to vary considerably. This was the case, as indicated by the large standard errors. Together, these data reinforce the unreliability of this task with nonclinical participants (Schmukle, 2005).

For the ECAT the primary measure was reaction time, and this did not change with repeated testing. This may be due to the low cognitive demand of the task and the ease of accessing self-referent stimuli—that is, there was no capacity for practice to improve performance. Evidence suggests that self-referent stimuli are processed automatically and faster than non-self-referent stimuli (Bargh, 1982; Geller & Shaver, 1976). In addition, there was no difference in reaction times to positive or negative words and no interaction between session and valence. Thus this measure appears to be resistant to practice effects, across all sessions and valence.

Practice effects were observed with the EREC for both positive and negative correct words, but only for positive incorrect words. The comparatively higher rate of false intrusions of positive (vs. negative) incorrect words during the first two sessions might suggest an initial positive bias that is blunted by practice. Regarding the practice effects on this task more generally, the words recalled in the emotional recall task were the same for each session. Hence, the large practice effects likely reflect both familiarity with the task procedure and familiarity with the items to be recalled. These issues could be addressed at least in part by the use of alternative stimulus sets for each test session. However, while the use of alternative stimuli reduces practice effects in some studies, the evidence remains inconsistent and is likely to be task specific and therefore requires specific testing (Benedict & Zgaljardic, 1998; Hinton-Bayre & Geffen, 2005).

For the EMEM task, no practice effects were observed for response bias. There was a significant difference in response to positive and negative words; however, this did not interact with session. Thus, like the ECAT task, the EMEM task appears to be resistant to practice effects, across all sessions and valence.

For all but one task there was an acceleration of reaction time with repeated testing, but for the last two sessions responding stabilized for all tasks. This pattern of results is consistent with findings from other studies of practice effects on cognitive test batteries (e.g., Falleti, Maruff, Collie, & Darby, 2006). This probably reflects the effects of familiarity with the task procedures on reaction time since there was no speed–accuracy trade-off for any task that might indicate a change in response strategy over time. Accuracy only improved with repeated testing for the FERT and the EREC. The FERT requires participants to categorize unfamiliar faces according to their emotional expression, and hence increased familiarity may have improved categorization accuracy on this task.

One consideration is whether the results observed in this study are comparable with observations in previous ETB studies. Compared to the results from Experiment 1 (data from the first test session in parentheses), healthy volunteers in previous ETB studies showed the following accuracy on the FERT: 48% (45%) to anger, 50% (54%) to disgust, 52% (52%) to fear, 62% (69%) to happy, 51% (54%) to sad, 68% (71%) to neutral, and 58% (60%) to surprise (Harmer et al., 2003; Harmer, Heinzen, O'Sullivan, Ayres, & Cowen, 2008; Harmer et al., 2004). Hence, the accuracy levels for each emotion observed in this study are comparable with those reported in previously published research. In addition, previous work has shown that healthy populations exhibit a positive emotional bias when responding on the ETB (Schmidt et al., 2015). This was the case with the FERT and EMEM tasks, whereby participants were significantly quicker and more accurate when presented with positive stimuli than with negative stimuli. Hence, these data replicate well established effects with the ETB.

The present results suggest that overall performance on the ETB tasks is stable after two sessions and that the ETB could be used for repeated test sessions with the inclusion of two practice sessions. However, an issue might be whether after two practice sessions, there is reduced sensitivity to detect significant effects of an experimental manipulation due to the induction of a rigid response set or floor or ceiling effects. Ceiling effects were likely observed for the EREC after two sessions because the number of items correctly recalled was 12, which may be at the limit of memory. The use of an alternative response set as previously discussed would address this issue. For the EMEM and FERT, stable performance was at levels where both increases and decreases in performance are likely to be detectable. Together, the results suggest good reliability and limited practice effects, which are potentially important findings for the use of ETB tasks in repeated assessment of depressed patients in clinical studies and clinical practice. In particular, the test–retest reliability and absence of practice effects for the FERT response bias measure are very encouraging, given its recent use in the early assessment of antidepressant response in a primary care study (Browning et al., 2015).

Based on these findings we would suggest that ETB researchers should consider two practice sessions when using the battery in future studies that have within-subjects designs to increase the reliability of the results. The absence of practice sessions could create uncertainty as to whether data may be subject to practice effects, possibly creating Type 1 or Type 2 errors.

