Characterisation of facial emotion recognition in bipolar disorder: focus on emotion mislabelling and neutral expressions

Short title: *Facial emotion recognition in bipolar disorder*

Lisa S. Furlong[1], Susan L. Rossell[2,3], James A. Karantonis[1,2], Vanessa L. Cropley[1], Matthew Hughes[2] and Tamsyn E. Van Rheenen[1,2*]

Affiliations

[1]Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne, Melbourne, Australia

[2]Centre for Mental Health, Faculty of Health, Arts and Design, School of Health Sciences, Swinburne University, Melbourne, Australia

[3]Department of Psychiatry, St Vincent's Hospital VIC, Australia

*Correspondence to: Tamsyn E. Van Rheenen, Level 3, Alan Gilbert Building, 161 Barry Street, Carlton, Victoria 3053, Australia. Email: tamsyn.van@unimelb.edu.au

Abstract:

Increasing evidence suggests that facial emotion recognition is impaired in bipolar disorder (BD). However, patient-control differences are small, owing to ceiling effects on the tasks used to assess them. The extant literature is also limited by a relative absence of attention toward identifying patterns of emotion misattribution or understanding whether neutral faces are mislabelled in the same way as ones displaying emotion. We addressed these limitations by comparing facial emotion recognition performance in BD patients and healthy controls on a novel and challenging task. Thirty-four outpatients with BD I and 32 demographically-matched healthy controls completed a facial emotion recognition task requiring the labelling of neutral and emotive faces displayed at low emotional intensities. Results indicated that BD patients were significantly less accurate at labelling faces than healthy controls, particularly if they displayed fear or neutral expressions. There were no between-group differences in response times or patterns of emotion mislabelling, with both groups confusing sad and neutral faces, though BD patients also mislabelled sad faces as angry. Task performance did not significantly correlate with mood symptom severity in the BD group. These findings suggest that facial emotion recognition impairments in BD extend to neutral face recognition. Emotion misattribution occurs in a similar, albeit exaggerated manner in patients with BD compared to healthy controls. Future behavioural and neuroimaging research should reconsider the use of neutral faces as baseline stimuli in their task designs.

Keywords:
Facial expression; bipolar disorder; social perception; cognition; affect; emotion processing

Data availability statement:
Research data are not shared.

Acknowledgements:
The authors would like to thank all the participants who took part in this study. This work was supported by the National Health and Medical Research Council [grant numbers 1088785, 1154651, 1177370]; the National Imaging Facility; Australian Rotary Health; the Ian Parker Bipolar Research Fund; Brunslea Park Estate; Swinburne University of Technology; and the University of Melbourne. The authors acknowledge the facilities and scientific assistance of the National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability, at the Swinburne Neuroimaging Facility, Swinburne University of Technology.

Introduction

Bipolar disorder (BD) is a severe psychiatric illness that is characterised by mood dysregulation and psychosocial dysfunction (American Psychiatric Association, 2013; Van Rheenen and Rossell, 2014a). Within the disorder's cognitive profile, social cognitive deficits are increasingly recognised as a feature (Miskowiak et al., 2019; Van Rheenen et al., 2019; Van Rheenen and Rossell, 2013). In particular, previous research shows that BD patients are less able to identify facial expressions when compared to healthy controls in terms of both accuracy and speed, especially for negative emotions (Miskowiak et al., 2019; Van Rheenen et al., 2017; Van Rheenen and Rossell, 2013; Vederman et al., 2012). This facial emotion recognition impairment appears to have trait relevance, since it has been reported in individuals with the disorder across symptomatic mood states (Getz et al., 2003; Lembke and Ketter, 2002; Soeiro-de-Souza et al., 2012) including psychosis (Daros et al., 2014), and also during euthymia (Bozikas et al., 2006).

Currently, the specific pattern of emotion misattribution in BD is unclear, as only a small number of studies have explicitly examined which emotions certain facial expressions are mistaken for. Within these studies, one reported that individuals with BD mislabelled angry facial expressions as fearful compared to healthy controls (Goghari and Sponheim, 2013). Two other studies reported that BD patients mislabelled sad facial expressions as fearful (Derntl et al., 2009) or happy (Yalcin-Siedentopf et al., 2014), and mislabelled disgusted expressions as surprised or angry (Yalcin-Siedentopf et al., 2014).

Studies examining the accuracy of neutral face recognition in BD are also very limited. Neutral facial expressions are often used as a baseline condition in facial emotion recognition tasks in which there is an implicit assumption that they are accurately recognised as displaying no emotion. Thus, it is important to determine whether individuals with BD do not ascribe emotion to faces in which there is none. A small number of studies have sought to examine this, with findings suggesting that individuals with BD may mislabel neutral stimuli as sad more often than healthy controls (Lee and Van Meter, 2020; Ruocco et al., 2014). Neutral faces have also been mislabelled as happy (Ruocco et al., 2014), and in paediatric BD, as hostile (Rich et al., 2006). However, these studies were either relatively small in size (Rich et al., 2006) or examined only one BD clinical state such as psychotic BD (Ruocco et al., 2014). As such, the trait-like nature of the impairment could not be clearly determined.

Further, while past studies of facial emotion recognition have detected differences between patients and healthy controls, the reported effects tend to be small. It has been argued that facial emotion recognition tasks containing prototypical 'full-blown' facial emotions may be prone to ceiling effects (Kessels et al., 2014). Further, some evidence suggests that BD patients require greater intensities of emotional expression to identify facial expressions compared to healthy controls (Brotman et al., 2008; Schaefer et al., 2010). Thus, the impairment in facial emotion recognition may be quite mild; and more robust effects may be observed when the performance of BD and healthy control groups are compared on tasks that use less intense emotional expressions as stimuli — as is arguably more ecologically valid — because they may be more challenging for individuals with BD.

