Adams T, Pounder Z, Preston S, Hanson A, Gallagher P, Harmer CJ, McAllister-Williams RH.

Test-retest reliability and task order effects of emotional cognitive tests in healthy subjects.

***Cognition and Emotion* 2016, 30(7), 1247-1259.**

Test-retest reliability and task order effects of emotional cognitive tests in healthy subjects

Thomas Adams[1,2], Zoe Pounder[1], Sally Preston[1,2], Andy Hanson[1], Peter Gallagher[1],

Catherine J. Harmer[3] & R. Hamish McAllister-Williams[1,2]

Author affiliations:
1 – Institute of Neuroscience, Newcastle University, UK
2 – Northumberland Tyne and Wear NHS Foundation Trust, UK
3 – Department of Psychiatry, University of Oxford, UK

Short title: Retest reliability of emotional tasks

Corresponding author: Dr R. Hamish McAllister-Williams

Mailing address: Institute of Neuroscience, Newcastle University,
Academic Psychiatry, Campus for Ageing and Vitality
Newcastle upon Tyne, NE4 5PL

E-mail address: r.h.mcallister-williams@ncl.ac.uk

Telephone: +44(0) 191 208 1370

Fax: +44(0) 191 208 1387

Abstract: 148 words

Manuscript: 6417

Figures: 4

Tables: 1

References: 34

Key words: Test-Retest reliability; attentional blink, Emotional test battery; Facial emotion recognition task

Abstract

Little is known of the retest reliability of emotional cognitive tasks or the impact of using different tasks employing similar emotional stimuli within a battery. We investigated this in healthy subjects. We found improved overall performance in an emotional attentional blink task (EABT) with repeat testing at one hour and one week compared to baseline, but the impact of an emotional stimulus on performance was unchanged. Similarly performance on a facial expression recognition task (FERT) was better one week after a baseline test, though the relative effect of specific emotions was unaltered. There was no effect of repeat testing on an emotional word categorising, recall and recognition task. We found no difference in performance in the FERT and EABT irrespective of task order. We concluded that it is possible to use emotional cognitive tasks in longitudinal studies and combine tasks using emotional facial stimuli in a single battery.

Background

‘Emotional processing’ is a generic term sometimes used to refer to the processing of emotional stimuli during a broad range of cognitive tasks. These include both the recognition of emotions and the impact of emotional stimuli on attention and memory. Cognitive tasks used to test emotional processing often use human facial expressions of emotion or emotionally valenced words. Such studies have shown a negative response bias towards sadness in individuals with major depression (Gotlib *et al.* 2004), so that positive (happy), neutral or ambiguous facial expressions tend to be evaluated as more sad or less happy, and with selective attention towards sad expressions and away from happy expressions, compared with healthy control groups (Bourke *et al.* 2010). In addition, there is some evidence of relatively better recall of negative emotionally valenced memories compared to positive and neutral memories (Dere *et al.* 2010). Antidepressant medication has been found to modify this process in both healthy volunteers and depressed individuals (Harmer *et al.* 2009; Tranter *et al.* 2009). It has been argued that this change in emotional processing with antidepressants is central to their mode of action in depression (Harmer *et al.* 2009). Changes in emotional processing have frequently been studied using an “emotional test battery” comprising tests of verbal memory for positive and negative valenced words, a facial expression recognition test (FERT) and an emotional dot probe task. This emotional test battery has now been used in a large number of studies of the effects of various psychotropic medication (Arnone *et al.* 2009; Harmer *et al.* 2008; Harmer *et al.* 2011).

The vast majority of behavioural studies investigating emotional processing in patients with depression, or the effect of antidepressant treatments, have used a between group, cross-sectional design. However some research questions necessitate repeat testing of subjects in longitudinal studies to examine the state/trait nature of changes in emotional processing with illness and effects of medication over time. Nevertheless, to our knowledge there is no data on the test-retest reliability of the emotional test battery. Elements of the emotional test battery, particularly the FERT or similar tasks, have been studied in longitudinal studies of medication in patient populations (Pavuluri *et al.* 2012; Tranter *et al.* 2009) though without the inclusion of control groups, thus making it impossible to know whether changes might have in fact been due to retesting. Changes in performance in tasks over time affecting their test-retest reliability might arise due to processes such as habituation to emotional stimuli and/or learning/practice effects.

Repeated exposure to emotional stimuli, including emotional faces (Ishai *et al.* 2004) and words (Protopopescu *et al.* 2005), has been reported to be associated with habituation in the amygdala response. It should be noted that some emotions such as anger have been reported to lead to sensitization, rather than habituation, of amygdala responses (Strauss *et al.* 2005). Nevertheless, potential amygdala habituation raises concerns about the feasibility of not only repeat testing of a particular emotional task, but also raises concerns about the effects of preceding tasks employing emotional stimuli on subsequent ones using the same or different stimuli in the same testing session. The time scales of reported amygdala habituation are generally over the course of seconds or minutes (Ishai *et al.* 2004). While there is little data to indicate how long such effects last for, there is some suggestion that there can be some recovery of amygdala response over the course of 20 mins (Britton *et al.* 2008). From an evolutionary perspective this makes sense. A fearful face may signal danger, but if the danger does not manifest itself, there is an advantage in dampening down the response to further fearful faces in the short term. There is no advantage, however, in maintaining a diminished response to fearful faces over the course of days, as the next time the danger may be real and present.

Of potentially greater relevance to the time scale being used in longitudinal studies of emotional processing in patients with depression and the study of effects of antidepressants are learning and practice effects. A recent meta-analysis of various cognitive tasks, not including emotional stimuli, found practice effects (being defined as improvements in accuracy or reductions in response time with repeat testing on the same task) influencing performance maintained over several years (Calamia *et al.* 2012). However there was a great deal of variation in the magnitude of effect between tasks suggesting that effect of retest interval may be unique for each neuropsychological test.

