

1 **Test-retest reliability and effects of repeated testing and satiety on performance of an**
2 **Emotional Test Battery**

3

4 Running Title: Effects of repeated testing and satiety on performance on the ETB

5

6 J. M. Thomas, S. Higgs, & C. T. Dourish.

7

8 Full name: Dr Jason Michael Thomas

9 Affiliation: University of Birmingham

10 Address: School of Psychology, University of Birmingham, Birmingham, B15 2TT, UK.

11 Email: thomasjm@bham.ac.uk

12 Tel: +44 (0) 121 41 44899

13

14 Full name: Dr Suzanne Higgs

15 Affiliation: University of Birmingham

16 Address: School of Psychology, University of Birmingham, Birmingham, B15 2TT, UK.

17 Email: s.higgs.1@bham.ac.uk

18 Tel: +44 (0) 121 41 44907

19

20 Full name: Dr Colin Trevor Dourish

21 Affiliation: P1vital

22 Address: P1vital, Manor House, Howbery Park, Wallingford, Oxfordshire, OX10 8BA, UK.

23 Email: cdourish@p1vital.com

24 Tel: +44 (0) 1865 522030

25

26 Correspondence should be addressed to Dr Jason Michael Thomas.

27 This research was conducted at School of Psychology, University of Birmingham.

28

29 **Disclosures**

30 Dr Colin Dourish is an employee and shareholder of P1vital Limited, Dr Suzanne Higgs is a

31 member of P1vital's Advisory Panel, and Jason Michael Thomas is funded by the Steve

32 Cooper P1vital-BBSRC PhD Studentship.

33

34 Word Count: 8670

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51 **Abstract**

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

70

71 **Keywords:** Emotional Test Battery, ETB, practice effects, satiety, cross-over design

72

73

74

75

76 **Introduction**

77 Computerised test batteries have been used extensively to investigate the effects of
78 behavioural and pharmacological interventions on cognitive function. For example, the
79 P1vital[®] Oxford Emotional Test Battery (ETB, e.g. Murphy, Downham, Cowen, & Harmer,
80 2008) has been used to detect early effects of antidepressant drugs on cognitive-emotional
81 functioning and has been validated over a number years (e.g. Harmer et al. 2003; Harmer,
82 Shelley, Cowen, & Goodwin, 2004; Horder, Cowen, Di Simplicio, Browning, & Harmer,
83 2009; Harmer et al. 2010) in healthy volunteers (Harmer, Bhagwagar, Cowen, & Goodwin,
84 2002) and in patients with depression (Harmer et al. 2009; Post et al. 2014; Browning et al.
85 2015).

86

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

97

98 The Faces Dot Probe Task (FDOT) involves the presentation of two faces, which are replaced
99 by a pair of dots (192 trials in total). On some trials, one of the faces has an emotional
100 expression (happy versus fearful). Participants must report the orientation of the pair of dots

101 (i.e. vertical versus horizontal) for each trial. For this task a vigilance score is calculated as
102 the primary measure. This is a measure of sustained attention for a given stimulus and is
103 derived by subtracting the reaction times from congruent trials (trials where the probe appears
104 in the same location as the stimulus) from incongruent trials (trials where the probe appears
105 in a different location from the stimulus). Accuracy and reaction times can also be calculated
106 to examine potential speed-accuracy trade-off.

107

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

117

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

124

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

137

138 The use of multiple stimulus sets or alternate test forms across test sessions can overcome
139 some of the issues associated with repeated testing because participants are unable to learn
140 responses to specific stimuli, but this does not address changes in performance over time due
141 to procedural learning (Roebuck-Spencer, Sun, Cernich, Farmer, & Bleiberg, 2007). Another
142 useful approach to examine whether the rate of change in performance in an experimental
143 group differs from that in a control or reference group is test–retest variability or
144 measurement error (Jacobson and Truax 1991). This can identify the variability over time that
145 is expected by chance or due to other factors such as practice. Such approaches can also be
146 used to compare the performance of individuals to that of a group, for example to assess
147 whether a patient is responding to treatment (Chelune, 2002). However, an issue with this
148 approach is that a reference group may not be well matched on individual difference variables
149 that affect the degree of learning or practice on the tasks. In this case, an effect attributed to

150 an intervention may be better explained by pre-existing differences in the rate of change
151 between groups (Wesnes and Pincock, 2002). One way of minimising these issues is to assess
152 normative change when performance has plateaued and test-retest reliability is stable.

153

154 The test-re-test reliability of specific tests has been evaluated and a meta-analysis of practice
155 effects for a range of neuropsychological tests revealed substantial practice effects for many
156 tasks although the size of the effects dependent on factors such as the age of the participants
157 and the length of the re-test interval (Calamia, Markon, & Tranel, 2013). Moreover, an
158 examination of the reliability of the dot-probe attentional task suggested that performance
159 was neither internally consistent nor stable in a non-clinical sample of participants
160 (Schmukle, 2005). These data underscore the importance of assessing the reliability of
161 specific cognitive tests (Heilbronner, et al. 2010). To date there has been no examination of
162 test-retest reliability or how many sessions are required for performance on the ETB tasks to
163 stabilise, although previous work suggests that practice effects on other cognitive tasks are
164 minimised after 2-3 sessions (Collie, Maruff, Darby, & McStephen, 2003). It has been
165 recommended that four pre-study training sessions in psychopharmacology should be adopted
166 as a standard procedure (McClelland 1987). Hence, the aim of Study 1 was to assess the test-
167 retest reliability and stability of performance on ETB measures over 4 test sessions. Such
168 information is needed if learning effects are to be precluded from clinical studies where
169 accurate baseline measures of cognitive performance are required. In addition, such data add
170 to the body of knowledge on practice effects for cognitive tasks assessing different domains
171 of function.

172

173 Another methodological issue that arises when testing the effects of an intervention on
174 cognitive function is the extent to which hunger and satiety should be controlled for prior to

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

187

188 **Study 1**

189 **Methods and Materials**

190 *Participants*

191 30 healthy women student volunteers (mean age = 18.9 years; mean body mass index, BMI =
192 21.5; mean national adult reading score, NART = 111) were recruited for the study from the
193 University of Birmingham. Informed consent was obtained and participants were given either
194 £20 cash or course credits upon completion. The study was approved by the University of
195 Birmingham Research Ethics Committee and was conducted in accordance with the ethical
196 standards laid down in the 1964 Declaration of Helsinki. Participants were excluded from the
197 study if they were under 18 or over 65 years of age and if they were not fluent English
198 speakers. Using a screening questionnaire, participants were excluded if they: had previously
199 taken part in an ETB study; were dyslexic; smokers; taking medication; had consumed a high

200 amount of caffeine (> 750mg; Winston, Hardwick, & Jaber, 2005) or alcohol (> 3 units;
201 NICE, 2010) in the last 24 hours; or had current or past depression, determined by using the
202 questions for assessing depression only, from the Structured Clinical Interview for DSM-IV
203 Axis I Disorders (SCID – Spitzer, Williams, Gibbon, & First, 2004).