In the current study, we addressed these limitations, by comparing BD patients to healthy controls on a novel and challenging facial emotion recognition task requiring the labelling of affective stimuli (fearful, angry, sad) displayed at low emotional intensities. We included neutral faces in all of our analyses, as this facial emotion condition has been relatively less studied in BD. Inbuilt into the task was a control measure assessing the processing of inanimate objects to determine the extent to which general visual processing of non-facial stimuli was similar between groups. This also permitted the analysis of neutral faces as an experimental emotion condition unto itself (rather than using the neutral face condition as a reference/control).

The specific aims of the study were to: i) replicate previous research showing that BD is associated with facial emotion labelling deficits related to task accuracy and response time when compared to healthy controls; ii) determine the pattern and extent to which certain emotions were being mislabelled for other emotions in each group; and iii) assess the relationship between facial emotion recognition and mood symptoms to explore whether potential deficits occurred independent of mood state in the BD cohort. We hypothesised that the BD group would demonstrate lower accuracies and increased response times compared to healthy controls for the labelling of facial affect (Hypothesis 1), though the assessment of impairments in the particular emotions included in the task (fearful, angry, sad and neutral), and the order in which performance decrements occurred was exploratory in nature (Research Question 1). We also explored the within- and between-group pattern of emotion mislabelling in BD (Research Question 2), and whether there were between-group differences in the labelling of non-facial control stimuli (Research Question 3). Finally, as previous research suggests that facial emotion recognition deficits may be a trait-like feature of BD (Bozikas et al., 2006; Van Rheenen and Rossell, 2014a; Vederman et al., 2012), we explored whether there were any significant associations between mood symptoms and task accuracy, response time, or emotion mislabelling in BD (Research Question 4).

Methods

This study complied with the ethical standards of the relevant Human Ethics Review Boards and with the Declaration of Helsinki.

Participants

This cross-sectional study comprised 66 participants, including *n* = 32 healthy controls and *n* = 34 outpatients with a DSM-IV-TR diagnosis of BD type 1 as confirmed by the Mini International Neuropsychiatric Interview (MINI; Sheehan et al., 1998). Participants were recruited as part of a neuroimaging study, the results of which are not the focus of this paper as we aimed to specifically assess the *behavioural* characteristics of this cohort in detail. Participants with BD were all clinically stable outpatients at the time of testing, with none meeting criteria for a mood episode at the time of assessment or in the three weeks prior. Current mood symptoms were assessed using the Young Mania Rating Scale (YMRS; Young et al., 1978), and Montgomery-Asberg Depression Rating Scale (MADRS; Montgomery and Asberg, 1979), from which n = 24 (70.6%) BD participants were considered to be affectively stable, as defined by MADRS scores of ≤12 (Calafiore et al., 2018) and YMRS scores of ≤10 (Keck et al., 2006). The remaining 10 BD participants (29.4%) displayed mild-moderate symptoms (i.e., 8 with MADRS scores >12 and 2 with YMRS scores >10). The number of patients using mood-stabilisers, antipsychotics and antidepressants at the time of assessment was also recorded.

Healthy control participants were matched as closely as possible to BD participants in terms of age. Control participants were assessed using the MINI to confirm no current or prior diagnosis of psychiatric illness. In all participants, the Wechsler Test of Adult Reading (WTAR; Wechsler, 2001) was used to estimate the premorbid intelligence quotient (IQ). Exclusion criteria for all participants included; a history of neurological and neurodegenerative conditions, visual and hearing impairments, history of severe head injury, substance abuse/dependence in the past three months, and difficulties with written or spoken English. Furthermore, BD participants were excluded if they had non-trivial medication changes in the two months prior to assessment, or benzodiazepine use in the 48 hours prior. Healthy controls were excluded if they had a personal or family history of psychiatric illness, or current/previous psychotropic medication use. Participants were recruited via pre-existing databases held by the research team, as well as general advertisements, social media and community support groups.

Facial emotion labelling task

The facial emotion recognition task was designed by the senior author (TVR) and completed by each participant within the context of a functional magnetic resonance imaging (MRI) scan, where only the behavioural data from this task was analysed for the purposes of this paper (see supplementary material for the reasoning behind this). The data of interest here included accuracy rates, response times, and patterns of emotion labelling.

Stimuli for the emotion conditions of the task consisted of photographs of faces displaying fearful, angry, sad and neutral expressions (5 female and 5 male), taken from the well validated Pictures of Facial Affect (Ekman & Friesen, 1976) (see supplementary material for the reasoning behind the selection of these expressions). For the sake of ease, all four of the facial expressions used in this task are henceforth referred to as 'emotions'. These stimuli were computerised black and white images cropped in an oval shape, showing only faces and excluding features such as hair, ears and facial accessories. All fearful, angry and sad face images were degraded through morphing software and displayed at 60% of the original emotional intensity to increase the possibility of detecting subtle emotion recognition impairments by reducing the likelihood of ceiling effects.

Stimuli for the control condition of the task included a) 10 black and white validated photographs of houses taken from the DalHouses image set and rated as having a high degree of face-likeness (Filliter et al., 2016); and b) 10 black-and-white front-facing car images taken from Google due to their face-likeness, and void of extraneous detail such as branding or visible number plates. Houses and (front-facing) cars were selected as they are familiar stimuli with perceptual similarities to faces in the sense that they have distinct and consistent featural components (e.g. windows or lights for eyes, doors or grills for mouth) that form part of a gestalt whole (eyes/windows/lights above mouth/door/grill). They are known to activate face selective areas of the brain, albeit to a lesser extent than faces themselves (Filliter et al., 2016; Gauthier et al., 2000; McGugin et al., 2012).

House and car images were manipulated in the GNU Image Manipulation Program to ensure image saturation was equivalent to that of the face stimuli, and were not cropped to an oval shape, to avoid image distortion.