The importance of investigating test-retest reliability of investigations, particularly those using emotional stimuli is illustrated by a number of recent publications regarding the reproducibility of fMRI data of amygdala activation by emotive faces over periods of 7 days to 6 months (Cao *et al.* 2014; Fournier *et al.* 2014; Plichta *et al.* 2012; Sauder *et al.* 2013; Stark *et al.* 2004). Some of these studies have suggested that the test-retest reliability of amygdala activation is not good (Sauder *et al.* 2013; Stark *et al.* 2004), and worse than in relation to non-emotional neuronal activation (Cao *et al.* 2014; Plichta *et al.* 2012). This may depend on the exact method of fMRI analysis (Fournier *et al.* 2014). Of relevance to the present studies, few of these reports detailed behavioural performance of subjects across time as opposed to neuronal activity. However it has been suggested that ratings of emotional faces (Stark *et al.* 2004), and response times to them (Plichta *et al.* 2012), do not change over a one week period.

The aim of the current series of investigations was to explore the effect of repeat testing of various cognitive tasks employing emotional stimuli on task performance. In particular we were interested in the test-retest reliability of the emotional test battery. In addition we were interested in examining the test-retest reliability of an emotional attentional blink task (EABT) and the impact this might have on performance in the other tasks if added to the emotional test battery, given that the version being employed utilises pictures of faces displaying emotions, from a different, but similar, set to those used in the FERT. To date there is a study suggesting good test-retest reliability of a non-emotional attentional blink task in healthy subjects (Kranczioch and Thorne, 2013), but we were unable to find any study examining the test-retest reliability of an EABT. Likewise, no literature has been identified regarding combining the EABT with other elements of the emotional test battery, especially the FERT which also uses emotive facial expressions as stimuli. In the first experiment described below participants repeated the EABT after an hour, and then a week, after first completing it. In the second study, participants completed the emotional test battery and then repeated it a week later. The aim of the third study was to examine the effect of order of administration of task on performance in the EABT and the FERT. There were no specific hypotheses for these three studies - instead, the null hypothesis was that there is no effect of repeat testing on outcome of studies, and no effect of order of task administration.

Methods

Participants

Healthy participants for each of the studies were recruited through Newcastle University and were aged between 18 and 60 years old. Potential participants were screened to exclude those with a history of mental illness (as assessed using the Mini-International Neuropsychiatric Interview (MINI; Sheehan *et al.* 1997), those with an IQ less 90 as measured by the National Adult Reading Test (NART; Nelson, 1982), and use of psychotropic medication or recreational drugs (excluding alcohol) in the 2 months prior to screening. Participants were unique to each study; no participants took part in more than one study.

Neuropsychological Tests of Emotional Processing

Emotional Attentional Blink Task (EABT)

The "attentional blink" is most commonly studied using a rapid serial visual presentation (RSVP) paradigm in which a series of visual stimuli are presented from brief periods of around 80-100ms. In the task, participants are instructed to attend to a specified target stimulus (T1) in the RSVP and then report whether they see a second target (T2). The attentional blink refers to the phenomenon whereby the rate of T2 detection is reduced when it is presented within around 200 to 500ms of T1 (Raymond *et al.* 1992). There is evidence that T2 is more likely to be seen if it is emotionally 'salient', such as angry or fearful faces (Milders *et al.* 2006). In the current study, participants completed a variation of this attentional blink-task including emotional facial expressions. Participants were shown RSVP strings of 22 scrambled faces containing two target stimuli (T1 and T2). T1 was an emotionally neutral face (equally male and female). T1 was randomly placed as either the 4th, 5th, 6th or 7th stimulus in the RSVP string. In 75% of RSVP trials a T2 was present being either a neutral or fearful face (equal numbers of male and female faces). The time lag between T1 and T2 was either 160ms (Lag 2) or 560ms (Lag7). The faces were taken from the Karolinska set of emotional facial expressions and matched as far as possible on their intensity and arousal using data from a previous validation study (Goeleven *et al.* 2008). In each RSVP string, each stimulus was presented for 80ms with no inter-stimulus interval. At the end of each string there was a 250ms blank screen, followed by two questions with 2000ms for each response. Participants were firstly asked to respond using two designated keys on a computer keyboard as to the gender of the face at T1 and secondly the number of faces seen (1 or 2). Following responses, there was a final 250ms blank screen before the next RSVP string began. Participants were shown 16 practice strings to help explain the task.

These were followed by a 30s break before the actual task began. The number of task strings was 176 for the EABT in Study 1 and Study3. The primary outcome for the EABT was the percentage of T2 stimuli detected given correct gender identification of the T1 stimulus, at lags 2 and 7. Participants were excluded if their T1 gender identification rate was less than the group mean minus twice the group standard deviation. This was to ensure that participants were attending to the T1 stimulus (Tibboel *et al.* 2011). Including practice strings, the EABT lasted around 25 minutes.

Facial Expression Recognition Test (FERT)

Stimuli for this test were taken from Young and colleagues' modifications (Young *et al.* 1997) to the standard Ekman & Friesen pictures of facial affect (Ekman and Frieson, 1976). The emotive faces were modified to vary in intensity from 0% (appearing neutral) to 100% (full emotion) at 10% intervals. Participants were presented with a photo of one of the six basic emotions (angry, fearful, disgusted, happy, sad or surprised) as well as neutral faces. For all 6 emotions, there were 4 examples (drawn from a bank of 10 individuals in total, each contributing to at least two facial expressions, plus neutral) at 10 levels of intensity. Each face was also presented with a neutral expression, giving a total of 250 stimuli. Stimuli were presented on screen for a total of 500ms and then replaced with a blank screen. Participants were then instructed to press a key on the keyboard according to which emotion they thought they saw. The accuracy was recorded (number of faces correctly identified divided by the total number of faces showing that emotion) as well as the number of misclassifications and reaction times. A signal detection analysis was also performed on the results to allow for response tendencies to assess discriminability (Grier, 1971; Harmer *et al.* 2008). The duration of the FERT was approximately 12 minutes.