204

205 *Design*

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

212

213 *Procedure*

214 Participants completed a consent form before completing the screening measures. They had
215 their height and weight measured for BMI calculation then completed: the NART (Nelson,
216 1982) as an estimate of verbal IQ; the SCID (questions relating to depression only), a lifestyle
217 questionnaire (including questions about age, gender, medical conditions, smoker status, etc)
218 and an alcohol and caffeine questionnaire (documenting intake during the last 24 hours).
219 Participants were then given visual analogue scales (VAS) with the following mood and
220 appetite items to rate on a scale from 0-100mm (0mm anchor = not at all, 100mm anchor =
221 extremely): ‘alertness’; ‘disgust’; ‘drowsiness’; ‘light-headed’; ‘anxiety’; ‘happiness’;
222 ‘nausea’; ‘sadness’; ‘withdrawn’; ‘faint’; ‘hungry’; ‘full’; ‘desire to eat’ and ‘thirst’. After
223 this, participants completed the ETB (which took approximately 60 minutes) and then the
224 Three Factor Eating Questionnaire (TFEQ - Stunkard and Messick, 1985) and the Beck

225 Depression Inventory (BDI - Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) in a
226 counterbalanced order. Finally, participants completed another VAS questionnaire.

227

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

233

234 *Data Analysis*

235 *General:* Within-subjects analysis of variance (ANOVA) was used to analyse the data.

236 Bonferroni correction was used for all post-hoc t-tests and violations of sphericity were
237 addressed using the Greenhouse-Geisser correction.

238 *VAS:* To establish a factor structure for the VAS, a principal components analysis (PCA) was
239 run with varimax rotation. Analysis of the 14 items provided 4 factors with eigenvalues > 1 ,
240 accounting for 66.64% of the variance. Items that loaded > 0.5 onto a factor were included,
241 resulting in 4 factors of 3 or more items: appetite (desire to eat, hungry, fullness and thirst);
242 negative physical effects (faint, lightheaded and nausea); arousal (alertness, happiness and
243 drowsiness); negative mood (anxiety, sadness and disgust). Withdrawn did not load > 0.5
244 onto any of the factors and was analysed separately. Scores for each of the factors were
245 calculated by summing the scores for all items in that factor and then dividing by the number
246 of items. Items with a negative scale, were inverted to match the other items.

247 *ETB Data:*

248 Effects of session are reported first, followed by task specific effects that were relevant to the
249 task but not to the experimental manipulation. These are presented to confirm the ability to

250 detect effects of emotion and or valence. Main effects and interactions (session x
251 valence/emotion) were followed with t-tests to further analyse the data. For sessions,
252 comparisons consisted of sessions 1 versus 2, 2 versus 3, and 3 versus 4.

253

254 *Intraclass Correlation Coefficients:* To examine test-retest reliability for ETB task measures,
255 intraclass correlation coefficients (ICCs) were calculated using a two-way mixed-effects
256 model for absolute level of agreement. ICCs were calculated between sessions 1 to 2, 2 to 3,
257 and 3 to 4 for the primary measures of interest for the ETB tasks (split by emotion): FERT
258 response bias; ECAT reaction times; EREC correct word recall; and EMEM response bias.
259 ICCs were not conducted on FDOT vigilance scores as healthy participants do not show an
260 emotional bias on this task, hence it would not be expected that this measure would be
261 reliable over time. Instead, accuracy and reaction times were examined for reliability. Across
262 measures, an ICC less than 0.40 was considered poor test–retest reliability, 0.40–0.75
263 adequate, and 0.75 or greater was considered good to very good (Weintraub et al. 2014).

264

265 **Results**

266 *Questionnaire Data*

267 BDI scores were in the low range (mean = 6.8, SE = 1.2), alcohol consumption prior to
268 testing was low (mean = 0.04 units, SE = 0.02) and caffeine consumption was well within the
269 defined study limit (mean = 187.2mg, SE = 20.5). ANOVA comparing these measures across
270 the four test sessions did not show any significant differences (all $p > 0.05$). For the TFEQ
271 measures, cognitive restraint, disinhibition and hunger scores were all in the normal range
272 (mean = 7.2, SE = 1.2; mean = 6.5, SE = 0.6; mean = 7.4, SE = 0.7) and did not differ
273 significantly between sessions (all $p > 0.05$). Analysis of VAS ratings revealed that there
274 were no effects of session, time, or interaction between these factors for the following (all $p >$

275 0.05); Appetite (mean = 44.8, SE = 1.6); Negative Physical Effects (mean = 5.9, SE = 1.6);
276 Negative mood (mean = 8.1, SE = 1.6); Withdrawn (mean = 7.9, SE = 2.0); However, for
277 arousal there was a main effect of session ($F(3, 87) = 3.12; p < 0.05$). Bonferroni corrected t-
278 tests comparing sessions were not significant, though the closest to significance was the
279 decrease in arousal from session 1 to session 3 ($t(29) = 2.70; p = 0.07$) (session 1 mean = 64.1,
280 SE = 2.7; session 2 mean = 57.6, SE = 2.8; session 3 mean = 57.0, SE = 2.9; session 4 mean
281 = 59.3, SE = 3.1). There was no effect of time or a significant interaction for this measure
282 (both $p > 0.05$)

283

284 *ETB Data*

285 For reaction time measures, only data for correct responses were used. All data were
286 examined for outliers (± 3 standard deviations from the mean), resulting in the removal of
287 1.1% of the total ETB data set.

288

289 *Intraclass Correlation Coefficients (ICCs)*

290 Average ICC scores across all four sessions ranged from 0.4-0.8 for 16 out of the 17
291 measures (94%), indicating adequate test-retest reliability for the majority of measures (Table
292 1). The only exception was the FDOT accuracy score for positive words which displayed an
293 average ICC of 0.3, indicating poor test-retest reliability.

294

295 **INSERT TABLE 1**

296

297 *Facial expression recognition task (FERT)*: Repeated-measures ANOVA with session (4
298 levels: 1, 2, 3 and 4) and emotion (7 levels: anger, disgust, fear, happy, neutral, sad and
299 surprise) as factors revealed that for response bias there was no effect of session ($F(3, 72) =$

300 1.25; $p > 0.05$ – Figure 1), but there was an effect of emotion ($F(4, 86) = 105.06$; $p < 0.001$)
301 and an interaction approaching significance ($F(5, 114) = 2.28$; $p = 0.05$ – Figure 1). Breaking
302 down the interaction by emotion, there was a main effect of session for anger, neutral and
303 surprise (all $p < 0.05$), but not for disgust, fear, happy and sad (all $p > 0.05$). Examining the
304 effect of session for anger, neutral and surprise, Bonferroni corrected t-tests showed a
305 significant increase in response bias to anger expressions from session 1 to session 2 (0.63
306 versus 0.71; $t(29) = 2.905$; $p < 0.05$ – Figure 1). There were no other significant effects for any
307 other emotions.