Experiment 2

Method

Participants

Thirty healthy women psychology students (mean age = 21.4 years; mean BMI = 20.0; mean NART = 117) were recruited from the University of Birmingham. Informed consent was obtained from all participants, who were compensated after the study with either course credits or £10 cash. The study was approved by the University of Birmingham Research Ethics Committee and was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. Exclusion criteria from Experiment 1 also applied to Experiment 2 (age range, fluency in English, prior ETB study participation, dyslexia and smoker status, medication use, caffeine and alcohol consumption, and depression). In addition, participants had to possess a BMI between 18.5 and 24.9, have no food allergies or diabetes, and score less than 10 on the restraint scale of the TFEQ to be recruited. This is because high levels of dietary restraint have been associated with impaired cognitive performance (Green, Rogers, Elliman, & Gatenby, 1994). Participants were also excluded from taking part if they had participated in Experiment 1; hence, none of the subjects included in Experiment 2 had taken part in Experiment 1.

Design

A between-subjects design with a single factor (satiety state) and two levels (satiated versus hungry) was used. Participants were randomly allocated to a condition with 15 participants in each group. Previous work has shown that 12–16 participants per group yielded significant effects on the ETB (Browning, Reid, Cowen, Harmer, & Goodwin, 2007; Harmer et al; 2004; Murphy et al., 2008). Similarly, Benton and colleagues (Benton & Parker, 1998) reported significant effects on memory with a fed versus fasted manipulation with approximately 16–17 participants per group, while Smith and colleagues (1991) reported significant effects on attention comparing fed and overfed groups of 12 and 11 participants, respectively. Hence, 15 participants per group appears adequate to detect an effect in this type of paradigm. Based on prior research indicating that mood effects can be reliably detected 60 min after food consumption (Macht & Dettmer, 2006; Smith, Leekam, Ralph, & McNeill, 1988), participants were tested on the ETB 60 minutes after consuming lunch or in a hungry state.

Cheese sandwich lunch

For lunch, participants were served a platter of cheese sandwiches: 16 quarters, arranged in two rows of eight quarters each. Each quarter sandwich serving contained 92.3 calories and weighed approximately 31 g. Participants were provided with a plate to eat from and were asked to eat as much as they wanted until they felt comfortably full. The platter was weighed before and after serving (along with any remnants left on the participant's plate) to determine total food intake in grams. Participants were also provided with a glass of water.

Procedure

Prior to attending the test session, participants were sent the TFEQ via email to ensure they were eligible for the study. Those who attended the test day (between 12 pm and 2 pm) were screened with a lifestyle questionnaire, a breakfast questionnaire (to ensure they had not consumed food since 8 pm the previous day), the SCID (questions relating to depression only), and the NART. Participants also completed an alcohol and caffeine screening questionnaire to assess their intake over the last 24 hours, before completing a set of VAS. VAS items were placed above the center of a 100-mm line, anchored with “not at all” (0 mm) and “extremely” (100 mm), and included the items: alert; disgusted; drowsy; light-headed; anxious; happy; nauseated; sad; withdrawn; faint; hungry; thirsty; full; and desire to eat.

Participants in the satiated condition were served a cheese sandwich lunch, after which they completed another VAS and a sandwich rating questionnaire. This questionnaire assessed liking of the sandwich, whether the meal was a typical size, and whether participants ate beyond comfortable fullness, using VAS scale items. Participants were then asked to wait in a test cubicle for an hour before administration of the ETB test; as noted above, mood effects have previously been detected an hour after eating. During this time they completed a VAS after 30 minutes and 60 minutes, the latter immediately prior to ETB testing. Participants were then asked to complete the ETB tasks, followed by a batch of questionnaires, including the Power of Food Scale (PFS; Lowe et al., 2009) as a measure of appetitive anticipation, the Barratt Impulsivity Scale (BIS 11; Patton, Stanford, & Barratt, 1995) as a measure of impulsive behavior, and the BDI to assess depression and mood. Participants then had their height and weight measured for calculation of BMI, were asked what they thought the aims of the study were, and were debriefed and thanked for their time. Participants in the hungry condition completed a similar procedure (also waiting an hour before testing on the ETB), but consumed the lunch of cheese sandwiches after completing the ETB tasks.

Data analysis

General. Between-subjects and mixed ANOVAs were used to analyze main effects of satiety state and interactions. Bonferroni correction was used for all post hoc t tests, and violations of sphericity were addressed using the Greenhouse–Geisser correction.

VAS. The factor structure derived from Experiment 1 was applied to the VAS data from Experiment 2.

ETB data. As with Experiment 1, effects of the manipulations are presented first, followed by task-specific effects (e.g., effects of emotion, or valence).

Results

Participant characteristics and subjective state questionnaires

Mean values for participant characteristics and subjective state questionnaires, split by hungry and satiated groups, are displayed in Table 2. Participants were young, with healthy BMI scores and good verbal IQs (NART). They were within the normal range of impulsiveness (BIS 11) and appetitive anticipation (PFS), and showed low scores on the BDI, indicating normal mood. Their TFEQ scores were within the low–normal range, and the mean amount of food consumed was within expectations for a lunch. Using independent t tests (hungry versus satiated) no significant differences were observed for any measure (all $p > .05$).