Emotional Categorisation Task

Stimuli for this task were 60 personality characteristic words, of which half were positive and half were negative as previously classified (Anderson, 1968). Each of these words was presented on screen and participants were asked to respond using the 'like' or 'dislike' keys on the keyboard according to whether they would like or dislike to be described in this way. Correct responses and reaction times were recorded. Participants were not informed that the words used in the emotional categorisation task would form part of a memory test subsequently (although in the test-retest study the emotional categorisation task was repeated a week later, so participants likely anticipated the recall task during their second visit). This task took less than 2 minutes to complete.

Emotion Recall Task

Immediately following the emotional categorisation task, participants were given two minutes to write down as many of the words from the task as they could remember. Outcomes assessed were number of words correctly recalled, number falsely recalled, and the percentage of correctly recalled words which were positive. This task took approximately 3 minutes to complete.

Emotional Recognition Task

Immediately following the emotional recall task, participants were shown 120 words on screen, 60 of which had been used in the emotional categorisation task and 60 of which were novel. The 60 novel words were personality characteristic words, again half of which were positive and half of which were negative. Participants responded using the keyboard according to whether they recognised the word from the emotional categorisation task. The number of correct positive and negative words identified, as well as the number of correct positive and negative rejections, were recorded. The outcome measure analysed was the percentage accuracy for positive and negative words calculated as the number correctly identified plus the number correctly rejected for each emotion. Reaction times were also recorded. The duration of this task was around 5 minutes.

Dot Probe

The dot probe is a computer task which aims to establish disengagement and biases in attention. Participants were shown a fixation cross for 500ms. This was removed and replaced with two distractor stimuli (above and below the location of the fixation cross). Distractor stimuli consisted of one neutral word and a positive or negative word. These were displayed for 500ms. In half of the trials, distractor stimuli were visible for 14ms and then replaced by a mask for 186ms which consisted of nonsense syllables, which were in turn replaced by the target stimuli. In unmasked trials, the distractor stimuli were removed after 500ms and replaced by the target stimuli. The target stimuli consisted of one or two stars which appeared either above or below the fixation cross, and either in the location of the emotional stimuli (congruent) or in the opposite location (incongruent). Participants were asked to press a key according to how many stars they saw. Reaction times were recorded as well as vigilance (the extent to which the emotive words drew the attention over the neutral words). Vigilance was calculated by subtracting the congruent trial reaction times from the incongruent trial reaction times.

Study 1: EABT Retest Study

20 healthy participants (10 male, mean age 23.8 years, range= 20-39, s.d. = 5.2) completed the EABT as described above (baseline), and then repeated the task one hour later and then again, one week later. All participants completed 408 trials of the task on each occasion. Participants were paid an honorarium of £30.

Study 2: Emotional Test Battery Retest Study

20 healthy participants completed study 2 (12 male, mean age 28.8 years, range = 18-52, s.d.= 10.0). This involved a first session consisting of screening and the emotional test battery (comprised of the FERT, emotional categorisation, recall and recognition tasks, and Dot Probe, in that order), and repeating the emotional test battery exactly a week later. Participants were paid an honorarium of £25.

Study 3: FERT-EABT Order Study

20 healthy participants (10 male, mean age 27.7 years, range = 19-60, s.d.= 11.7) completed the FERT and the EABT in one session. They were randomly allocated to complete either the FERT or the EABT first (10 in each group). All participants completed 220 trials of the EABT task. Participants were paid an honorarium of £10.

Procedure

In all studies participants provided written informed consent. The series of studies was approved by the Research Ethics Committee of the Faculty of Medicine, Newcastle University, UK.

Statistical analysis

Data were analysed using Statistical Package for the Social Sciences (SPSS), version 17.0. In Study 1, baseline data were first examined for the presence of an attentional blink in the EABT and an effect of emotion on this was examined using a 2 x 2 repeated measures ANOVA with factors of lag (2 and 7) and stimulus emotion (neutral and fear). As previously described, the effect of a variable on the attentional blink is indicated by an interaction between the variable, in this case emotion, and lag (MacLean and Arnell, 2012). The effect of repeat testing was then examined in an omnibus 3 x 2 x 2 ANOVA with factors of test period (baseline, one hour later and one week), lag and emotion. The analysis of data in Study 2 was similar with ANOVA factors of test period (baseline and one week) and emotion (varying levels depending on task and described in the results section). In Study 3, the EABT was analysed using a mixed measures ANOVA with factors of emotion and lag. The between subjects factor was the order in which the neuropsychological tests were administered. For any significant effect of a factor with more than one level, post-hoc ANOVAs were conducted to identify the source of the significant effect. Effects were judged to be significant if $p < 0.05$. All data is quoted as means ± standard deviations. Greenhouse-Geisser corrections were applied when ANOVAs violated Mauchly's test of sphericity (and are indicated by degrees of freedom being described to one decimal place).

In order to identify whether the changes observed in Studies 1 and 2 were consistent between individuals, a two way, mixed Intraclass Correlation (ICC) analysis of reliability was performed. In accordance with Fleiss (1986), reliability was considered ‘excellent’ if the ICC coefficient was greater than 0.75 and ‘fair to good’ if between 0.4 and 0.75 (Fleiss, 1986). Confidence intervals and full details of ICC values are listed in Table 1.