308

309 **INSERT FIGURE 1**

310

311 For accuracy, there were main effects of session ($F(3, 78) = 5.65$; $p < 0.01$ – Figure 2) and
312 emotion ($F(3, 79) = 16.85$; $p < 0.01$), but no significant interaction ($p > 0.05$ – Figure 2).
313 Bonferroni corrected t-tests on the effect of session revealed that accuracy increased from
314 session 1 to 2 (55.7% versus 58.2%; $t(27) = -2.86$; $p < 0.05$), but did not differ significantly
315 between sessions 2 to 3 and 3 to 4 (both $p > 0.05$). Following up the effect of emotion,
316 accuracy in categorising anger (45.4%), disgust (53.8%), fear (51.6%), sadness (53.7%) and
317 surprise (59.9%) was lower than for neutral faces (70.8%) (all $p < 0.01$), while accuracy for
318 happy faces (69.2%) was not significantly different from accuracy for neutral faces ($p >$
319 0.05).

320

321 **INSERT FIGURE 2**

322

323 For reaction time there were main effects of session ($F(3, 69) = 28.53$; $p < 0.001$ – Figure 3)
324 and emotion ($F(3, 80) = 27.91$; $p < 0.001$) but no significant interaction ($p > 0.05$ – Figure 3).

325 Reaction times significantly decreased between sessions 1 and 2 (1331.6ms versus 1239.3ms;
326 $t(25) 3.63; p < 0.01$) and 2 and 3 (1242.9ms versus 1164.1ms; $t(27) 3.46; p < 0.01$), but not
327 between session 3 and 4 ($p > 0.05$). For the effect of emotion, reaction times to expressions
328 of anger (1322.1ms), disgust (1205.6ms), fear (1452.2ms), sadness (1184.2ms) and surprise
329 (1241.3ms) were significantly slower than to neutral faces (1049.7ms) (all $p < 0.01$), while
330 reaction times to happy faces (1055.0ms) and neutral faces did not differ ($p > 0.05$).

331

332 **INSERT FIGURE 3**

333

334 *Faces dot probe task (FDOT)*: Repeated-measures ANOVA with session (4 levels: 1, 2, 3
335 and 4), emotion (2 levels: fear and happy) and masking (2 levels: masked and unmasked) as
336 factors revealed that for vigilance scores, there was no main effect of session ($F(3, 78) =$
337 $1.13; p > 0.05$ – See Figure 4), emotion ($F(1, 26) = 0.74; p > 0.05$), or mask ($F(1, 26) = 0.05;$
338 $p > 0.05$), nor any significant interactions (all $p > 0.05$). The same repeated-measures
339 ANOVA was used for accuracy and reaction times, however, the factor of congruence was
340 added (2 levels: congruent and incongruent). For accuracy, there was a main effect of
341 masking on accuracy (masked faces = 96.7% versus unmasked faces = 96.1%; ($F(1, 25) =$
342 $4.31; p < 0.05$), but no effect of session (see Figure 4), emotion (fear versus happy) or
343 congruence (congruent versus incongruent probe location), nor any interactions (all $p >$
344 0.05). For reaction time, there was a main effect of session ($F(2, 56) = 10.86; p < 0.001$), an
345 interaction between emotion and session ($F(3, 75) = 3.95; p < 0.05$), and a four-way
346 interaction between masking, emotion, congruence and session ($F(3, 75) = 2.76; p < 0.05$).
347 Breaking down the four-way interaction by emotion, there were main effects of session for
348 reaction times to both fearful and happy expressions ($F(3, 78) = 10.62; p < 0.001; F(2, 61) =$
349 $10.52; p < 0.001$), but no other main effects or significant interactions (all $p > 0.05$).

350 Bonferroni corrected paired t-tests showed that response times reduced from sessions 1 to 2
351 for both emotions (happy, session 1 = 610.6ms vs. session 2 = 581.0ms, $p < 0.01$; fear,
352 session 1 = 614.7ms vs. session 2 = 587.0ms, $p < 0.01$ – see Figure 4). There was also a trend
353 for reaction times to fearful faces to decrease between sessions 3 and 4 (583.6 vs. 571.5; $p =$
354 0.06).

355

356 **INSERT FIGURE 4**

357

358 *Emotional categorisation task (ECAT):*

359 Repeated-measures ANOVA with session (4 levels: 1, 2, 3 and 4) and valence (2 levels:
360 positive and negative) as factors revealed that for reaction times there was no effect of
361 session (Figure 5), valence, or an interaction between session and valence (all $p > 0.05$). For
362 accuracy there was an effect of session ($F(2, 57) = 3.53$; $p < 0.05$), however, Bonferroni
363 corrected paired t-tests comparing sessions (1 versus 2; 2 versus 3; and 3 versus 4) were not
364 significant (all $p > 0.05$ – see Figure 5). The nearest to significance was the comparison
365 between session 3 and 4 (94.3% versus 93.1%, respectively; $p = 0.7$). There was also an
366 effect of valence on accuracy, whereby negative words were categorised more accurately
367 than positive words (mean = 95.6%, SE = 0.7 vs. mean = 93.5%, SE = 1.1; $F(1, 25) = 6.76$; p
368 = 0.07). There was no significant interaction between valence and session ($p > 0.05$).

369

370 **INSERT FIGURE 5**

371

372 *Emotional recall task (EREC):* Repeated-measures ANOVA with session (4 levels: 1, 2, 3
373 and 4) and valence (2 levels: positive and negative) as factors revealed a main effect of
374 session on the number of words correctly recalled ($F(3, 84) = 46.12$; $p < 0.001$). Bonferroni
375 corrected t-tests showed that accuracy increased from session 1 to 2 and session 2 to 3 (both p

376 < 0.001 – Figure 6), but did not change between sessions 3 and 4 ($p > 0.05$). There was also a
377 main effect of valence for the number of words correctly recalled (negative words = 8.1
378 versus positive words = 9.8; $F(1, 28) = 15.70$; $p < 0.001$), but no significant interaction
379 between valence and session ($F(3, 84) = 1.88$; $p > 0.05$).

380

381 For the number of incorrectly recalled words, there was a main effect of session ($F(3, 81) =$
382 8.59 ; $p < 0.001$), a main effect of valence ($F(1, 27) = 13.62$; $p < 0.01$), and an interaction
383 between session and valence ($F(3, 81) = 6.59$; $p < 0.001$). Breaking down the interaction by
384 valence, there was no effect of session for incorrectly recalled negative words ($F(3, 84) =$
385 0.56 ; $p > 0.05$), but there was an effect of session for incorrectly recalled positive words ($F(3,$
386 $84) = 13.13$; $p < 0.001$). Bonferroni corrected t-tests showed significant decreases in positive
387 words falsely recalled from session 1 to 2 ($t(29) = 2.71$; $p < 0.05$) and session 2 to 3 ($t(28)$
388 $= 2.64$; $p < 0.05$), but no difference between session 3 and 4 ($t(28) = 1.22$; $p > 0.05$ – see Figure
389 6).