Table 2. Participant characteristics and subjective state questionnaires from Experiment 2.

	Condition	
Measure	Hungry	Satiated
Age (years)	19.7 (0.3)	20.3 (0.5)
Body Mass Index (BMI)	21.5 (0.6)	21.4 (0.5)
National Adult Reading Test (NART)	116.3 (1.1)	117.1 (1.3)
Barratt Impulsivity Scale (BIS)	63.3 (2.0)	68.2 (3.0)
Power of Food Scale (PFS)	38.2 (2.4)	37.4 (3.1)
Beck Depression Inventory (BDI)	5.8 (0.9)	7.8 (1.5)
TFEQ		
Cognitive Restraint	6.2 (0.8)	6.3 (0.8)
Disinhibition	5.3 (0.7)	7.1 (1.0)
Hunger	5.4 (1.0)	7.3 (0.9)
Amount eaten (grams)	193.6 (16.7)	188.5 (15.5)

Note. TFEQ = Three Factor Eating Questionnaire. Standard error of the mean in parentheses.

Visual analogue scales

VAS scores were entered into mixed ANOVAs with the factor of satiety state (satiated versus hungry) and time (pre versus post manipulation). For appetite there was a main effect of satiety state and time, and a significant interaction between satiety state and time (all $p < .001$). Comparing pre- versus post-manipulation ratings separately for each group, appetite significantly decreased over time in the satiated group ($p < .001$), but not in the hungry group ($p > .05$; see Table 3). For arousal there was a main effect of time ($p < .05$), whereby arousal decreased slightly (63.6 mm to 58.3 mm), but there was no effect of satiety state or a significant interaction (both $p > .05$). For negative physical effects, there was no effect of satiety state or time (both $p > .05$), but there was a trend for an interaction between satiety state and time ($p = .07$); however, follow-up t tests did not reveal any significant effects (both $p > .05$). For negative mood and withdrawn, there were no effects of satiety state or time, or a significant interaction between satiety state and time (all $p > .05$).

ETB data

For reaction time measures, only data for correct responses were used. All data were examined for outliers (±3 standard deviations from the mean), resulting in the removal of 1.1% of the total ETB data set.

Table 3. Visual Analogue Scale mean scores split by satiety state and time.

	Hungry		Satiated	
VAS item	Pre-manipulation	Post-manipulation	Pre-manipulation	Post-manipulation
Appetite[a],[b,c]	74.3 (3.8)	76.7 (4.0)	77.3 (3.8)	21.5 (4.0)
Arousal[b]	64.0 (4.2)	55.6 (4.2)	63.1 (4.2)	61.0 (4.2)
Negative Physical Effects	15.8 (4.1)	18.9 (4.1)	15.2 (4.1)	6.7 (4.1)
Negative Mood	11.8 (2.5)	8.6 (2.0)	6.2 (2.5)	4.8 (2.0)
Withdrawn	17.2 (4.7)	18.6 (4.2)	13.3 (4.7)	9.5 (4.2)

Note. VAS = Visual Analogue Scale. Standard error of the mean in parentheses.
[a]Main effect of satiety state. [b]Main effect of time. [c]Interaction between satiety state and time.

Facial expression recognition task (FERT). A mixed ANOVA with satiety state (2 levels: satiated and hungry) and emotion (7 levels: anger, disgust, fear, happy, neutral, sad, and surprise) as factors revealed that for response bias there was no effect of satiety state (satiated = 0.62, hungry = 0.64), $F(1, 28) = 0.45$, $p > .05$, an effect of emotion, $F(2, 59) = 125.03$, $p < .001$, and no significant interaction, $F(6, 168) = 0.52$, $p > .05$ (Figure 8). Bonferroni corrected t tests on the main effect of emotion showed that participants were significantly biased towards anger (0.75), disgust (0.76), fear (0.76), happy (0.94), sad (0.69), and surprise (0.74) faces, compared to neutral (–0.23; all $p < .001$).

For accuracy, there was no effect of satiety state ($p > .05$), a main effect of emotion, $F(3, 91) = 29.45$, $p < .001$, and no interaction ($p > .05$; see Figure 9). Bonferroni corrected t tests on the effect of emotion showed that the accuracy for each emotion (anger = 46.0%, disgust = 54.8%, fear = 46.7%, happy = 61.8%, sad = 46.8%, and surprise = 58.0%) was significantly lower than that for neutral facial expressions (78.3%; all $p < .01$). Analysis of reaction time data also revealed no effect of satiety state ($p > .05$), a main effect of emotion, $F(6, 156) = 21.41$, $p < .001$, and no interaction between emotion and satiety state ($p > .05$; see Figure 9). For the effect of emotion, reaction times to expressions of anger (1504.8 ms), disgust (1300.2 ms), fear (1614.5 ms), sadness (1414.6 ms), and surprise (1387.5 ms) were significantly slower than those to neutral faces (1124.6 ms; all $p < .01$), while reaction times to happy faces (1179.6 ms) were not significantly different from those to neutral faces ($p > .05$).