Results

Study 1

One participant was excluded on the criteria that their rate of correct identification of T1 gender (56%) was more than two standard deviations below the mean (82%). Of the remaining 19 participants, the mean IQ was 114.2 (range 105-120, s.d. = 4.7). At baseline, as can be seen in figure 1A, detection of T2 stimuli in the EABT was greater at lag 7 compared to lag 2 and when T2 was a fearful compared to a neutral face. A preliminary repeated measures ANOVA of baseline data revealed a significant effect of lag ($F(1, 18) = 59.97$, $p < 0.001$, $\eta_p^2 = 0.77$, 95% CI 0.51, 0.86) and an emotion by lag interaction ($F(1, 18) = 8.13$, $p = 0.011$, $\eta_p^2 = 0.31$, 95% CI 0.02, 0.55) indicative of an attentional blink per se, and that this is modified by emotion. An omnibus ANOVA of T2 detection rates across all three test periods confirmed a lag by emotion interaction ($F(1, 18) = 21.63$, $p < 0.001$, $\eta_p^2 = 0.55$, 95% CI 0.19, 0.71). In addition there was a significant main effect of test period ($F(2, 36) = 7.05$, $p = 0.003$, $\eta_p^2 = 0.28$, 95% CI 0.05, 0.46) but no test period by lag interaction ($F(1.3, 23.9) = 2.09$, $p = 0.138$, $\eta_p^2 = 0.10$, 95% CI 0.00, 0.33). As can be seen in figures 1B and 1C, where T2 detection rates just for neutral and fearful T2 stimuli are shown respectively, this significant effect of test period was due to performance in the EABT improving at one hour, with little further change at one week, compared to baseline. This was confirmed by post-hoc ANOVA that revealed a significant effect of test period ($F(1, 18) = 13.07$, $p = 0.002$ $\eta_p^2 = 0.42$, 95% CI 0.08, 0.63) when comparing baseline with one hour later, but a lack of effect when comparing one hour with one week ($F(1, 18) < 0.001$, $p = 0.997$, $\eta_p^2 = 0.00$, 95% CI 0.00, 0.00).

Figure 1 near here

While emotion of T2 and test period were both significant in the omnibus ANOVA, there was no test period by emotion ($F(2, 36) = 0.791$, $p = 0.461$, $\eta_p^2 = 0.04$, 95% CI 0.00, 0.18) or test period by emotion by lag ($F(1.6, 28.0)=0.183$, $p=0.779$, $\eta_p^2 = 0.01$, 95% CI 0.00, 0.13) interaction, demonstrating that while overall performance in the EABT improved with repeat testing, the effect of a fearful face at T2 was not altered. This is illustrated in figure 1D that shows the difference in T2 detection rates for fear and neutral at lags 2 and 7 over the three test periods.

Reliability for T2 detection between baseline and one hour was either fair or excellent, ranging from an ICC coefficient of 0.68 for neutral images at lag 7 to 0.83 for fear faces at Lag 2. Between one hour and one week, ICC coefficients were also between fair and excellent, ranging from 0.58 for fear faces at lag 7 to 0.91 for neutral faces at Lag 2 (see Table 1 for full results). Performance on the NART was found to correlate only with Neutral-Neutral identification at week 1.

Table 1 near here

Study 2

The mean IQ of participants in Study 2 was 115.9 (range 106-127, s.d = 6.3). NART scores were not found to correlate with any of the outcome measures for this study.

FERT

One participant was excluded from the analysis as they did not correctly identify any faces as being ‘Disgust’. This left 19 participants in the FERT analysis. Overall, participants achieved a higher discrimination index on their second test period (figure 2A). Repeated measures ANOVA with within subject factors of test period (baseline and 1 week) and stimulus emotion (neutral, anger, disgust, fear, happy, sad and surprised) was conducted. This showed significant effects for test period ($F(1, 18) = 7.43$, $p = 0.014$, $\eta_p^2 = 0.29$, 95% CI 0.01, 0.54) and emotion ($F(2.5, 44.2) = 16.40$, $p < 0.001$, $\eta_p^2 = 0.48$, 95% CI 0.24, 0.61) but there was no test period by emotion interaction ($F(3.8, 68.9) = 0.219$, $p = 0.921$, $\eta_p^2 = 0.012$, 95% CI 0.00, 0.05). Post hoc paired t-tests showed that there were significant improvements in discrimination index for disgust ($t(18) = -2.10$, $p = 0.050$, $d = -0.99$) and happy ($t(18) = -3.62$, $p = 0.002$, $d = -1.7$) faces at the second test period compared to the first. There was no significant difference for any of the other emotions. To further examine if there was an effect of repeat testing on the effect of emotion in the FERT, the difference between the discrimination index for neutral faces and emotive faces was calculated (figure 2B). Repeated measures ANOVA showed a significant main effect of emotion ($F(2.8, 51.1) = 19.16$, $p < 0.001$, $\eta_p^2 = 0.52$, 95% CI 0.29, 0.63), but not of test period ($F(1, 18) = 0.140$, $p = 0.713$, $\eta_p^2 = 0.01$, 95% CI 0.00, 0.20) or test period by emotion interaction ($F(3.1, 54.9) = 0.242$, $p = 0.870$, $\eta_p^2 = 0.01$, 95% CI 0.00, 0.07).

Figure 2 near here

There was a general trend towards fewer misclassifications in the FERT at one week (figure 2C). Repeated measures ANOVA showed significant effects for test period ($F(1, 18) = 7.84$, $p = 0.012$, $\eta_p^2 = 0.30$, 95% CI 0.02, 0.55), emotion ($F(3.5, 62.6) = 206.27$, $p < 0.001$, $\eta_p^2 = 0.92$, 95% CI 0.88, 0.94) and an emotion by test period interaction ($F(2.6, 47.2) = 2.24$, $p = 0.045$, $\eta_p^2 = 0.11$, 90% CI 0.00, 0.22). Post hoc analysis indicated that this latter finding was driven by significantly lower rates of misclassification of anger (t(18) = 2.22 , p = 0.039, d = 1.05), and a trend for neutral faces (t(18) = 1.98, p = 0.063, d = 0.93), at week 1 compared to baseline. There was no significant difference between test periods for any of the other emotions.