390

391 **INSERT FIGURE 6**

392

393 *Emotional recognition memory task (EMEM):*

394 Repeated-measures ANOVA with session (4 levels: 1, 2, 3 and 4) and valence (2 levels:
395 positive and negative) as factors revealed that for response bias there was no effect of session
396 ($F(3, 84) = 1.24$; $p = 0.3$ – Figure 7), but there was a main effect of valence whereby
397 participants showed a greater response bias to negative words compared to positive (0.37
398 versus -0.14; $F(1, 28) = 140.99$; $p < 0.001$). There was no interaction between valence and
399 session ($p > 0.05$). For accuracy there was no effect of session ($F(3, 84) = 0.22$; $p > 0.05$ –
400 Figure 7), but there was a main effect of valence whereby positive words were recalled more

401 accurately than negative (mean = 83.8%, SE = 1.5 vs. mean = 68.7%, SE = 2.1; $F(1, 28) =$
402 $79.45; p < 0.001$). There was no interaction between valence and session ($p > 0.05$). For
403 reaction time, there was a main effect of session ($F(2, 59) = 4.51; p < 0.05$). Follow-up t-tests
404 (Bonferroni corrected) showed that reaction times significantly decreased between sessions 1
405 and 2 ($t(27) = 3.75; p < 0.01$ – Figure 7), however, there were no significant differences
406 between sessions 2 and 3, or 3 and 4 (both $p > 0.05$). An effect of valence was also noted for
407 reaction time whereby responses were quicker to positive words than negative words (mean =
408 929.3ms , SE = 39.9 vs. mean = 1022.1ms , SE = 43.0; $F(1, 26) = 52.89; p < 0.001$). There
409 was no interaction between valence and session ($p > 0.05$).

410

411 **INSERT FIGURE 7**

412

413 **Discussion**

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

418

419 The validity of using the ETB in repeated-measures designs rests on the assumption of
420 reliable test-retest results over sessions. Here we confirm that test-retest reliability scores for
421 the majority of the ETB measures were adequate, with many tasks yielding ICCs of 0.7 or
422 0.8. These data are comparable with the results of a recent meta-analysis reporting the mean
423 test-retest reliability of a range of cognitive tasks to be around 0.7 or higher (Calamia et al.
424 2013). Of the four measures showing poor test-retest reliability, FDOT accuracy scores
425 (positive and negative) were particularly unreliable, however, this is comparable to previous

426 work reporting a lack of internal consistency and stability in non-clinical samples with this
427 task (Schmukle, 2005). Reliability for the other two measures (EREC correct positive words
428 and EMEM negative response bias) reached adequate reliability for the final two sessions
429 (0.4 and 0.7, respectively), hence with the exception of the FDOT, all measures exhibit
430 reasonable reliability after the first two sessions.

431

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

444

445 For the ECAT the primary measure was reaction time and this did not change with repeated
446 testing. This may be due to the low cognitive demand of the task and the ease of accessing
447 self-referent stimuli; i.e. there was no capacity for practice to improve performance. Evidence
448 suggests that self-referent stimuli are processed automatically and faster than non-self-
449 referent stimuli (Bargh, 1982; Geller and Shaver, 1976). In addition, there was no difference
450 in reaction times to positive or negative words, and no interaction between session and

451 valence. Thus this measure appears to be resistant to practice effects, across all sessions and
452 valence.

453

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

465

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

470

471 For all but one task there was an acceleration of reaction time with repeated testing, but for
472 the last two sessions responding stabilised for all tasks. This pattern of results is consistent
473 with findings from other studies of practice effects on cognitive test batteries (e.g. Falletti,
474 Maruff, Collie, & Darby, 2006). This probably reflects the effects of familiarity with the task
475 procedures on reaction time since there was no speed-accuracy trade-off for any task that

476 might indicate a change in response strategy over time. Accuracy only improved with
477 repeated testing for the FERT and the EREC. The FERT requires participants to categorise
478 unfamiliar faces according to their emotional expression and hence increased familiarity may
479 have improved categorisation accuracy on this task.

480

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

494

495 The present results suggest that overall performance on the ETB tasks is stable after 2
496 sessions and that the ETB could be used for repeated test sessions with the inclusion of two
497 practice sessions. However, an issue might be whether after two practice sessions, there is
498 reduced sensitivity to detect significant effects of an experimental manipulation due to the
499 induction of a rigid response set or floor or ceiling effects. Ceiling effects were likely
500 observed for the EREC after two sessions because the number of items correctly recalled was

501 12 which may be at the limit of memory. The use of an alternative response set as previously
502 discussed would address this issue. For the EMEM and FERT, stable performance was at
503 levels where both increases and decreases in performance are likely to be detectable.
504 Together, the results suggest good reliability and limited practice effects, which are
505 potentially important findings for the use of ETB tasks in repeated assessment of depressed
506 patients in clinical studies and clinical practice. In particular, the test-retest reliability and
507 absence of practice effects for the FERT response bias measure are very encouraging, given
508 its recent use in the early assessment of antidepressant response in a primary care study
509 (Browning et al. 2015).

510

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

516

517 **Study 2**

518 **Methods and Materials**

519 *Participants*

520 30 healthy women psychology students (mean age = 21.4 years; mean BMI = 20.0; mean
521 NART = 117) were recruited from the University of Birmingham. Informed consent was
522 obtained from all participants, who were compensated after the study with either course
523 credits or £10 cash. The study was approved by the University of Birmingham Research
524 Ethics Committee and was conducted in accordance with the ethical standards laid down in
525 the 1964 Declaration of Helsinki. Exclusion criteria from Study 1 also applied to Study 2 (age

526 range, fluency in English, prior ETB study participation, dyslexia and smoker status,
527 medication use, caffeine and alcohol consumption and depression). In addition, participants
528 had to possess a BMI between 18.5 and 24.9, have no food allergies or diabetes, and score
529 less than 10 on the restraint scale of the TFEQ to be recruited. This is because high levels of
530 dietary restraint have been associated with impaired cognitive performance (Green, Rogers,
531 Elliman, & Gatenby, 1994). Participants were also excluded from taking part if they had
532 participated in Study 1; hence, none of the subjects included in Study 2 had taken part in
533 Study 1.

534

535 *Design*

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

548

549

550

551 *Cheese Sandwich Lunch*

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

558

559 *Procedure*

560 Prior to attending the test session, participants were sent the TFEQ via email to ensure they
561 were eligible for the study. Those who attended the test day (between 12pm and 2pm) were
562 screened with a lifestyle questionnaire, a breakfast questionnaire (to ensure they had not
563 consumed food since 8pm the previous day) the SCID (questions relating to depression only)
564 and the NART. Participants also completed an alcohol and caffeine screening questionnaire
565 to assess their intake over the last 24 hours, before completing a set of VAS. VAS items were
566 placed above the centre of a 100mm line, anchored with "not at all" (0mm) and "extremely"
567 (100mm), and included the items: alert; disgusted; drowsy; light-headed; anxious; happy;
568 nauseated; sad; withdrawn; faint; hungry; thirsty; full; and desire to eat.

569

570 Participants in the satiated condition were served a cheese sandwich lunch after which they
571 completed another VAS and a sandwich rating questionnaire. This questionnaire assessed
572 liking of the sandwich, whether the meal was a typical size, and whether participants ate
573 beyond comfortable fullness, using VAS scale items. Participants were then asked to wait in a
574 test cubicle for an hour before administration of the ETB test; as noted above, mood effects
575 have previously been detected an hour after eating. During this time they completed a VAS

576 after 30 minutes and 60 minutes, the latter immediately prior to ETB testing. Participants
577 were then asked to complete the ETB tasks, followed by a batch of questionnaires, including
578 the Power of food Scale as a measure of appetitive anticipation (PFS, Lowe et al. 2009), the
579 Barratt Impulsivity Scale as a measure of impulsive behaviour (BIS 11– Patton, Stanford, &
580 Barratt, 1995) and the BDI to assess depression and mood. Participants then had their height
581 and weight measured for calculation of BMI, were asked what they thought the aims of the
582 study were, debriefed and thanked for their time. Participants in the hungry condition
583 completed a similar procedure (also waiting an hour before testing on the ETB), but
584 consumed the lunch of cheese sandwiches after completing the ETB tasks.