Emotive and control stimuli were presented one at a time on a black background; with an average stimulus onset asynchrony of 4 seconds (including an initial 500ms fixation cross, a 500ms stimulus presentation, and a jittered inter-stimulus interval of between 1.5-5.5 seconds). The stimulus presentation time of 500ms was selected to ensure rapid initial impressions of the facial expression were captured. Stimuli were displayed within blocks that made up five different conditions (e.g. stimulus categories); fear targets, angry targets, sad targets, neutral targets, and house targets (control stimuli).

At the beginning of each block, a target word (e.g., ‘Angry’) was displayed, and participants were required to identify stimuli matching the target word (targets) and stimuli not matching the target word (nontargets/foils). Six target stimuli and six nontarget/foil stimuli were presented within each block. Blocks were repeated eight times across the task, to achieve a total of *n* = 48 fear, angry, sad, neutral and house targets (6 targets repeated 8 times = 48 stimuli) and *n* = 48 fear, angry, sad, neutral, house non-targets (6 non-targets repeated 8 times = 48 stimuli) per condition. The total experiment time was 40 minutes across 4 runs of the task (40 blocks total, comprising 10 blocks per run in which each block condition [fear, angry, sad, neutral, and house targets] was presented twice). The runs were broken up with rest periods (during which structural neuroimaging scans were acquired) to reduce participant burden.

The condition order of each block across the whole task was counter-balanced so that each condition preceded and followed every other condition at least once, but not more than three times (i.e., the order of conditions presented within each run was different). Using a button-box, participants were instructed to press one button to label the stimulus as a target, and another button to label the stimulus as a nontarget. They were instructed respond as quickly and accurately as possible. Participants undertook one practice trial prior to completing the task.

Statistical analysis

All statistical analyses were conducted using SPSS (version 24). The mean accuracy for each participant was determined for each condition as a function of *sensitivity* (True Positive: correct recognition of the target as the target) plus *specificity* (True Negative: correct rejection of a foil as not the target) divided by the total number of stimuli across the task and converted into a percentage. Mean response times were calculated as an average of the sensitivity and specificity for each condition across the task in milliseconds. To capture the mislabelling of emotions, the false alarm data was used; defined as the proportion of emotion foils mislabelled as emotion targets, converted into a percentage. The assumptions of normality, linearity, univariate and multivariate outliers and multicollinearity were assessed with no violations. The assumption of homogeneity of variance-covariance matrices (as measured by the Box's M test) was violated, though this test is robust to violations when sample sizes between groups are similar (Tabachnick and Fidell, 2013), as was the case in this study.

Descriptive Analysis: Independent samples *t*-tests or χ^2 tests were used to assess between-group differences in age, sex and estimated premorbid IQ. Because of group differences in premorbid IQ, bivariate correlations were conducted to check their effect on all measures of interest (Supplementary Table 1). As there was no correlation whatsoever between premorbid IQ and any of these variables within either the BD or healthy control group, premorbid IQ was not included as a covariate in the analyses specified below.

Analysis Set 1: To address Research Question 1 and Hypothesis 1, we first conducted two separate mixed repeated measures analyses of variance (ANOVAs), one for accuracy and one for response time, to assess differences between BD and healthy control groups in the *pattern* of emotion recognition accuracies or response times across emotion conditions. For the accuracy analysis, significant interactions between the between-subject factor (group) and the within-subject factor (emotion condition) were followed up with appropriate post-hoc tests, with Bonferroni correction used to control the Type I error rate related to multiple comparisons (p < .05/4 emotion conditions = p = .013). As there were no significant interactions for response time, post-hoc tests were not conducted[1].

Analysis Set 2: To determine whether there were between-group differences in the degree to which each emotion was mislabelled (Research Question 2), we first conducted a one-way multivariate analysis of variance (MANOVA) on the false alarm data, that is the percentage of false alarms per condition (three possible foils for each of the four emotion targets = 12 dependent variables; group [BD and healthy controls] = 1 independent variable). As the emotion mislabelling analyses were exploratory in nature, we also explored within-group patterns of emotion mislabelling regardless of whether or not there were between-group differences in the degree of emotion mislabelling on these variables. To do this we conducted repeated measures ANOVAs for each of the four emotion foils to assess the pattern of emotion mislabelling within each group separately (fear, angry, sad, neutral, for

[1] As response times were averaged across the four runs of the task, we conducted a preliminary check to determine whether there were group differences across task runs using a repeated measures ANOVA. The interaction between group and run was not significant. The results of this analysis are presented in Supplementary Table 3.

each group = eight ANOVAs in total), specifying percentage false alarms for each of the targets as the within-subject factor. For example, if neutral was the foil, the false alarm data for fear, angry and sad targets were used to investigate whether neutral faces were being preferentially mislabelled as fearful, angry or sad. Pairwise comparisons between these false alarm emotion targets were then considered if the omnibus within-group ANOVA was significant. In these cases, Bonferroni correction was applied to account for multiple comparisons within each group ($p < .05/3$ comparisons for each emotion for each group = $p =$ .017). Further, as this study involved a series of target present or target absent decisions, signal detection was conducted as an additional analysis to assess the degree of discrimination and bias in task-based decision-making. False alarm and hit rates were converted into z-scores and signal detection was applied to the data using the following formulas (where H = hit rate; F = false alarm rate; d' = d-prime, a measure of discrimination; c = criterion location, a measure of bias; Huang and Ferreira, 2020):

$$d' = z(H) - z(F)$$

$$c = - 1/2 * (z(H) + z(F))$$

Two mixed repeated measures ANOVAs, one for d' and one for c, were then conducted to evaluate emotion condition*group interaction effects and to compare discrimination and bias between groups.

Analysis 3: To address Research Question 3, a one-way MANOVA was used to assess between-group differences in mean accuracies and response times for the control condition, which comprised the house and car stimuli.