Participants responded faster to stimuli in the FERT at the second time period compared to baseline (figure 2D). Repeated measures ANOVA showed significant effects of test period (F(1, 18) = 15.78, p < 0.001, η_p^2 = 0.47, 95% CI 0.11, 0.66), emotion (F(2.6, 47.5) = 8.00, p < 0.001, η_p^2 = 0.31, 95% CI 0.08, 0.46), but there was no emotion by test period interaction (F(3.3, 59.4) = 1.8, p = 0.104, η_p^2 = 0.10, 95% CI 0.00, 0.21). Paired t-tests showed that this was because reaction times were less at the second time period for all emotions: anger (t (18) = 2.73, p = 0.014, d = 1.29); disgust: (t(18) = 2.34, p = 0.031, d = 1.10); fear: (t(18) = 2.77, p = 0.013, d = 1.3); happy: (t(18) = 2.43, p = 0.026, d = 1.15); sad: (t(18) = 3.82, p = 0.001, d = 1.8); surprise: (t(18) = 3.64, p = 0.002, d = 1.72); but with the exception of neutral facial stimuli (t(18) = 0.875, p = 0.393, d = 0.41).

The reliability for the discrimination scores was fair to excellent for the six emotions, ranging from an ICC coefficient of 0.58 for anger faces to 0.98 for disgust faces (see table 1). Neutral faces however had a lower reliability coefficient of 0.36. Misclassifications for the FERT were also analysed for reliability, with neutral face misclassifications being the least reliable with an ICC coefficient of 0.59 and sad faces being the most reliable with 0.87.

Emotional Categorisation Task

All participants' data was included in the analysis. emotional categorisation task responses were analysed using a repeated measures ANOVA including within subject factors of test period and emotion (positive and negative). With regard to emotion labelling, performance neared maximal levels on both test periods (scores out of 30 of 29.0 ± 0.9 for positive and 28.5 ± 1.9 for negative words at baseline and 29.3 ± 0.9 for positive and 28.9 ± 1.5 for negative words at 1 week). There was no significant effects of emotion (*F*(1, 19) = 1.45, p = 0.243, η_p^2 = 0.07, 95% CI 0.00, 0.33), test period (*F*(1, 19) = 2.16, p = 0.158, η_p^2 = 0.10, 95% CI 0.00, 0.36) or test period by emotion interaction (*F*(1, 19) = 0.00, p = 0.999, $\eta_p^2 < 0.001$, 95% CI 0.00, 0.00). With regard to emotional categorisation task response times (figure 3A), there was likewise no significant effect of test period (*F*(1, 19) = 0.38, p = 0.545, η_p^2 = 0.02, 95% CI 0.00, 0.24), emotion (*F*(1, 19) = 2.01, p = 0.173, η_p^2 = 0.10, 95% CI 0.00, 0.36) or test period by emotion interaction (*F*(1, 19) = 0.13, p = 0.726, η_p^2 = 0.01, 95% CI 0.00, 0.19). Categorization of positive words had low reliability with an ICC score of 0.23, whereas negative words were more reliable with a 'fair' ICC score of 0.60.

Emotional Recall Task

All participants' data was included in the analysis. As can be seen in figure 3B, and confirmed by repeated measures ANOVA, participants recalled more positive than negative words on both test periods (significant effect of emotion: (*F*(1, 19) = 5.40, p = 0.031, η_p^2 = 0.22, 90% CI 0.01, 0.44) and recalled more words at week one compared to baseline (significant effect of test period: (*F*(1, 19) = 26.78, $p < 0.001$, η_p^2 = 0.59, 95% CI 0.24, 0.74). However, there was no emotion by test period interaction (*F*(1, 19) = 0.62, p = 0.442, η_p^2 = 0.03, 95% CI 0.00, 0.26) due to the increase in recall between baseline and week 1 of positive and negative words being similar. Post hoc analysis comparing the proportion of correctly identified words which were positively valenced showed no difference between baseline and week 1 (t(19) = -0.466, p = 0.646, d = -0.21). Reliability was fair for the emotional recall task with an ICC score of 0.54.

Figure 3 near here

Emotional Recognition Task

All participants' data was included in the analysis. As shown in figure 3C, participants were more accurate at week 1 than baseline, and with positive compared to negative words. Repeated measures ANOVA showed significant effects of test period (*F*(1, 19) = 9.95, $p = 0.005$, $\eta_p^2 = 0.34$, 95% CI 0.04, 0.57), emotion (*F*(1, 19) = 5.19, $p = 0.034$, $\eta_p^2 = 0.22$, 90% CI 0.01, 0.43) but no test period by emotion interaction (*F*(1, 19) = 0.45, $p = 0.512$, $\eta_p^2 = 0.02$, 95% CI 0.00, 0.25). With regard to reaction times in the emotional recognition task (figure 3D) repeated measures ANOVA showed a significant effect of emotion (*F*(1, 19) = 34.89, $p < 0.001$, $\eta_p^2 = 0.65$, 95% CI 0.32, 0.78), but no effect of test period (*F*(1, 19) = 1.34, $p = 0.261$, $\eta_p^2 = 0.07$, 95% CI 0.00, 0.32) or test period by emotion interaction (*F*(1, 19) = 2.57, $p = 0.125$, $\eta_p^2 = 0.12$, 95% CI 0.00, 0.38). Reliability for the emotional recognition was good with an ICC score of 0.75.