585

586 *Data Analysis*

587 *General:* Between-subjects and mixed analysis of variance (ANOVA) were used to analyse
588 main effects of satiety state and interactions. Bonferroni correction was used for all post-hoc
589 t-tests, and violations of sphericity were addressed using the Greenhouse-Geisser correction.

590 *VAS:* The factor structure derived from Study 1 was applied to the VAS data from Study 2.

591 *ETB Data:* As with Study 1, effects of the manipulations are presented first, followed by task
592 specific effects (e.g. effects of emotion, or valence).

593

594 **Results**

595 *Participant Characteristics and Subjective State Questionnaires*

596 Mean values for participant characteristics and subjective state questionnaires, split by
597 hungry and satiated groups, are displayed in Table 2. Participants were young, with healthy
598 BMI scores and good verbal IQs (NART). They were within the normal range of
599 impulsiveness (BIS 11) and appetitive anticipation (PFS), and showed low scores on the BDI,
600 indicating normal mood. Their TFEQ scores were within the low-normal range and the mean

601 amount of food consumed was within expectations for a lunch. Using independent t-tests
602 (hungry versus satiated) no significant differences were observed for any measure (all $p >$
603 0.05).

604

605 **Insert Table 2**

606

607 *Visual Analogue Scales*

608 VAS scores were entered into mixed ANOVAs with the factor of satiety state (satiated versus
609 hungry) and time (pre versus post-manipulation). For appetite there was a main effect of
610 satiety state, time, and a significant interaction between satiety state and time (all $p < 0.001$).
611 Comparing pre versus post-manipulation ratings separately for each group, appetite
612 significantly decreased over time in the satiated group ($p < 0.001$), but not in the hungry
613 group ($p > 0.05$) (see Table 3). For arousal there was a main effect of time ($p < 0.05$),
614 whereby arousal decreased slightly (63.6mm to 58.3mm), but there was no effect of satiety
615 state or a significant interaction (both $p > 0.05$). For negative physical effects, there was no
616 effect of satiety state or time (both $p > 0.05$), but, there was a trend for an interaction between
617 satiety state and time ($p = 0.07$), however, follow-up t-tests did not reveal any significant
618 effects (both $p > 0.05$). For negative mood and withdrawn, there were no effects of satiety
619 state, time, or a significant interaction between satiety state and time (all $p > 0.05$).

620

621 **Insert Table 3**

622

623 *ETB Data*

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

627 *Facial expression recognition task (FERT):* A mixed ANOVA with satiety state (2 levels:
628 satiated and hungry) and emotion (7 levels: anger, disgust, fear, happy, neutral, sad and
629 surprise) as factors revealed that for response bias there was no effect of satiety state (satiated
630 = 0.62, hungry = 0.64; $F(1, 28) = 0.45$; $p > 0.05$), an effect of emotion ($F(2, 59) = 125.03$; p
631 < 0.001) and no significant interaction ($F(6, 168) = 0.52$; $p > 0.05$ – Figure 8). Bonferroni
632 corrected t-tests on the main effect of emotion showed that participants were significantly
633 biased towards anger (0.75), disgust (0.76), fear (0.76), happy (0.94) sad (0.69) and surprise
634 (0.74) faces, compared to neutral (-0.23) (all $p < 0.001$).

635

636 **INSERT FIGURE 8**

637

638 For accuracy, there was no effect of satiety state ($p > 0.05$), a main effect of emotion ($F(3$
639 $91) = 29.45$; $p < 0.001$), and no interaction ($p > 0.05$ –see Figure 9). Bonferroni corrected t-
640 tests on the effect of emotion showed that the accuracy for each emotion (anger = 46.0%,
641 disgust = 54.8%, fear = 46.7%, happy = 61.8 %, sad = 46.8 %, and surprise = 58.0 %) was
642 significantly lower compared to neutral (78.3%) (all $p < 0.01$). Analysis of reaction time data
643 also revealed no effect of satiety state ($p > 0.05$), a main effect of emotion ($F(6, 156) = 21.41$;
644 $p < 0.001$), and no interaction between emotion and satiety state ($p > 0.05$ – see Figure 9).

645 For the effect of emotion, reaction times to expressions of anger (1504.8ms), disgust
646 (1300.2ms), fear (1614.5ms), sadness (1414.6ms) and surprise (1387.5ms) were significantly
647 slower than to neutral faces (1124.6ms) (all $p < 0.01$), while reaction times to happy faces
648 (1179.6ms) were not significantly different from those to neutral faces ($p > 0.05$).

649

650 **INSERT FIGURE 9**

651

652 *Faces Dot Probe Task (FDOT)*: A mixed ANOVA with satiety state (2 levels: satiated and
653 hungry), emotion (2 levels: fear and happy) and masking (2 levels: masked and unmasked)
654 revealed that for vigilance scores there was no main effect of satiety state (hungry = -7.07
655 (SE = 4.27), satiated = 1.59 (SE = 4.41); $F(1, 27) = 1.99$; $p > 0.05$), emotion (fear = -3.85
656 (SE = 3.88), happy = -1.63 (SE = 5.03); $F(1, 27) = 0.12$; $p > 0.05$), or mask (masked = -3.32
657 (SE = 3.80), unmasked = -2.16 (SE = 5.03); $F(1, 27) = 0.03$; $p > 0.05$), nor any significant
658 interactions (all $p > 0.05$) (see Table 4). The same mixed ANOVA was used for accuracy and
659 reaction times, however, the factor of congruence was added (2 levels: congruent and
660 incongruent). For both measures, there was no main effect of satiety state (hungry versus
661 satiated; see Table 4), emotion (fear versus happy faces), masking (masked versus
662 unmasked), or congruency (congruent versus incongruent probe location) and no significant
663 interactions between these factors (all $p > 0.05$).

664

665 **Insert Table 4**

666

667 *Emotional categorisation task (ECAT)*: A mixed ANOVA with satiety state (2 levels: satiated
668 and hungry) and valence (2 levels: positive and negative) showed there was no effect of
669 satiety state, valence, nor an interaction between satiety state and valence (positive versus
670 negative words) for ECAT accuracy (all $p > 0.05$; see Table 4). Analysis of ECAT reaction
671 time showed no effect of satiety state ($p > 0.05$), a trend towards a main effect of valence
672 with quicker times for positive versus negative words ($F(1, 28) = 4.16$; $p = 0.05$), and no
673 interaction ($p > 0.05$).

674

675 *Emotional recall task (EREC)*: A mixed ANOVA with satiety state (2 levels: satiated and
676 hungry) and valence (2 levels: positive and negative) revealed that for words correctly

677 recalled, there was no effect of satiety state ($p > 0.05$), a main effect of valence with more
678 positive words recalled versus negative ($F(1, 28) = 54.24; p < 0.001$; see Table 4), and no
679 significant interaction ($p > 0.05$). For words incorrectly recalled, there was also no effect of
680 satiety state ($p > 0.05$), an effect of valence with more positive words recalled versus negative
681 ($F(1, 28) = 15.97; p < 0.001$; see Table 4), and no significant interaction ($p > 0.05$).