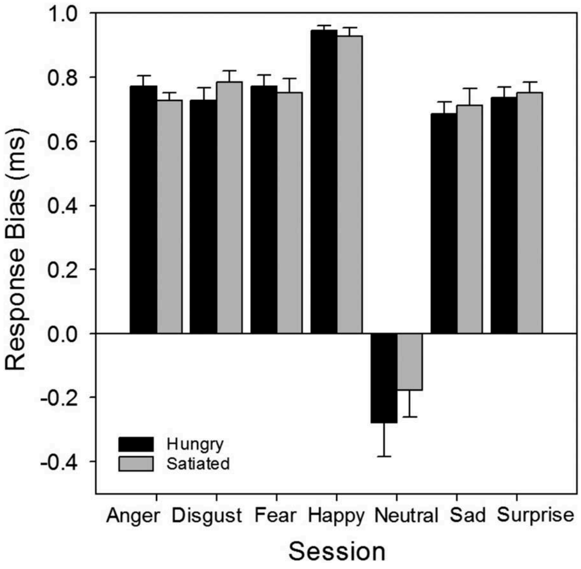

Figure 8. Facial Expression Recognition Task (FERT): response bias, split by satiety state and emotion. Error bars represent standard error of the mean.

Faces Dot Probe Task (FDOT). A mixed ANOVA with satiety state (2 levels: satiated and hungry), emotion (2 levels: fear and happy), and masking (2 levels: masked and unmasked) revealed that for vigilance scores there was no main effect of satiety state [hungry = –7.07 (SE = 4.27), satiated = 1.59 (SE = 4.41); $F(1, 27) = 1.99$, $p > .05$], emotion [fear = –3.85 (SE = 3.88), happy = –1.63 (SE = 5.03); $F(1, 27) = 0.12$, $p > .05$], or mask [masked = –3.32 (SE = 3.80), unmasked = –2.16 (SE = 5.03); $F(1, 27) = 0.03$, $p > .05$], nor any significant interactions (all $p > .05$; see Table 4). The same mixed ANOVA was used for accuracy and reaction times; however, the factor of congruence was added (2 levels: congruent and incongruent). For both measures, there was no main effect of satiety state (hungry versus satiated; see Table 4), emotion (fear versus happy faces), masking (masked versus unmasked), or congruency (congruent versus incongruent probe location) and no significant interactions between these factors (all $p > .05$).

Emotional categorization task (ECAT). A mixed ANOVA with satiety state (2 levels: satiated and hungry) and valence (2 levels: positive and negative) showed there was no effect of satiety state or valence, nor an interaction between satiety state and valence (positive versus negative words) for

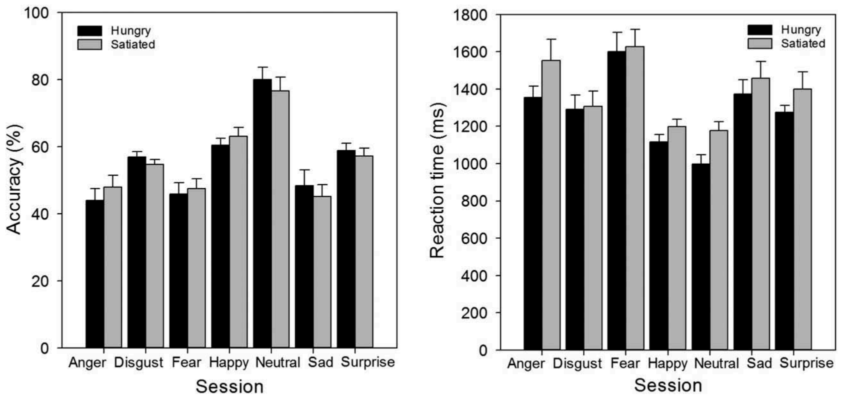

Figure 9. Facial Expression Recognition Task (FERT): accuracy (left) and reaction times (right) split by satiety state and emotion. Error bars represent standard error of the mean.

ECAT accuracy (all $p > .05$; see Table 4). Analysis of ECAT reaction time showed no effect of satiety state ($p > .05$), a trend towards a main effect of valence with quicker times for positive versus negative words, $F(1, 28) = 4.16$, $p = .05$, and no interaction ($p > .05$).

Emotional recall task (EREC). A mixed ANOVA with satiety state (2 levels: satiated and hungry) and valence (2 levels: positive and negative) revealed that for words correctly recalled, there was no effect of satiety state ($p > .05$), a main effect of valence with more positive than negative words recalled, $F(1, 28) = 54.24$, $p < .001$ (see Table 4), and no significant interaction ($p > .05$). For words incorrectly recalled, there was also no effect of satiety state ($p > .05$), an effect of valence with more positive words recalled versus negative, $F(1, 28) = 15.97$, $p < .001$ (see Table 4), and no significant interaction ($p > .05$).