Dot Probe

Repeated measures ANOVA of the vigilance scores showed no significant effects of test period (*F*(1, 19) = 1.73, $p = 0.204$, $\eta_p^2 = 0.08$, 95% CI 0.00, 0.34), emotion (*F*(1, 19) = 0.58, $p = 0.454$, $\eta_p^2 = 0.03$, 95% CI 0.00, 0.26) or emotion by test period interaction (*F*(1, 19) = 0.194, $p = 0.664$, $\eta_p^2 = 0.01$, 95% CI 0.00, 0.21). Results for reaction times also showed no significant effects of test period (*F*(1, 19) = 2.52, $p = 0.129$, $\eta_p^2 = 0.12$, 95% CI 0.00, 0.38), emotion (*F*(1, 19) = 1.03, $p = 0.322$, $\eta_p^2 = 0.05$, 95% CI 0.00, 0.30) or test period by emotion interaction (*F*(1, 19) = 2.00, $p = 0.174$, $\eta_p^2 = 0.10$, 95% CI 0.00, 0.36). ICC analysis also showed low levels of reliability between participants' results from one week compared to the next, with ICC scores ranging from -0.04 for positive, masked stimuli to 0.13 for negative unmasked stimuli (see table for full results and confidence intervals).

Study 3

All participants' data was included in the analysis. The mean IQ for participants in Study 3 as measured by the NART was 116.2 (range 105-126, s.d. = 6.4). There were no significant differences in IQ between the two participant groups. Repeated measures ANOVA of the FERT data, including the between subject factor of order of tasks, showed significant effects of stimulus emotion ($F(3, 38) = 14.89$, $p < 0.001$, $\eta_{p}^{2} = 0.54$, 95% CI 0.27, 0.66) and a trend towards an emotion by task order interaction ($F(3.4, 60.7) = 2.28$, $p = 0.081$, $\eta_{p}^{2} = 0.11$, 95% CI 0.00, 0.23). Figure 4A shows the FERT discrimination index results for the participants in the two groups: those who completed the FERT before the EABT and those who completed the tasks in the reverse order. Order had a significant effect on the discrimination index for neutral faces (t(18) = 2.23, $p = 0.039$, $d = 1.05$) but not anger (t(18)= -0.766, $p = 0.453$, $d = -0.36$), disgust (t(18) = -0.317, $p = 0.755$, $d = -0.15$), fear (t(18) = -0.733, $p = 0.473$, $d = -0.35$), happy (t(18) = -0.109, $p = 0.914$, $d = -0.05$), sad (t(18) =1.301, $p = 0.210$, $d = 0.61$) or surprise (t(18) < 0.001, $p = 1.000$, $d < 0.001$).

A similar repeated measures ANOVA of the EABT data (figure 4B) showed significant effects of emotion ($F(1, 18) = 37.57$, $p < 0.001$, $\eta_{p}^{2} = 0.68$, 95% CI 0.35, 0.80), lag ($F(1,18) = 20.12$, $p < 0.001$, $\eta_{p}^{2} = 0.53$, 95% CI 0.17, 0.70) and lag by emotion interaction ($F(1, 18) = 10.40$, $p = 0.005$, $\eta_{p}^{2} = 0.37$, 95% CI 0.04, 0.59), but no effect of order of administration of task on lag or emotion. IQ as measured by the NART was found to correlate only with the discrimination index for sad faces ($r = 0.618$, $p = 0.004$). No other correlation between NART and any other outcome variable was found in this study.

Discussion

This series of studies provides important data relating to the test-retest reliability of emotional cognitive tests in healthy participants. Study 1 showed that repeat testing of an EABT using emotional faces results in an improvement in overall task performance in terms of a reduction in the magnitude of the attentional blink. This represents an improved ability to re-orientate to a second stimulus presented shortly after a stimulus to which the participants are attending. This improvement was found one hour after the baseline assessment with no further improvement one week later. In contrast to the overall performance on the task, the impact of stimulus emotion on the magnitude of the attentional blink was not altered by repeat testing. Study 2 examined an emotional test battery including a range of tasks employing visual stimuli of emotional faces and emotional words and revealed similar findings; in general, performance on facial expression recognition and memory tasks was improved in terms of accuracy and/or response times when tasks were repeated after one week. However, again there was no evidence of repeat testing resulting in changes in the effect of stimulus emotion on the task. Study 3 demonstrated that the order of administration of two tasks (FERT and EABT) both employing facial expression visual stimuli (from different stimuli sets) had no consistent influence on performance on either.

The ICC scores (see table 1) across time periods for Studies 1 and 2 was generally high, suggesting a good to high retest reliability level for these measures. This is consistent with what little behavioural data there is for performance on the FERT in fMRI studies (Plichta *et al.* 2012; Stark *et al.* 2004). The only exception to this pattern was the Dot Probe task, which did not yield any significant effects of stimulus emotion or test period. Previous studies (Schmukle, 2005) have also found the task to have low reliability, with performance possibly being related to current mood state.

It is worth noting that the majority of studies investigating habituation of amygdala response to emotional stimuli discussed in the introduction found changes that occur over a time scale of seconds or minutes. The current research has investigated the considerably longer period of 7 days in studies 1 and 2. The results from these studies suggest that although repeat testing may increase overall performance on the tasks used, the effect of stimulus emotion on performance was not altered. This research complements that of Britton (Britton *et al.* 2008) who found evidence of recovery of amygdala habituation to emotional stimuli after just 20 minutes, as well as studies of neuropsychological testing (Calamia *et al.* 2012) which provides evidence for practice effects on neuropsychological tasks being maintained over periods of days or weeks. It is therefore possible that learning is occurring in neural networks responsible for generalised cognitive functions rather than habituation to emotional stimuli.