682

683 *Emotional recognition memory task (EMEM)*: A mixed ANOVA with satiety state (2 levels:
684 satiated and hungry) and valence (2 levels: positive and negative) showed that for response
685 bias, there was an effect of satiety state ($F(1, 28) = 10.25; p < 0.01$), an effect of valence (F
686 ($1, 28) = 64.02; p < 0.001$), and a significant interaction ($F(1, 28) = 5.59; p < 0.05$ –see Table
687 4). Breaking down the interaction by emotion, response bias to the positive words was
688 significantly lower in satiated compared to hungry individuals (-0.34 versus 0.12; $t(28) = 3.24$;
689 $p < 0.01$). There was no significant difference in response bias between satiated and hungry
690 individuals to the negative words (0.35 versus 0.49; $t(28) = 1.78; p > 0.05$). Accuracy scores
691 showed no effect of satiety state ($p > 0.05$), a main effect of valence with better accuracy for
692 positive versus negative words ($F(1, 27) = 59.97; p < 0.001$; see Table 4), and no significant
693 interaction ($p > 0.05$). Analysis of reaction time also showed no effect of satiety state ($p >$
694 0.05), an effect of valence with quicker times for positive versus negative words ($F(1, 28) =$
695 $54.24; p < 0.001$ – see Table 4), and no significant interaction ($p > 0.05$).

696

697 **Discussion**

698 We report the first investigation of eating to satiety on performance for each of the ETB
699 tasks. Eating to satiety has only limited effects on ETB task performance, affecting EMEM
700 response bias only. These data suggest that a robust satiety manipulation has very limited

701 effects on ETB performance and therefore satiety state is unlikely to be a significant
702 confound in ETB studies.

703

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

717

718 It is possible that the lack of wider effects of satiety on the ETB is related to the food used in
719 this study. For instance, a study by Macht and Dettmer (2006) reported that both apple and
720 chocolate consumption elevated mood in healthy women, but the effect of chocolate
721 consumption was greater than the effect of apple consumption. Hence, it is possible that
722 highly palatable or energy dense foods have greater effects on mood than less palatable or
723 less energy dense foods. This suggestion is supported by evidence that foods with a high
724 energy content have greater effects on mood than food with a lower energy content (Macht,
725 Gerer, & Ellgring, 2003). Thus, the use of a food that is more palatable or energy dense than

726 bland cheese sandwiches may have elicited greater effects on emotion, which could have
727 affected performance on additional ETB tasks. However, this is only of potential concern for
728 ETB studies if food is provided immediately before testing. It may also be the case that the
729 EMEM response bias is a particularly sensitive measure, as it has good resolution
730 (milliseconds versus percentage, number of words, etc.) and low noise (very low standard
731 error values), which could explain why effects were not observed on more tasks and
732 measures.

733

734 Another possibility is that despite selecting a sample size that should have been adequate to
735 detect effects of satiation, the study was underpowered. By calculating effect sizes (Cohen's
736 *d*) and conducting power analyses (G-power 3.1; power = 90%, $\alpha = 0.05$) it was possible to
737 determine how many additional participants would be required to detect an effect of satiation
738 for each ETB task measure. The lowest number of additional participants required was 96
739 (for EMEM accuracy) and the highest was 51,177 (for ECAT reaction times). The average
740 number of additional participants required (across all tasks and measures) was 7251 and the
741 average effect size was 0.14 (range = 0.01 to 0.29). Thus, given the high number of
742 participants required to detect a significant effect, it is unlikely that we have incorrectly
743 accepted the null hypothesis that there is no effect of satiation on most ETB tasks. In
744 addition, significant effects of the valence of the emotional stimuli were observed, confirming
745 effects observed in previous studies with the ETB. This adds further weight to the conclusion
746 that the study was sufficiently powered to detect significant effects on performance.

747

748 As a measure of internal consistency between studies, scores for the primary measures used
749 in studies 1 and 2 can be compared. Thus, compared to the results from Study 1 (in
750 parentheses), volunteers in study 2 showed the following response bias scores for the FERT:

751 anger 0.75 (0.62), disgust 0.76 (0.70), fear 0.76 (0.70), happy 0.94 (0.94), neutral -0.23
752 (0.02), sad 0.70 (0.71) and surprise 0.74 (0.71). Hence, response bias score were similar for
753 the majority of emotions across both studies. For FDOT vigilance scores, results varied
754 between the two studies as expected: happy -1.63 (0.87) and fear -3.85 (-0.98). ECAT
755 reaction times were comparable across both studies: positive 795.1ms (837.4ms) and negative
756 826.9ms (808.1ms); as was EREC correct word recall: positive 7.1 (6.5) and negative 4.9
757 (5.7). Finally, ECAT response bias scores were also similar across both studies: positive -
758 0.11 (-0.20) and negative 0.42 (0.34). Thus, the primary measures from the ETB tasks show
759 good consistency between studies 1 and 2, with the exception of FDOT response bias.

760

761 **Conclusion**

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

775

776 **Acknowledgements**

777 This work was supported by P1 vital, the BBSRC under BB/G016739/1 and the University of
778 Birmingham. The authors would like to thank Professor Catherine Harmer and Dr Michael
779 Browning for their helpful comments and suggestions, and Miss Kim Verlaers and Miss Wen
780 Dong for their assistance with data collection.

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801 **References**

802 Bargh, J. A. (1982). Attention and automaticity in the processing of self-relevant information.

803 *Journal of Personality and Social Psychology*, 43, 425–436. doi: 10.1037/0022-

804 3514.43.3.425.

805

806 Beck, A. T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for

807 measuring depression. *Archives of General Psychiatry*, 4, 561–571. doi:

808 10.1001/archpsyc.1961.01710120031004.

809

810 Benedict, R. H., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations

811 of memory tests with and without alternate forms. *Journal of Clinical and Experimental*

812 *Neuropsychology*, 20, 339-52. doi: 10.1076/jcen.20.3.339.822

813

814 Benton, D., & Parker, P. Y. (1998). Breakfast, blood glucose, and cognition. *American*

815 *Journal of Clinical Nutrition*, 67, 772S-778S.

816

817 Browning, M., Reid, C., Cowen P. J., Harmer, C. J., & Goodwin, G. M. (2007). A single dose

818 of citalopram increases fear recognition in healthy subjects. *Journal of Psychopharmacology*,

819 21, 684–690. doi: 10.1177/0269881106074062.

820

821 Browning, M., Kingslake, J., Dourish, C.T., Harmer, C.J., Brammer, M., Goodwin, G.M., &

822 Dawson G. R. (2015). A Precision Medicine Approach to Antidepressant Treatment in

823 Depression. *Journal of Psychopharmacology*, 29, (8) Suppl, A40.

824 Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological
825 measures: meta-analyses of test-retest correlations. *Clin Neuropsychol*, *27*, 1077-105. doi:
826 10.1080/13854046.2013.809795.