Emotional recognition memory task (EMEM). A mixed ANOVA with satiety state (2 levels: satiated and hungry) and valence (2 levels: positive and negative) showed that for response bias, there was an effect of satiety state, $F(1, 28) = 10.25$, $p < .01$, an effect of valence, $F(1, 28) = 64.02$, $p < .001$, and a significant interaction, $F(1, 28) = 5.59$, $p < .05$ (see Table 4). Breaking down the interaction by emotion, response bias to the positive words was significantly lower in satiated than in hungry individuals (−0.34 versus 0.12), $t(28) = 3.24$, $p < .01$. There was no significant difference in response bias between satiated and hungry individuals to the negative words (0.35 versus 0.49), $t(28) = 1.78$, $p > .05$. Accuracy scores showed no effect of satiety state ($p > .05$), a main effect of valence with better accuracy for positive than for negative words, $F(1, 27) = 59.97$, $p < .001$ (see Table 4), and no significant interaction ($p > .05$). Analysis of reaction time also showed no effect of satiety state ($p > .05$), an effect of valence with quicker times

Table 4. Vigilance score, response bias, accuracy, reaction times, and number of correct and incorrect words recalled for ETB tasks, split by negative and positive stimuli and hungry and satiated states.

		Negative		Positive	
ETB task	Measure	Hungry	Satiated	Hungry	Satiated
Faces Dot Probe (FDOT)	Vigilance score	−8.63 (5.4)	0.93 (5.6)	−5.50 (7.0)	2.25 (7.2)
	Accuracy (%)	95.7 (1.0)	94.8 (1.0)	95.2 (1.0)	94.9 (1.1)
	Reaction time (ms)	630.8 (14.8)	642.1 (15.3)	631.9 (15.9)	643.4 (16.4)
Emotional Categorization (ECAT)	Accuracy (%)	96.7 (1.0)	97.4 (1.0)	97.4 (1.0)	95.0 (1.0)
	Reaction time (ms)	834.7 (41.7)	819.1 (41.7)	785.2 (37.5)	805.0 (37.5)
Emotional Recall (EREC)	Correct words[b]	5.1 (0.7)	4.7 (0.7)	7.2 (0.7)	7.0 (0.7)
	Incorrect words[b]	0.6 (0.2)	0.5 (0.2)	1.7 (0.4)	2.1 (0.4)
Emotional Recognition Memory (EMEM)	Response bias[a,b,c]	0.49 (0.1)	0.35 (0.1)	0.12 (0.1)	−0.34 (0.1)
	Accuracy (%)[b]	65.3 (3.3)	66.9 (3.5)	79.8 (2.8)	85.0 (2.8)
	Reaction time (ms)[b]	1081.3 (62.5)	1093.1 (62.5)	915.7 (44.0)	912.1 (44.0)

Note. ETB = Emotional Test Battery. Standard error of the mean in parentheses.
[a]Main effect of satiety state ($p < .01$). [b]Main effect of valence ($p < .001$). [c]Interaction between satiety state and valence ($p < .05$).

for positive than for negative words, $F(1, 28) = 54.24$, $p < .001$ (see Table 4), and no significant interaction ($p > .05$).

Discussion

We report the first investigation of eating to satiety on performance for each of the ETB tasks. Eating to satiety has only limited effects on ETB task performance, affecting EMEM response bias only. These data suggest that a robust satiety manipulation has very limited effects on ETB performance, and therefore satiety state is unlikely to be a significant confound in ETB studies.

Participants who were asked to eat a sandwich lunch until satiated reported a decrease in appetite, compared to participants who were not given lunch. Satiation did not significantly affect questionnaire based measures of mood; however, it significantly reduced response bias on the EMEM task to positive, but not negative, words. This is particularly interesting as the initial categorization of these words on the ECAT task was not affected by satiety state, nor was free recall performance on the EREC, suggesting that the effect is specific to recognition memory. While there is evidence that the consumption of food can decrease positive emotional responses (Smith et al., 1991) and enhance recognition memory for words (Smith et al., 1994), there has been no investigation of how satiety affects emotional biases within recognition memory. Hence, this appears to be the first evidence to suggest that satiation may blunt a positive bias in emotional recognition memory. Therefore, in studies where EMEM performance is an outcome variable of interest, monitoring hunger may be a prudent course of action.