The strengths and weaknesses of the current studies should be highlighted. These studies have investigated an area previously neglected in the scientific literature, attaining high rates of reliability and with some consistency between the results of studies 1 and 2. However the results from these studies may only pertain to the specific emotional processing tasks that were conducted and may not be generalizable to other tasks or emotions. The studies only included behaviour measures with no assessment of underlying neural activity. As such, it is not possible to comment on whether the lack of effect of repeat testing on the effects of stimulus emotion in the tasks studied might be mirrored on a neural level with a lack of habituation in amygdala activity. However, the retest effects on the tasks studied were generally evident in terms of overall performance rather than changes in the impact of stimulus emotion on performance. The studies conducted were all in healthy volunteers. It is important not to extrapolate the data without care into patient populations where there is evidence of differential habituation effects compared to healthy controls (Sladky *et al.* 2012).

It is also important to state that, given the relatively small sample sizes for each of these studies, care must be taken in interpreting the lack of effect of retesting on the effect of emotional stimuli, or the lack of effect of task order. With regards to the EABT, the numerically largest difference from the baseline measure was at one hour for the lag 2 condition. The mean difference in the effect of fear, one hour minus baseline was 2.94% (s.d. = 16.18) giving a Cohen's effect size of d = 0.18. This suggests that if there is an effect of repeat testing on the effect of emotion in the EABT a sample size of 320 would be needed to detect this with a power of 90% and an α of 0.05 (G-power 3.1). With regards to the FERT, figure 2D suggests some possible difference in the effects of emotion on response times over time, especially for sad faces. Subtracting the response time for neutral from that for sad faces, the difference between this measure at baseline and repeat testing at one week is 214 seconds (s.d. = 566) giving d = 0.38 and a required sample size to show a significant effect of 76 (same parameters as above). These two effect size and power calculations suggest that while we can't rule out the possibility that we are falsely accepting the null hypothesis (that there is no effect of repeat testing on the effect of stimulus emotion), this seems unlikely.

These data suggest therefore that it is feasible to administer cognitive tasks employing emotional stimuli on repeated occasions without an alteration in the effect of stimulus emotion on the task, and that order of tasks within a session is not critical. It also suggests that emotional tasks have good reliability though care may need to be taken in the use of the dot-probe measure. These results open up new possibilities in terms of repeat testing and longitudinal studies, such as studying an individual's processing both before and after a depressive episode, or before and after treatment with an antidepressant. The practice effects seen would suggest that it may be of benefit for participants to practise some tasks in order to maximise their performance before implementing an intervention. An example may be the attentional blink task, where performance on the task itself improves after the first testing but then remains stable, while the effect of emotion doesn't change. However, caution should be observed as practice may result increasingly fixed performance, reducing the tendency for any experimental manipulations to alter outcomes.

Conclusion

There was a reasonably consistent improvement in accuracy and response times in the tests examined, which is likely due to an effect of practice. It is interesting to see that there is no consistent pattern of changes in the effect of stimulus emotion on attention, facial recognition or memory over time, nor evidence of clear interaction between stimuli used in different tasks. This does suggest that it is feasible to use multiple emotional tasks during a single testing session, and to undertake repeat testing, at least over the time periods investigated here. The order of administration was not found to influence the performance on tests including emotional facial stimuli.

Acknowledgements

Sally Preston and Tom Adams were both funded through Research Capability Funding (RCF) from Northumberland, Tyne and Wear NHS Foundation Trust.

Figure Legends

Figure 1. Study 1: effect of repeat testing on the emotional attentional blink task (EABT). **A.** Percentage detection rates of the second target stimulus (T2) at lags 2 and 7 at baseline. T2, when present, was either a neutral or fearful face. **B**. Percentage detection of neutral T2 stimuli at baseline and when the EABT was repeated after 1 hour and one week. **C**. Percentage detection of fearful T2 stimuli at baseline and when the EABT was repeated after 1 hour and 1 week. **D**. The difference between percentage detection of fearful T2 and neutral T2 stimuli (a measure of the effect of a fearful stimuli on task performance) at baseline and when the EABT was repeated after 1 hour and 1 week. All data plotted as means with error bars representing the standard error of the mean.

Figure 2. Study 2: effect of repeat testing on the facial expression recognition test (FERT) element of the emotional test battery . **A**. Discrimination index for the identification neutral and each of the six standard emotions, plotted at baseline and when the task was repeated one week later. **B**. The difference in discrimination index of each of the six emotions minus the discrimination index for neutral, to illustrate the effect of each emotion on recognition, plotted for baseline and when repeated one week later. **C**. Number of misclassifications of each emotion at baseline and one week later. **D**. Response times when responding to each of the emotional stimuli at baseline and after one week. All data plotted as means with error bars representing the standard error of the mean.

Figure 3. Study 2: effect of repeat testing on the emotional categorisation, recall and recognition task elements of the emotional test battery. A. Response times in the emotional categorisation task for positive and negatively valenced words, plotted at baseline and when the task was repeated one week later. **B**. Number of positive and negatively valenced words recalled in the emotional recall task, plotted at baseline and when the task was repeated one week later. **C**. Percentage accuracy in the emotional recognition task for positive and negatively valenced words, plotted at baseline and when the task was repeated one week later. **D.** Response times in the emotional recognition task for positive and negatively valenced words, plotted at baseline and when the task was repeated one week later. All data plotted as means with error bars representing the standard error of the mean.

Figure 4. Study 3: effect of task order when the EABT and FERT were run consecutively with the order randomly determined for each subject. **A**. FERT discrimination index plotted for neutral and each of the six standard emotions, for each of the two orders of tasks. **B**. Neutral and fearful T2 stimulus detection in the EABT, for each of the two orders of tasks. All data plotted as means with error bars representing the standard error of the mean.

References

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *J. Pers. Soc. Psychol.* **9**, 272-279.

Arnone, D., Horder, J., Cowen, P. J. & Harmer, C. J. (2009). Early effects of mirtazapine on emotional processing. *Psychopharmacology (Berl)* **203**, 685-691.

Bourke, C., Douglas, K. & Porter, R. (2010). Processing of facial emotion expression in major depression: a review. *Australian & New Zealand Journal of Psychiatry* **44**, 681-696.