Analysis 4: To address Research Question 4, MADRS and YMRS scores were correlated with the accuracy, response time and false alarm data in the BD group. This was done using Pearson's bivariate correlations with Bonferroni correction ($p < .05/20$

correlations = p = .003), to explore if current mood symptom severity had any effect on the variables of interest.

For the outcomes of all analyses, effect sizes are reported as either Cohen's *d* (0.2 small, 0.5 medium, 0.8 large) or η_p^2 (0.01 small, 0.09 medium, 0.25 large). Figures were produced using the R package ggplot2 (Wickham, 2009) and Microsoft Excel.

Results

Descriptive Analysis: Demographic and clinical characteristics of the sample

The demographic and clinical characteristics of the sample are presented in Table 1. There were no significant differences in age ($t(64) = 1.07$, $p = .289$) or sex ($\chi^2 = 2.92$, $p = .087$) between BD and healthy control groups. There was a significant group difference in premorbid IQ ($t(63) = -2.72$, $p = .008$), which was higher in the healthy control group (M = 115.19, SD = 8.83) compared to the BD group (M = 107.79, SD = 12.71). The BD group had significantly higher scores than the healthy control group on the MADRS ($t(64) = 5.06$, $p < .001$, M = 9.09, SD = 8.03 and M = 1.69, SD = 2.04 respectively), and on the YMRS ($t(64) = 5.42$, $p < .001$, M = 4.15, SD = 3.57 and M = .59, SD = 1.04 respectively).

[INSERT TABLE 1 ABOUT HERE.]

Analysis Set 1: Between-group differences in emotion labelling

Accuracy means and standard deviations for emotion labelling for each group are displayed in Figure 1 as well as Supplementary Table 2, and response time means and standard deviations are displayed in Supplementary Table 3[2].

[2] Post-hoc analyses were conducted to investigate emotion recognition performance by psychotropic medication status in the BD group. There were no significant differences in accuracy or response times between those taking and not taking different classes of medication. See the supplementary material for further information.

Accuracy: A significant main effect of emotion condition was evident (Wilks' lambda = .70, *F* (3, 62) = 8.99, $p < .001$, $\eta_p^2 = .30$), with emotion recognition accuracy across conditions occurring in the descending order of fear (*M* = 80.05, SD = 13.74), neutral (*M* = 76.66, SD = 14.18), angry (*M* = 76.34, SD = 14.58), and sad (*M* = 74.71, SD = 14.93). There was a significant group effect (*F* (4, 61) = 8.23, $p < .001$, Wilks' lambda = .65, $\eta_p^2 = .35$), with BD patients labelling facial expressions less accurately than healthy controls overall (Cohen's *d* = 0.80). A significant interaction between emotion condition and group was also apparent (Wilks' lambda = .82, *F* (3,62) = 4.65, *p* = .005) with both groups labelling fear the most accurately (BD: *M* = 75.15, SD = 15.43; healthy control: *M* = 85.25, SD = 9.38), but BD patients demonstrating accuracy across the other conditions in the descending order of angry (*M* = 72.18, SD = 15.85), sad (*M* = 71.63, SD = 14.99), and neutral (*M* = 69.61, SD = 15.93). In contrast, healthy controls demonstrated accuracy across the other conditions in the descending order of neutral (*M* = 84.15, SD = 6.28), angry (*M* = 80.76, SD = 11.8), and sad (*M* = 77.99, SD = 14.39). Follow-up analyses indicated that BD patients were less accurate than healthy controls at labelling neutral (*F* (1, 64) = 23.25, $p < .001$, $\eta_p^2 = .27$), fear (*F* (1, 64) = 10.17, *p* = .002, $\eta_p^2 = .14$) and angry emotions (*F* (1, 64) = 6.16, *p* = .016, $\eta_p^2 = .09$), although there was no significant between-group difference for the labelling of sad emotions (*F* (1, 64) = 3.09, *p* = .084, $\eta_p^2 = .46$). Only the neutral and fear group differences survived Bonferroni correction (*p* = .013), however, qualitatively worse performance was evident in the BD group for all emotion conditions (Cohen's *d* range = -.43 to -1.20). This is illustrated in Figure 1.

Response time: There was a significant main effect of emotion condition (Wilks' lambda = .66, *F* (3, 62) = 10.70, $p < .001$, $\eta_p^2 = .34$), with response times in both groups increasing across emotion conditions in the order of angry (*M* = 937.09, SD = 180.20), neutral (*M* = 952.41, SD = 189.28), fear (*M* = 984.87, SD = 214.60), and sad (*M* = 993.88, SD = 192.37). No significant omnibus group effects were evident (*F* (4, 61) = 1.14, *p* = .349, Wilks' lambda = 0.93, η_p^2 = .07), nor was there a significant emotion condition*group interaction (Wilks' lambda = .95, *F* (3,62) = 1.01, *p* > .05).

[INSERT FIGURE 1 ABOUT HERE.]

Analysis Set 2: Examination of emotion mislabelling

Between-group differences in false alarms: There was no significant omnibus (*F* (12, 53) = 1.72, *p* = .090; Wilks' lambda = .72), and no individual emotion condition (all *p's* >.05) differences between BD and healthy control groups in the *degree* of mislabelling emotion foils as emotion targets. The omnibus effect size was large, η_p^2 = .28, and individual emotion condition false alarm effects ranged from Cohen's *d* = -.42 to Cohen's *d* = .58 (Table 2).

Within-group patterns of emotion mislabelling: Emotion mislabelling means and standard deviations for the BD and healthy control groups are presented in Table 2 and Figure 2.