It is possible that the lack of wider effects of satiety on the ETB is related to the food used in this study. For instance, a study by Macht and Dettmer (2006) reported that both apple and chocolate consumption elevated mood in healthy women, but the effect of chocolate consumption was greater than the effect of apple consumption. Hence, it is possible that highly palatable or energy-dense foods have greater effects on mood than less palatable or less energy-dense foods. This suggestion is supported by evidence that foods with a high energy content have greater effects on mood than food with a lower energy content (Macht, Gerer, & Ellgring, 2003). Thus, the use of a food that is more palatable or energy dense than bland cheese sandwiches may have elicited greater effects on emotion, which could have affected performance on additional ETB tasks. However, this is only of potential concern for ETB studies if food is provided immediately before testing. It may also be the case that the EMEM response bias is a particularly sensitive measure, as it has good resolution (milliseconds versus percentage, number of words, etc.) and low noise (very low standard error values), which could explain why effects were not observed on more tasks and measures.

Another possibility is that despite selecting a sample size that should have been adequate to detect effects of satiation, the study was underpowered. By calculating effect sizes (Cohen's d) and conducting power analyses (G-power 3.1; power = 90%, $\alpha = .05$) it was possible to determine how many additional participants would be required to detect an effect of satiation for each ETB task measure. The lowest number of additional participants required was 96 (for EMEM accuracy), and the highest was 51,177 (for ECAT reaction times). The average number of additional participants required (across all tasks and measures) was 7251, and the average effect size was 0.14 (range = 0.01 to 0.29). Thus, given the high number of participants required to detect a significant effect, it is unlikely that we have incorrectly accepted the null hypothesis that there is no effect of satiation on most ETB tasks. In addition, significant effects of the valence of the emotional stimuli were observed, confirming effects observed in previous studies with the ETB. This adds further weight to the conclusion that the study was sufficiently powered to detect significant effects on performance.

As a measure of internal consistency between studies, scores for the primary measures used in Experiments 1 and 2 can be compared. Thus, compared to the results from Experiment 1 (in parentheses), volunteers in Experiment 2 showed the following response bias scores for the FERT: anger 0.75 (0.62), disgust 0.76 (0.70), fear 0.76 (0.70), happy 0.94 (0.94), neutral −0.23 (0.02), sad 0.70 (0.71), and surprise 0.74 (0.71). Hence, response bias scores were similar for the majority of emotions across both studies. For FDOT vigilance scores, results varied between the two studies as expected: happy −1.63 (0.87) and fear −3.85 (−0.98). ECAT reaction times were comparable across both studies: positive 795.1 ms (837.4 ms) and negative 826.9 ms (808.1 ms); as was EREC correct word recall: positive 7.1 (6.5) and negative 4.9 (5.7). Finally, ECAT response bias scores were also similar across both studies: positive −0.11 (−0.20) and negative 0.42 (0.34). Thus, the primary measures from the ETB tasks show good consistency between Experiments 1 and 2, with the exception of FDOT response bias.

Conclusion

In conclusion, we report adequate test–retest reliability for the ETB, confirming that the battery can be reliably used in repeated measures designs. We report evidence of practice effects for four out of five ETB tasks but provide further evidence that testing is stable after two sessions, suggesting that the ETB can be reliably used in repeated measures designs after initial training. Finally, we show that satiety state has only limited effects on performance on the ETB and, hence, is unlikely to be a confounding factor in ETB studies. Further work with alternative stimuli sets is proposed as a potential means to reduce practice effects. In addition, as these studies were conducted with lean healthy female participants, further work is necessary to investigate whether these effects generalize to other populations (e.g., men, individuals of varying weight and health status, etc.). These results are particularly important for the potential use of the ETB in clinical trials and clinical practice as they suggest that after initial training, the ETB is a robust and reliable measure of cognitive and emotional processing.

Acknowledgements

The authors would like to thank Catherine Harmer and Michael Browning for their helpful comments and suggestions, and Kim Verlaers and Wen Dong for their assistance with data collection.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by P1vital; the Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/G016739/1]; and the University of Birmingham.

ORCID

Jason Michael Thomas http://orcid.org/0000-0001-7013-8994

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bargh, J. A. (1982). Attention and automaticity in the processing of self-relevant information. *Journal of Personality and Social Psychology*, *43*, 425–436. doi:10.1037/0022-3514.43.3.425

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. doi:10.1001/archpsyc.1961.01710120031004

Benedict, R. H., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*, *20*, 339–352. doi:10.1076/jcen.20.3.339.822

Benton, D., & Parker, P. Y. (1998). Breakfast, blood glucose, and cognition. *American Journal of Clinical Nutrition*, *67*, 772S–778S.

Browning, M., Kingslake, J., Dourish, C. T., Harmer, C. J., Brammer, M., Goodwin, G. M., & Dawson, G. R. (2015). A precision medicine approach to antidepressant treatment in depression. *Journal of Psychopharmacology*, 29(8, Suppl.), A40.