Britton, J. C., Shin, L. M., Barrett, L. F., Rauch, S. L. & Wright, C. I. (2008). Amygdala and fusiform gyrus temporal dynamics: responses to negative facial expressions. *BMC. Neurosci.* **9**, 44.

Calamia, M., Markon, K. & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* **26**, 543-570.

Cao, H., Plichta, M. M., Schafer, A., Haddad, L., Grimm, O., Schneider, M., Esslinger, C., Kirsch, P., Meyer-Lindenberg, A. & Tost, H. (2014). Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. *Neuroimage.* **84**, 888-900.

Dere, E., Pause, B. M. & Pietrowsky, R. (2010). Emotion and episodic memory in neuropsychiatric disorders. *Behavioural Brain Research* **215**, 162-171.

Ekman, P. & Frieson, W. (1976). *Pictures of facial affect.* Consulting Psychologists Press: Palo Alto, CA.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments.* John Wiley Sons: New York.

Fournier, J. C., Chase, H. W., Almeida, J. & Phillips, M. L. (2014). Model specification and the reliability of FMRI results: implications for longitudinal neuroimaging studies in psychiatry. *PLoS. One.* **9**, e105169.

Goeleven, E., De Raedt, R., Leyman, L. & Verschuere, B. (2008). The Karolinska directed emotional faces: A validation study. *Cognition and Emotion* **22**, 1094-1118.

Gotlib, I. H., Krasnoperova, E., Yue, D. N. & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of Abnormal Psychology* **113**, 121-135.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psychological Bulletin* **75**, 424-429.

Harmer, C. J., de, B. C., Dawson, G. R., Dourish, C. T., Waldenmaier, L., Adams, S., Cowen, P. J. & Goodwin, G. M. (2011). Agomelatine facilitates positive versus negative affective processing in healthy volunteer models. *J. Psychopharmacol.* **25**, 1159-1167.

Harmer, C. J., Goodwin, G. M. & Cowen, P. J. (2009). Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *Br. J. Psychiatry* **195**, 102-108.

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.

Ishai, A., Pessoa, L., Bikle, P. C. & Ungerleider, L. G. (2004). Repetition suppression of faces is modulated by emotion. *Proc. Natl. Acad. Sci. U. S. A* **101**, 9827-9832.

Kranczioch, C. & Thorne, J. D. (2013). Simultaneous and preceding sounds enhance rapid visual targets: Evidence from the attentional blink. *Adv. Cogn Psychol.* **9**, 130-142.

MacLean, M. H. & Arnell, K. M. (2012). A conceptual and methodological framework for measuring and modulating the attentional blink. *Atten. Percept. Psychophys.* **74**, 1080-1097.

Milders, M., Sahraie, A., Logan, S. & Donnellon, N. (2006). Awareness of faces is modulated by their emotional meaning. *Emotion.* **6**, 10-17.

Nelson, H. E. (1982). *The National Adult Reading Test (NART): Test Manual.* NFER Publishing Co., Windsor, England.

Pavuluri, M. N., Passarotti, A. M., Fitzgerald, J. M., Wegbreit, E. & Sweeney, J. A. (2012). Risperidone and divalproex differentially engage the fronto-striato-temporal circuitry in pediatric mania: a pharmacological functional magnetic resonance imaging study. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 157-170.

Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A. B., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P. & Meyer-Lindenberg, A. (2012). Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *Neuroimage.* **60**, 1746-1758.

Protopopescu, X., Pan, H., Tuescher, O., Cloitre, M., Goldstein, M., Engelien, W., Epstein, J., Yang, Y., Gorman, J., Ledoux, J., Silbersweig, D. & Stern, E. (2005). Differential time courses and specificity of amygdala activity in posttraumatic stress disorder subjects and normal control subjects. *Biological Psychiatry* **57**, 464-473.

Raymond, J. E., Shapiro, K. L. & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* **18**, 849-860.

Sauder, C. L., Hajcak, G., Angstadt, M. & Phan, K. L. (2013). Test-retest reliability of amygdala response to emotional faces. *Psychophysiology* **50**, 1147-1156.

Schmukle, S. (2005). Unreliability of the dot probe task. *European Journal of Personalit* **19**, 595-605.

Sheehan, D. V., Lecrubier, Y., Sheitman, B., Janavs, J., Weiller, E., Keskiner, A., Schinka, J., Knapp, E., Sheehan, M. F. & Dunbar, G. C. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry* **12**, 232-241.

Sladky, R., Hoflich, A., Atanelov, J., Kraus, C., Baldinger, P., Moser, E., Lanzenberger, R. & Windischberger, C. (2012). Increased neural habituation in the amygdala and orbitofrontal cortex in social anxiety disorder revealed by FMRI. *PLoS. One.* **7**, e50050.

Stark, R., Schienle, A., Walter, B., Kirsch, P., Blecker, C., Ott, U., Schafer, A., Sammer, G., Zimmermann, M. & Vaitl, D. (2004). Hemodynamic effects of negative emotional pictures - a test-retest analysis. *Neuropsychobiology* **50**, 108-118.

Strauss, M. M., Makris, N., Aharon, I., Vangel, M. G., Goodman, J., Kennedy, D. N., Gasic, G. P. & Breiter, H. C. (2005). fMRI of sensitization to angry faces. *Neuroimage.* **26**, 389-413.

Tibboel, H., Van, B. B. & De, H. J. (2011). Is the emotional modulation of the attentional blink driven by response bias? *Cogn Emot.* **25**, 1176-1183.

Tranter, R., Bell, D., Gutting, P., Harmer, C., Healy, D. & Anderson, I. M. (2009). The effect of serotonergic and noradrenergic antidepressants on face emotion processing in depressed patients. *Journal of Affective Disorders* **118**, 87-93.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A. & Perrett, D. I. (1997). Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition* **63**, 271-313.