Neutral face mislabelling: In the BD group, there was a significant effect of emotion target on the mislabelling of neutral faces (Wilks' lambda = .42, *F* (2, 32) = 22.53, *p* < .001, η_p^2 = .59). Post-hoc pairwise comparisons indicated that after Bonferroni correction for multiple comparisons, neutral faces were mistaken for sad more frequently than fearful (false alarm mean difference of 9.74%, *p* < .001), and for angry more frequently than fearful (false alarm mean difference of 8.82%, *p* < .001). Neutral faces were not preferentially mislabelled as angry or sad (false alarm mean difference of .92%, *p* = 1.00).

In the healthy control group, there was a significant effect of emotion target on the mislabelling of neutral faces (Wilks' lambda = .48, *F* (2, 30) = 16.23, *p* < .001, η_p^2 = .52).

The same pattern was present as in BD (sad more frequently than fearful, false alarm mean difference of 5.08%, $p = .016$; and angry more than fearful (false alarm mean difference of 11.52%, $p < .001$), except that the contrast between angry and sad was significant (angry more than sad; false alarm mean difference of 6.45%, $p = .011$).

Fearful face mislabelling: There was no significant effect of emotion target on the mislabelling of fearful faces in the BD group (Wilks' lambda = .90, F (2, 32) = 1.76, $p = .189$, $\eta_p^2 = .10$); but there was in the healthy control group (Wilks' lambda = .50, F (2, 30) = 14.91, $p < .001$, $\eta_p^2 = .50$). Post-hoc analyses indicated that fearful faces were mislabelled as angry more than neutral (false alarm mean difference of 15.82%, $p < .001$), and sad more than neutral (false alarm mean difference of 17.77%, $p < .001$). Fearful faces were not preferentially mislabelled as angry or sad (false alarm mean difference of 1.95%, $p = 1.00$).

Angry face mislabelling: There was no significant effect of emotion target on the mislabelling of angry faces in either the BD group (Wilks' lambda = .99, F (2, 32) = .09, $p = .911$, $\eta_p^2 = .01$), or the healthy control group (Wilks' lambda = 0.86, F (2, 30) = 2.50, $p = .099$, $\eta_p^2 = .14$).

Sad face mislabelling: In the BD group, there was a significant effect of emotion target on the mislabelling of sad faces (Wilks' lambda = .57, F (2, 32) = 12.22, $p < .001$, $\eta_p^2 = .43$). Post-hoc pairwise comparisons indicated that sad faces were mistaken for neutral more than angry (false alarm mean difference of 18.93%, $p < .001$). Sad faces were not preferentially mislabelled as fearful or neutral (false alarm mean difference of 12.68%, $p = .030$), or fearful or angry (false alarm mean difference of 6.25%, $p = .128$).

In the healthy control group, there was a significant effect of emotion target on the mislabelling of sad faces (Wilks' lambda = .62, F (2, 30) = 9.19, $p = .001$, $\eta_p^2 = .38$). Post-hoc pairwise comparisons indicated that sad faces were mistaken for neutral more than angry (false alarm mean difference of 14.45%, $p = .001$), and fearful more than angry (false alarm mean difference of 12.31%, p = .005). Sad faces were not preferentially mislabelled as fearful or neutral (false alarm mean difference of 2.15%, p = 1.00).

[INSERT TABLE 2 ABOUT HERE.]

[INSERT FIGURE 2 ABOUT HERE.]

Signal detection: The results of the signal detection analysis, including mean d' and c values, are presented in Table 3.

Discrimination (d'): The BD group had a higher false alarm rate than hit rate, as indicated by their negative mean d' values. There was no significant main effect of emotion condition, Wilks' lambda = 1.00, F (2, 63) = .00, p = 1.00, η_p^2 = .00, nor was there a significant interaction between emotion condition and group, Wilks' lambda = .99, F (2, 63) = .18, p = .832, η_p^2 = .01). The main effect comparing the two groups was not significant, F (1, 64) = 1.2, p = .276, η_p^2 = .02.

Bias (c): There was no significant main effect of emotion condition, Wilks' lambda = 1.00, F (11, 54) = .00, p = 1.00, η_p^2 = .00, nor was there a significant interaction between emotion condition and group, Wilks' lambda = .87, F (11, 54) = .76, p = .678, η_p^2 = .13). There was a significant main effect of group, F (1, 64) = 17.83, p < .001, η_p^2 = .22. The BD group had positive mean c values for all emotion target / foil combinations, and vice versa for the healthy control group, indicating that the BD group were generally *conservative* in their decision-making bias and the healthy control group were *liberal* in their decision-making bias.

[INSERT TABLE 3 ABOUT HERE.]

Analysis 3: Control condition

There were no significant between-group differences in accuracy (BD: *M* = 90.50, SD = 21.91; healthy controls: *M* = 95.28, SD = 13.94) or response time (BD: *M* = 615.94, SD = 138.50; healthy controls: *M* = 595.75, SD = 115.81) for the labelling of houses (*F* (2, 61) = .56, *p* = .573, Wilks' lambda = .98, Cohen's *d* = - .26 and .16 respectively).

Analysis 4: Correlations between facial emotion recognition performance and mood symptoms

In the BD group, bivariate correlation analyses indicated no significant associations between mood symptom severity (MADRS and YMRS scores) and either accuracy, response time or false alarm variables after Bonferroni correction (adjusted α = .003; Supplementary Table 7).

Discussion

The aim of this study was to build on previous work on facial emotion recognition in BD patients compared to healthy controls by examining accuracy and response times across emotions and exploring patterns of emotion misattribution. Our findings largely support previous studies (de Brito Ferreira Fernandes et al., 2016; Derntl et al., 2009; Seidel et al., 2012; Vederman et al., 2012), where specifically, and in line with Hypothesis 1, we found the BD group to be less accurate in recognising facial emotions overall compared to healthy controls. This was particularly the case for neutral and fearful facial recognition, for which large patient-healthy control effect sizes were evident.