Browning, M., Reid, C., Cowen, P. J., Harmer, C. J., & Goodwin, G. M. (2007). A single dose of citalopram increases fear recognition in healthy subjects. *Journal of Psychopharmacology*, *21*, 684–690. doi:10.1177/0269881106074062

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist*, *27*, 1077–1105. doi:10.1080/13854046.2013.809795

Chelune, G. J. (2002). Making neuropsychological outcomes research consumer friendly: A commentary on Keith et al. (2002). *Neuropsychology*, *16*, 422–425. doi:10.1037/0894-4105.16.3.422

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*, *9*, 419–428. doi:10.1017/S1355617703930074

Dye, L., Lluch, A., & Blundell, J. E. (2000). Macronutrients and mental performance. *Nutrition*, *16*, 1021–1034. doi:10.1016/S0899-9007(00)00450-0

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week, and one month test–retest intervals. *Journal of Clinical and Experimental*

Neuropsychology, *28*, 1096–1112. doi:10.1080/13803390500205718

Geller, V., & Shaver, P. (1976). Cognitive consequences of self-awareness. *Journal of Experimental Social Psychology*, *12*, 99–108. doi:10.1016/0022-1031(76)90089-5

Gibson, E. L., & Green, M. W. (2002). Nutritional influences on cognitive function: Mechanisms of susceptibility. *Nutrition Research Reviews*, *15*, 169–206. doi:10.1079/NRR200131

Goldberg, T. E., Keefe, R. S., Goldman, R. S., Robinson, D. G., & Harvey, P. D. (2010). Circumstances under which practice does not make perfect: A review of the practice effect literature in schizophrenia and its relevance to clinical treatment studies. *Neuropsychopharmacology*, *35*, 1053–1062. doi:10.1038/npp.2009.211

Green, M. W., Rogers, P. J., Elliman, N. A., & Gatenby, S. J. (1994). Impairment of cognitive performance associated with dieting and high levels of dietary restraint. *Physiology and Behavior*, *55*, 447–452. doi:10.1016/0031-9384(94)90099-X

Harmer, C. J., Bhagwagar, Z., Cowen, P. J., & Goodwin, G. M. (2002). Acute administration of citalopram facilitates memory consolidation in healthy volunteers. *Psychopharmacology*, *163*, 106–110. doi:10.1007/s00213-002-1151-x

Harmer, C. J., Bhagwagar, Z., Perrett, D. I., Vollm, B. A., Cowen, P. J., & Goodwin, G. M. (2003). Acute SSRI administration affects the processing of social cues in healthy volunteers. *Neuropsychopharmacology*, *28*, 148–152. doi:10.1038/sj.npp.1300004

Harmer, C. J., de Bodinat, C., Dawson, G. R., Dourish, C. T., Waldenmaier, L., Adams, S.,… Goodwin, G. M. (2010). Agomelatine facilitates positive versus negative affective processing in healthy volunteer models. *Journal of Psychopharmacology*, *25*, 1159–1167. doi:10.1177/0269881110376689

Harmer, C. J., Heinzen, J., O'Sullivan, U., Ayres, R. A., & Cowen, P. J. (2008). Dissociable effects of acute antidepressant drug administration on subjective and emotional processing measures in healthy volunteers. *Psychopharmacology*, *199*, 495–502. doi:10.1007/s00213-007-1058-7

Harmer, C. J., O'Sullivan, U., Favaron, E., Massey-Chase, R., Ayres, R., Reinecke, A,… Cowen, P. J. (2009). Effect of acute antidepressant administration on negative affective bias in depressed patients. *American Journal of Psychiatry*, *166*, 1178–1184. doi:10.1176/appi.ajp.2009.09020149

Harmer, C. J., Shelley, N. C., Cowen, P. J., & Goodwin, G. M. (2004). Increased positive versus negative affective perception and memory in healthy volunteers following selective serotonin and norepinephrine reuptake inhibition. *American Journal of Psychiatry*, *161*, 1256–1263. doi:10.1176/appi.ajp.161.7.1256

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*, 1267–1278. doi:10.1080/13854046.2010.526785

Hinton-Bayre, A., & Geffen, G. (2005). Comparability, reliability, and practice effects on alternate forms of the Digit Symbol Substitution and Symbol Digit Modalities tests. *Psychological Assessment*, *17*, 237–241. doi:10.1037/1040-3590.17.2.237

Horder, J., Cowen, P. J., Di Simplicio, M., Browning, M., & Harmer, C. J. (2009). Acute administration of the cannabinoid CB1 antagonist rimonabant impairs positive affective memory in healthy volunteers. *Psychopharmacology*, *205*, 85–91. doi:10.1007/s00213-009-1517-4

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. doi:10.1037/0022-006X.59.1.12

Kane, R. L., & Kay, G. G. (1992). Computerized assessment in neuropsychology: A review of tests and test batteries. *Neuropsychology Review*, *3*, 1–117. doi:10.1007/BF01108787