827

828 Chelune, G. J. (2002). Making neuropsychological outcomes research consumer friendly: A
829 commentary on Keith et al. (2002). *Neuropsychology*, *16*, 422–425. doi: 10.1037/0894-
830 4105.16.3.422.

831

832 Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the
833 cognitive test performance of neurologically normal individuals assessed at brief test–retest
834 intervals. *Journal of the International Neuropsychological Society*, *9*, 419-428. doi:
835 10.1017/S1355617703930074.

836

837 Dye L., Lluch A., & Blundell J. E. (2000). Macronutrients and mental performance. *Nutrition*
838 *16*, 1021–1034. doi: 10.1016/S0899-9007(00)00450-0.

839

840 Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with
841 the repeated assessment of cognitive function using the CogState battery at 10-minute, one
842 week, and one month test-retest intervals. *Journal of Clinical and Experimental*
843 *Neuropsychology*, *28*, 1096–1112. doi: 10.1080/13803390500205718.

844

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

847

848 Gibson, E.L., & Green, M. W. (2002). Nutritional influences on cognitive function:
849 mechanisms of susceptibility. *Nutrition research reviews*, *15*, 169-206. doi:
850 10.1079/NRR200131.
851

852 Goldberg, T. E., Keefe, R. S., Goldman, R. S., Robinson, D. G., & Harvey, P. D. (2010).
853 Circumstances under which practice does not make perfect: a review of the practice effect
854 literature in schizophrenia and its relevance to clinical treatment studies.
855 *Neuropsychopharmacology*, *35*, 1053-1062. doi: 10.1038/npp.2009.211.
856

857 Green, M. W., Rogers, P. J., Elliman N. A., & Gatenby, S. J. (1994). Impairment of cognitive
858 performance associated with dieting and high levels of dietary restraint. *Physiology and*
859 *Behavior*, *55*, 447-452. doi: 10.1016/0031-9384(94)90099-X.
860

861 Harmer, C. J., Bhagwagar, Z., Cowen, P. J., & Goodwin, G. M. (2002). Acute administration
862 of citalopram facilitates memory consolidation in healthy volunteers. *Psychopharmacology*
863 *(Berl)*, *163*, 106 – 110. doi: 10.1007/s00213-002-1151-x.
864

865 Harmer, C. J., Bhagwagar, Z., Perrett, D. I., Vollm, B. A., Cowen, P. J., & Goodwin, G. M.
866 (2003). Acute SSRI administration affects the processing of social cues in healthy volunteers.
867 *Neuropsychopharmacology*, *28*, 148–152. doi:10.1038/sj.npp.1300004.
868

869 Harmer, C. J., Shelley, N. C., Cowen, P. J., & Goodwin, G. M. (2004). Increased positive
870 versus negative affective perception and memory in healthy volunteers following selective

871 serotonin and norepinephrine reuptake inhibition. *American Journal of Psychiatry*, *161*,
872 1256–1263. doi: 10.1176/appi.ajp.161.7.1256.

873

874 Harmer, C. J., Heinzen, J., O’Sullivan, U., Ayres, R. A., & Cowen, P. J. (2008). Dissociable
875 effects of acute antidepressant drug administration on subjective and emotional processing
876 measures in healthy volunteers. *Psychopharmacology*, *199*, 495–502. doi: 10.1007/s00213-
877 007-1058-7.

878

879 Harmer, C. J., O’Sullivan, U., Favaron, E., Massey-Chase, R., Ayres, R., Reinecke, A., ...
880 Cowen, P. J. (2009). Effect of acute antidepressant administration on negative affective bias
881 in depressed patients. *American Journal of Psychiatry*, *166*, 1178-1184. doi:
882 10.1176/appi.ajp.2009.09020149.

883

884 Harmer, C. J., de Bodinat, C., Dawson, G. R., Dourish, C. T., Waldenmaier, L., Adams, S.,
885 ... Goodwin, G. M. (2010). Agomelatine facilitates positive versus negative affective
886 processing in healthy volunteer models. *Journal of Psychopharmacology*, *25*, 1159-67. doi:
887 10.1177/0269881110376689.

888

889 Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010).
890 Official position of the American Academy of Clinical Neuropsychology on serial
891 neuropsychological assessments: The utility and challenges of repeat test administrations in
892 clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*, 1267–1278. doi:
893 10.1080/13854046.2010.526785.

894 Hinton-Bayre, A., & Geffen, G. (2005). Comparability, reliability, and practice effects on
895 alternate forms of the Digit Symbol Substitution and Symbol Digit Modalities tests.

896 *Psychological Assessment, 17*, 237-41. doi: 10.1037/1040-3590.17.2.237

897

898 Horder, J., Cowen, P. J., Di Simplicio, M., Browning, M., & Harmer, C. J. (2009). Acute
899 administration of the cannabinoid CB1 antagonist rimonabant impairs positive affective
900 memory in healthy volunteers. *Psychopharmacology, 205*, 85–91. doi: 10.1007/s00213-009-
901 1517-4.

902

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

906

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

909

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

913

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

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

919

920 McClelland, G. R. (1987). The effects of practice on measures of performance. *Human*
921 *Psychopharmacology*, *210*, 109–18. doi: 10.1002/hup.470020206.

922

923 Murphy, S. E., Downham, C., Cowen, P. J., & Harmer, C. J. (2008). Direct effects of
924 diazepam on emotional processing in healthy volunteers. *Psychopharmacology (Berl)*, *199*,
925 503 – 513. doi: 10.1007/s00213-008-1082-2.

926

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

930

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

932

933 Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt
934 impulsiveness scale. *Journal of Clinical Psychology*, *51*, 768–774. doi: 10.1002/1097-
935 4679(199511)51:63.0.CO;2-1.

936

937 Post, A., Smart, T., Krikke, J., Witkin, J., Statnick, M., Harmer, C., ... Mohs R. (2014). The
938 efficacy and safety of LY2940094, a selective nociceptin receptor antagonist, in patients with

939 major depressive disorder: A randomized, double-blind, placebo-controlled study.
940 *Neuropsychopharmacology*, 39, S346-347.

941

942 Roebuck-Spencer, T., Sun, W., Cernich, A. N., Farmer, K., & Bleiberg, J. (2007). Assessing
943 change with the Automated Neuropsychological Assessment Metrics (ANAM): issues and
944 challenges. *Archives of Clinical Neuropsychology*, 22, 79-87. doi:10.1016/j.acn.2006.10.011.

945

946 Schmidt, K., Cowen, P.J., Harmer, C. J., Tzortzis, G., Errington, S., & Burnet, P. W. (2015).
947 Prebiotic intake reduces the waking cortisol response and alters emotional bias in healthy
948 volunteers. *Psychopharmacology (Berl)*, 232, 1793-801. doi: 10.1007/s00213-014-3810-0.

949

950 Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*,
951 19, 595-605. doi: 10.1002/per.554.

952

953 Smith, A., Leekam, S., Ralph, A. & McNeill, G. (1988). The influence of meal composition
954 on post-lunch changes in performance efficiency and mood. *Appetite*, 10, 195-203. doi:
955 10.1016/0195-6663(88)90012-8.