Medium between-group effect sizes were also evident for the recognition of angry and sad faces, with patients performing worse than healthy controls. However, these effects did not survive correction for multiple comparisons. Notably, the effect sizes evident in this study are comparatively larger than previous research. This may be explained by the nature of the task at hand, as we used degraded facial emotion stimuli displayed at 60% of the original intensity. This appears to have been effective in increasing task difficulty and increasing sensitivity to the previously documented *subtle* between-group differences in task performance.

Contrary to expectations, response times did not differ between groups, although effect sizes were in the small-medium range favouring better performance in the healthy controls. Nonetheless, the absence of response time differences and relatively smaller between-group effects suggests that reduced facial emotion recognition accuracy in the BD group was not due to a speed versus accuracy trade off. There were also no between-group differences on the control condition, indicating that general visual processing ability was not compromised in the BD cohort. We also did not find any significant correlations between mood symptom severity scores and the emotion recognition measures of interest, consistent with past work indicating that facial emotion recognition deficits are a trait-like feature of BD.

The reduced accuracy for recognising neutral faces in the BD group is an important finding because neutral faces are commonly used as a baseline or contrast condition in facial emotion recognition paradigms, particularly those involving functional neuroimaging (Chen et al., 2006; Dima et al., 2013; Grotegerd et al., 2014; Jogia et al., 2008; Marchand et al., 2011). Our findings suggest that BD patients misinterpret neutral faces as being emotive to a much greater degree than healthy individuals, in line with previous preliminary evidence (Gur et al., 1992; Ruocco et al., 2014). It is possible that this is due to an inherent bias of BD patients toward the perception of emotion in the environment even in its absence, which fits with evidence indicating heightened emotional reactivity in BD patients compared to healthy controls (Miskowiak et al., 2019). An alternative explanation is that individuals with BD are hypersensitive to the trustworthiness of faces and thus more prone to emotion misattribution errors, since it has been shown that trustworthy neutral faces are more likely to be evaluated as happy, and untrustworthy neutral faces as angry (Oosterhof and Todorov, 2008, 2009). Irrespective of the driving factors behind the effect however, the findings of neutral face recognition deficits in BD in this study raise questions about the validity of findings from existing studies of BD in which neutral faces are used as a control condition; and suggest that the use of neutral face stimuli as a baseline condition in facial affect perception in BD should be reconsidered in future research.

Overall, there were no significant *between-group* differences in the magnitude by which neutral, sad, angry or fearful faces were confused for each other, although the between-group effect sizes for the mislabelling of fear as neutral (biased toward patients) or sad (biased toward healthy controls) were in the medium range. Within-group analysis indicated that emotion mislabelling patterns within the BD group were similar to the emotion mislabelling patterns of the healthy control group, with some minor differences (see supplementary material for a description). In combination with the accuracy findings, the generally comparable within-group patterns suggest that there is a spectrum on which facial emotion mislabelling occurs; such that individuals with BD appear to mislabel emotions in a similar, albeit exaggerated manner, compared to healthy individuals. In other words, individuals with BD appear to make the same type of emotion mislabelling mistakes as healthy individuals, but they make these errors more often. This is important as it suggests that training programs designed to improve emotional intelligence in the general population by improving facial emotion recognition may be successfully adapted for use in BD.

As suggested by the signal detection findings, individuals with BD may mistake more facial expressions than they correctly identify; possibly due to a conservative bias in their decision making. To our knowledge, we believe that no previous studies have conducted signal detection analysis on facial emotion processing tasks in BD. Our findings are in contrast to another study that conducted a signal detection analysis of a sustained attention task, and found no difference in response bias between euthymic BD and healthy controls (Harmer et al., 2002). Conservative decision making in BD may therefore be specific to *social* cognitive function, as suggested by our findings, rather than a global cognitive trait. Further research is required to clarify these discrepancies in social versus non-social cognitive response bias in BD.

Our findings of exaggerated emotion misattribution may explain some of the interpersonal and social impairments associated with BD, such as emotional reactivity (Henry et al., 2012). Although there is some indirect existing literature on the impact of emotion recognition impairments on psychosocial outcomes, it is sparse and inconclusive (Ospina et al., 2018; Van Rheenen and Rossell, 2014b). Given the larger effect sizes evident in the current study compared to past work, it appears that more sensitive tasks are needed to elicit the apparently subtle facial emotion processing impairments of BD. Use of such tasks may thus help in clarifying the presence and nature of associations between impaired emotion recognition and psychosocial outcomes in BD, which will aid in understanding the extent to which the deficits apparent in this study are clinically meaningful.

The current study had some limitations that should be considered. First, the sample size was relatively modest, owing to its place within an expensive neuroimaging project in which funding was only available for a certain number of participants. Second, we used a task that took 40 minutes for participants to complete. This was performed in the context of a functional MRI scanning session in which stimulus blocks were broken up into four stimulus runs interspersed with other scan acquisitions during which participants rested. Nonetheless, this is a relatively long testing paradigm that requires a degree of sustained attention. Since group by run interactions were not significant (see Supplementary Table 3), and to account for potential fatigue, we used an average of the scores across the four stimulus runs of the task. Given BD participants showed similar response times and similar patterns of emotion mislabelling as the healthy controls, it appears that this group was attending to the task and not disproportionately affected by fatigue. Further, the stimulus presentation in this task was a relatively brief 500ms in duration. Though this is different to some previous research that used longer stimulus presentation times, we reason that our task may be more ecologically valid, in that real-world facial emotions are not fixed for long durations. We acknowledge that this task design could increase burden for participants with cognitive processing deficits, however, the response window was jittered up to 5.5 seconds to capture slower psychomotor reactions. The absence of response time differences between groups suggests that task duration did not disproportionately affect performance in the BD group.