Lowe, M. R., Butryn, M. L., Didie, E. R., Annunziato, R. A., Thomas, J. G., Crerand, C. E., & Halford, J. (2009). The Power of Food Scale. A new measure of the psychological influence of the food environment. *Appetite*, *53*, 114–118. doi:10.1016/j.appet.2009.05.016

Macht, M., & Dettmer, D. (2006). Everyday mood and emotions after eating a chocolate bar or an apple. *Appetite*, *46*, 332–336. doi:10.1016/j.appet.2006.01.014

Macht, M., Gerer, J., & Ellgring, H. (2003). Emotions in overweight and normal-weight women immediately after eating foods differing in energy. *Physiology & Behavior*, *80*, 367–374. doi:10.1016/j.physbeh.2003.08.012

McClelland, G. R. (1987). The effects of practice on measures of performance. *Human Psychopharmacology*, *210*, 109–118. doi:10.1002/hup.470020206

Murphy, S. E., Downham, C., Cowen, P. J., & Harmer, C. J. (2008). Direct effects of diazepam on emotional processing in healthy volunteers. *Psychopharmacology*, *199*, 503–513. doi:10.1007/s00213-008-1082-2

National Institute for Health and Care Excellence. (2010). Alcohol-use disorders: Preventing harmful drinking. Retrieved from http://www.nice.org.uk/guidance/ph24/resources/guidance-alcoholuse-disorders-preventing-harmful-drinking-pdf

Nelson, H. E. (1982). *The National Adult Reading Test* (NART): *Test manual*. Windsor: NFER-Nelson.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*, 768–774. doi:10.1002/1097-4679(199511)51:63.0.CO;2-1

Post, A., Smart, T., Krikke, J., Witkin, J., Statnick, M., Harmer, C.,… Mohs, R. (2014). The efficacy and safety of LY2940094, a selective nociceptin receptor antagonist, in patients with major depressive disorder: A randomized, double-blind, placebo-controlled study. *Neuropsychopharmacology*, *39*, S346–S347.

Roebuck-Spencer, T., Sun, W., Cernich, A. N., Farmer, K., & Bleiberg, J. (2007). Assessing change with the Automated Neuropsychological Assessment Metrics

(ANAM): Issues and challenges. *Archives of Clinical Neuropsychology*, *22*, 79–87. doi:10.1016/j.acn.2006.10.011

Schmidt, K., Cowen, P. J., Harmer, C. J., Tzortzis, G., Errington, S., & Burnet, P. W. (2015). Prebiotic intake reduces the waking cortisol response and alters emotional bias in healthy volunteers. *Psychopharmacology*, *232*, 1793–1801. doi:10.1007/s00213-014-3810-0

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

Smith, A., Kendrick, A., Maben, A., & Salmon, J. (1994). Effects of breakfast and caffeine on cognitive performance, mood and cardiovascular functioning. *Appetite*, *22*, 39–55. doi:10.1006/appe.1994.1004

Smith, A., Leekam, S., Ralph, A., & McNeill, G. (1988). The influence of meal composition on post-lunch changes in performance efficiency and mood. *Appetite*, *10*, 195–203. doi:10.1016/0195-6663(88)90012-8

Smith, A., Ralph, A., & McNeill, G. (1991). Influences of meal size on post-lunch changes in performance efficiency, mood, and cardiovascular function. *Appetite*, *16*, 85–91. doi:10.1016/0195-6663(91)90034-P

Spitzer, R. L, Williams, J. B., Gibbon, M., & First, M. B. (2004). *Structured Clinical Interview for the DSM-IV* (SCID-I/P). New York, NY: Biometrics Research, New York State Psychiatric Institute.

Strauss, M. M., Makris, N., Aharon, I., Vangel, M. G., Goodman, J., Kennedy, D. N.,. . . Breiter, H. C. (2005). fMRI of sensitization to angry faces. *NeuroImage*, *26*, 389–413. doi:10.1016/j.neuroimage.2005.01.053

Stunkard, A. J., & Messick, S. (1985). The three-factor eating questionnaire to measure dietary restraint disinhibition and hunger. *Journal of Psychosomatic Research*, *29*, 71–83. doi:10.1016/0022-3999(85)90010-8

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Slotkin, J., & Gershon, R. (2014). The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *Journal of the International Neuropsychological Society*, *20*, 1–12. doi:10.1017/S1355617714000320

Wesnes, K., & Pincock, C. (2002). Practice effects on cognitive tasks: A major problem? *The Lancet Neurology*, *1*, 473. doi:10.1016/S1474-4422(02)00236-3.

Winston, A. P., Hardwick, E., & Jaberi, N. (2005). Neuropsychiatric effects of caffeine. *Advances in Psychiatric Treatment*, *11*, 432–439. doi:10.1192/apt.11.6.432