956

957 Smith, A., Ralph, A., & McNeill, G. (1991). Influences of meal size on post-lunch changes in
958 performance efficiency, mood, and cardiovascular function. *Appetite*, 16, 85-91. doi:
959 10.1016/0195-6663(91)90034-P.

960

961 Smith, A., Kendrick, A., Maben, A., & Salmon, J. (1994). Effects of breakfast and caffeine
962 on cognitive performance, mood and cardiovascular functioning. *Appetite*, 22, 39-55. doi:
963 10.1006/appe.1994.1004.

964

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

968

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

972

973 Stunkard, A. J. & Messick, S. (1985). The three-factor eating questionnaire to measure
974 dietary restraint disinhibition and hunger. *Journal of Psychosomatic Research*, 29, 71-83. doi:
975 10.1016/0022-3999(85)90010-8.

976

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

981

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

984 Winston, A. P., Hardwick, E., & Jaber, N. (2005). Neuropsychiatric effects of caffeine.

985 *Advances in Psychiatric Treatment*, 11, 432–9. doi: 10.1192/apt.11.6.432.

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005 **Table 1** Intraclass Correlation Coefficients (ICCs) for ETB tasks split by emotion over sessions

Task and Measure	Intraclass Correlation Coefficients (ICCs)			
	Session 1 - Session 2	Session 2 - Session 3	Session 3 - Session 4	Average ICC
FERT Response Bias – Anger	0.6***	0.7***	0.8***	0.7
FERT Response Bias – Disgust	0.6***	0.7***	0.8***	0.7
FERT Response Bias – Fear	0.4**	0.8***	0.7***	0.6
FERT Response Bias – Happy	0.4*	0.4*	0.5**	0.4
FERT Response Bias – Neutral	0.5**	0.6***	0.6***	0.6
FERT Response Bias – Sad	0.8***	0.7***	0.8***	0.8
FERT Response Bias – Surprise	0.7***	0.8***	0.8***	0.8
FDOT Accuracy – Positive ^a	0.4*	0.3	0.5**	0.4
FDOT Accuracy – Negative ^a	0.6***	0.3	0.1	0.3
FDOT Reaction Times – Positive	0.5***	0.7***	0.8***	0.7
FDOT Reaction Times – Negative	0.6***	0.6***	0.8***	0.7
ECAT Reaction Times - Positive	0.7***	0.8***	0.7***	0.7
ECAT Reaction Times - Negative	0.6***	0.7***	0.8***	0.7
EREC Correct Words – Positive ^a	0.2*	0.7***	0.7***	0.5
EREC Correct Words – Negative	0.5***	0.5***	0.5**	0.5
EMEM Response Bias – Positive	0.5**	0.5**	0.6***	0.6
EMEM Response Bias – Negative ^a	0.4**	0.2	0.4*	0.4

1006

1007 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

1008 ^a measures with ICCs < 0.4

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

Table 2 Participant Characteristics & Subjective State Questionnaires from Study 2 (standard error of the mean)

Measure	Condition	
	Hungry	Satiated
Age	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)
TFEQ Disinhibition	5.3 (0.7)	7.1 (1.0)
TFEQ Hunger	5.4 (1.0)	7.3 (0.9)
Amount Eaten (grams)	193.6 (16.7)	188.5 (15.5)

1020

1021 Three Factor Eating Questionnaire (TFEQ)

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

Table 3 Visual Analogue Scale mean scores split by satiety state and time (standard error of the mean)

VAS Item	Hungry		Satiated	
	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)

1045 ^a = Main effect of satiety state; ^b = Main effect of time; ^c = Interaction between satiety state and time

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

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 (standard error of the mean)

ETB Task	Measure	Negative		Positive	
		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	630.8 (14.8)	642.1 (15.3)	631.9 (15.9)	643.4 (16.4)
Emotional Categorisation (ECAT)	Accuracy	96.7 (1.0)	97.4 (1.0)	97.4 (1.0)	95.0 (1.0)
	Reaction Time	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 ^b	1081.3 (62.5)	1093.1 (62.5)	915.7 (44.0)	912.1 (44.0)

^a Main effect of satiety state ($p < 0.01$) ^b Main effect of valence ($p < 0.001$)

^c Interaction between satiety state and valence ($p < 0.05$)

1065
1066
1067
1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081 **Figure Captions**

1082

1083 **Figure 1** Facial expression recognition task (FERT): response bias, split by emotion and test session (**left**), and
1084 split by session only (**right**). To the presentation of anger expressions only, response bias increased from session
1085 1 to session 2. Error bars represent standard error of the mean. $*p < 0.05$

1086

1087 **Figure 2** Facial expression recognition task (FERT): accuracy, split by emotion and test session (**left**), and split
1088 by session only (**right**). There was an overall effect of session, whereby accuracy increased from session 1 to
1089 session 2. Error bars represent standard error of the mean. $*p < 0.05$

1090

1091 **Figure 3** Facial expression recognition task (FERT): reaction times, split by emotion and test session (**left**), and
1092 split by session only (**right**). There was an overall effect of session, whereby accuracy increased from session 1
1093 to session 2 and session 2 to session 3. Error bars represent standard error of the mean. $**p < 0.01$

1094

1095 **Figure 4** Faces dot probe task (FDOT): vigilance score (**left**), accuracy (**centre**) and reaction times (**right**) to
1096 happy and fearful expressions for the four test sessions. Reaction times to both happy and fearful faces
1097 decreased significantly from session 1 to session 2. Error bars represent standard error of the mean. $**p < 0.01$

1098

1099 **Figure 5** Emotional categorisation task (ECAT): reaction times (**left**) and accuracy (**right**) to positive and
1100 negative words for the four test sessions. Error bars represent standard error of the mean.

1101

1102 **Figure 6** Emotional recall task (EREC): Correctly recalled words split by valence and session (**left**) split by
1103 session only (**centre**) and incorrectly recalled words split by valence and session (**right**). Number of words
1104 correctly recalled increased from sessions 1 to 2 and 2 to 3, but not 3 to 4. For positive words incorrectly
1105 recalled, there was a significant decrease from sessions 1 to 2 and 2 to 3, but again, no change between sessions
1106 3 to 4. Error bars represent standard error of the mean. $*p < 0.05$, $***p < 0.01$

1107

1108

1109 **Figure 7** Emotional recognition memory task (EMEM): A) Response bias split by valence and session (**top left**)
1110 and valence only (**top right**); B) Accuracy split by valence and session (**middle left**) and valence only (**middle**
1111 **right**); C) Reaction times split by valence and session (**bottom left**) and session only (**bottom right**). There was
1112 a significant response bias towards negative words compared to positive words (but no main effect of session, p
1113 = 0.3); positive words were recognised with greater accuracy compared to negative words; and reaction times
1114 significantly decreased between the first and second session. Error bars represent standard error of the mean.
1115 ** ($p < 0.01$) *** ($p < 0.001$).

1116

1117 **Figure 8** Facial expression recognition task (FERT): response bias, split by satiety state and emotion. Error bars
1118 represent standard error of the mean.

1119

1120 **Figure 9** Facial expression recognition task (FERT): accuracy (**left**) and reaction times (**right**) split by satiety
1121 state and emotion. Error bars represent standard error of the mean.

1122

1123

1124

1125