Third, the BD sample comprised some patients with mild-moderate clinical symptoms, which hampers inferences regarding whether the findings observed here are trait-like in nature. Nonetheless, the absence of correlations between mood rating scores and task performance suggests that the accuracy of facial emotion recognition is independent of mood symptoms. This supports previous research that has suggested that facial emotion deficits may be a trait feature of BD (Samame et al., 2015; Van Rheenen and Rossell, 2014a).

Further, while we examined the effect of medication type on facial emotion recognition task performance, we did not specifically assess the effect of medication *dose* on this variable. Future research would do well to replicate this study in a medication-naïve sample. In addition, though premorbid IQ was estimated using the WTAR, we did not assess the *current* IQ of participants in this study. This is a limitation of the current study, though it should be noted that there was no association between premorbid IQ and facial emotion recognition task performance in this cohort. It is also important to consider the lack of ethnic diversity of actors in the Pictures of Facial Affect; this is a limitation common to many studies using these stimuli in the area of facial emotion recognition.

Finally, caution is warranted in generalising these findings beyond their application to the emotions examined in this study. For example, we intentionally omitted positive facial expressions in the design of our task to limit confounds related to differences in the complexity of positive and negative faces and the strategies used to process them (see Van Rheenen and Rossell, 2013 for a discussion). Future research should consider including positively valenced stimuli, such as happy faces, however, to elucidate potential biases in positively *or* negatively valenced mislabelling of neutral faces. Future research would also do well to explicitly replicate and further probe the neutral face processing impairments seen here in BD patients.

In sum, the findings from the current study suggest that facial emotion recognition is impaired in BD. In particular, BD appears to be associated with deficits in the accurate recognition of neutral faces, as well as fearful, angry, and sad faces. Similar, but exaggerated patterns of emotion mislabelling are evident in BD patients compared to healthy controls. Future emotion processing research in BD should take caution in using neutral faces in control conditions of experimental tasks, and should aim to explicitly characterise the neutral facial recognition deficits seen here in BD using neuroimaging techniques. Such research, in combination with other work replicating the emotion misattribution patterns seen in this study, will facilitate a more complete understanding of facial emotion recognition deficits in BD. This will go some way in optimising psychosocial treatments for the disorder, for example, by facilitating the formulation and refinement of facial emotion recognition training programs for BD.

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, 5th ed. APA, Washington, DC.

Bozikas, V.P., Tonia, T., Fokas, K., Karavatos, A., Kosmidis, M.H., 2006. Impaired emotion processing in remitted patients with bipolar disorder. J. Affect. Disord. 91, 53-56. DOI: 10.1016/j.jad.2005.11.013.

Brotman, M.A., Skup, M., Rich, B.A., Blair, K.S., Pine, D.S., Blair, J.R., Leibenluft, E., 2008. Risk for bipolar disorder is associated with face-processing deficits across emotions. J. Am. Acad. Child Adolesc. Psychiatry 47, 1455-1461. DOI: 10.1097/CHI.0b013e318188832e.

Calafiore, D., Rossell, S.L., Van Rheenen, T.E., 2018. Cognitive abilities in first-degree relatives of individuals with bipolar disorder. J. Affect. Disord. 225, 147-152. DOI: 10.1016/j.jad.2017.08.029.

Chen, C.H., Lennox, B., Jacob, R., Calder, A., Lupson, V., Bisbrown-Chippendale, R., Suckling, J., Bullmore, E., 2006. Explicit and implicit facial affect recognition in manic and depressed States of bipolar disorder: a functional magnetic resonance imaging study. Biol. Psychiatry 59, 31-39. DOI: 10.1016/j.biopsych.2005.06.008.

Daros, A.R., Ruocco, A.C., Reilly, J.L., Harris, M.S., Sweeney, J.A., 2014. Facial emotion recognition in first-episode schizophrenia and bipolar disorder with psychosis. Schizophr. Res. 153, 32-37. DOI: 10.1016/j.schres.2014.01.009.

de Brito Ferreira Fernandes, F., Gigante, A.D., Berutti, M., Amaral, J.A., de Almeida, K.M., de Almeida Rocca, C.C., Lafer, B., Nery, F.G., 2016. Facial emotion recognition in euthymic patients with bipolar disorder and their unaffected first-degree relatives. Compr. Psychiatry 68, 18-23. DOI: 10.1016/j.comppsych.2016.03.001.

Derntl, B., Seidel, E.M., Kryspin-Exner, I., Hasmann, A., Dobmeier, M., 2009. Facial emotion recognition in patients with bipolar I and bipolar II disorder. Br. J. Clin. Psychol. 48, 363-375. DOI: 10.1348/014466509X404845.

Dima, D., Jogia, J., Collier, D., Vassos, E., Burdick, K.E., Frangou, S., 2013. Independent modulation of engagement and connectivity of the facial network during affect processing by CACNA1C and ANK3 risk genes for bipolar disorder. JAMA Psychiatry 70, 1303-1311. DOI: 10.1001/jamapsychiatry.2013.2099.

Ekman, P., Friesen, W.V., 1976. Pictures of facial affect. Consulting Psychologists Press, Palo Alto, CA.

Filliter, J.H., Glover, J.M., McMullen, P.A., Salmon, J.P., Johnson, S.A., 2016. The DalHouses: 100 new photographs of houses with ratings of typicality, familiarity, and degree of similarity to faces. Behav. Res. Methods 48, 178-183. DOI: 10.3758/s13428-015-0561-8.

Gauthier, I., Skudlarski, P., Gore, J.C., Anderson, A.W., 2000. Expertise for cars and birds recruits brain areas involved in face recognition. Nat. Neurosci. 3, 191-197. DOI: 10.1038/72140.

Getz, G.E., Shear, P.K., Strakowski, S.M., 2003. Facial affect recognition deficits in bipolar disorder. J. Int. Neuropsychol. Soc. 9, 623-632. DOI: 10.1017/S1355617703940021.

Goghari, V.M., Sponheim, S.R., 2013. More pronounced deficits in facial emotion recognition for schizophrenia than bipolar disorder. Compr. Psychiatry 54, 388-397. DOI: 10.1016/j.comppsych.2012.10.012.

Grotegerd, D., Stuhrmann, A., Kugel, H., Schmidt, S., Redlich, R., Zwanzger, P., Rauch, A.V., Heindel, W., Zwitserlood, P., Arolt, V., Suslow, T., Dannlowski, U., 2014. Amygdala excitability to subliminally presented emotional faces distinguishes unipolar and bipolar depression: an fMRI and pattern classification study. Hum. Brain Mapp. 35, 2995-3007. DOI: 10.1002/hbm.22380.

Gur, R.C., Erwin, R.J., Gur, R.E., Zwil, A.S., Heimberg, C., Kraemer, H.C., 1992. Facial emotion discrimination: II. Behavioral findings in depression. Psychiatry Res. 42, 241-251. DOI: 10.1016/0165-1781(92)90116-k.

Harmer, C.J., Clark, L., Grayson, L., Goodwin, G.M., 2002. Sustained attention deficit in bipolar disorder is not a working memory impairment in disguise. Neuropsychologia. 40, 1586-1590. DOI: 10.1016/S0028-3932(02)00019-2.

Henry, C., Phillips, M., Leibenluft, E., M'Bailara, K., Houenou, J., Leboyer, M., 2012. Emotional dysfunction as a marker of bipolar disorders. Front Biosci (Elite Ed) 4, 2622-2630. DOI: 10.2741/e578.

Huang, Y., Ferreira, F., 2020. The Application of Signal Detection Theory to Acceptability Judgments. Front. Psychol. 11, 73. DOI: 10.3389/fpsyg.2020.00073.

Jogia, J., Haldane, M., Cobb, A., Kumari, V., Frangou, S., 2008. Pilot investigation of the changes in cortical activation during facial affect recognition with lamotrigine monotherapy in bipolar disorder. Br. J. Psychiatry 192, 197-201. DOI: 10.1192/bjp.bp.107.037960.

Keck, P.E., Jr., Calabrese, J.R., McQuade, R.D., Carson, W.H., Carlson, B.X., Rollin, L.M., Marcus, R.N., Sanchez, R., Aripiprazole Study, G., 2006. A randomized, double-blind, placebo-controlled 26-week trial of aripiprazole in recently manic patients with bipolar I disorder. J. Clin. Psychiatry 67, 626-637. DOI: 10.4088/jcp.v67n0414.

Kessels, R.P., Montagne, B., Hendriks, A.W., Perrett, D.I., de Haan, E.H., 2014. Assessment of perception of morphed facial expressions using the Emotion Recognition Task: normative data from healthy participants aged 8-75. J. Neuropsychol. 8, 75-93. DOI: 10.1111/jnp.12009.

Lee, P., Van Meter, A., 2020. Emotional body language: Social cognition deficits in bipolar disorder. J. Affect. Disord. 272, 231-238. DOI: 10.1016/j.jad.2020.03.114.

Lembke, A., Ketter, T.A., 2002. Impaired recognition of facial emotion in mania. Am. J. Psychiatry 159, 302-304. DOI: 10.1176/appi.ajp.159.2.302.

Marchand, W.R., Lee, J.N., Garn, C., Thatcher, J., Gale, P., Kreitschitz, S., Johnson, S., Wood, N., 2011. Aberrant emotional processing in posterior cortical midline structures in bipolar II depression. Prog. Neuropsychopharmacol. Biol. Psychiatry 35, 1729-1737. DOI: 10.1016/j.pnpbp.2011.05.017.

McGugin, R.W., Gatenby, J.C., Gore, J.C., Gauthier, I., 2012. High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. Proc. Natl. Acad. Sci. U. S. A. 109, 17063-17068. DOI: 10.1073/pnas.1116333109.

Miskowiak, K.W., Seeberg, I., Kjaerstad, H.L., Burdick, K.E., Martinez-Aran, A., Del Mar Bonnin, C., Bowie, C.R., Carvalho, A.F., Gallagher, P., Hasler, G., Lafer, B., Lopez-Jaramillo, C., Sumiyoshi, T., McIntyre, R.S., Schaffer, A., Porter, R.J., Purdon, S., Torres, I.J., Yatham, L.N., Young, A.H., Kessing, L.V., Van Rheenen, T.E., Vieta, E., 2019. Affective cognition in bipolar disorder: A systematic review by the ISBD targeting cognition task force. Bipolar Disord 21, 686-719. DOI: 10.1111/bdi.12834.

Montgomery, S.A., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382-389. DOI: 10.1192/bjp.134.4.382.

Oosterhof, N.N., Todorov, A., 2008. The functional basis of face evaluation. Proc. Natl. Acad. Sci. U. S. A. 105, 11087-11092. DOI: 10.1073/pnas.0805664105.

Oosterhof, N.N., Todorov, A., 2009. Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. Emotion 9, 128-133. DOI: 10.1037/a0014520.

Ospina, L.H., Nitzburg, G.C., Shanahan, M., Perez-Rodriguez, M.M., Larsen, E., Latifoglu, A., Burdick, K.E., 2018. Social cognition moderates the relationship between neurocognition and community functioning in bipolar disorder. J. Affect. Disord. 235, 7-14. DOI: 10.1016/j.jad.2018.03.013.

Rich, B.A., Vinton, D.T., Roberson-Nay, R., Hommer, R.E., Berghorst, L.H., McClure, E.B., Fromm, S.J., Pine, D.S., Leibenluft, E., 2006. Limbic hyperactivation during processing of neutral facial expressions in children with bipolar disorder. Proc. Natl. Acad. Sci. U. S. A. 103, 8900-8905. DOI: 10.1073/pnas.0603246103.

Ruocco, A.C., Reilly, J.L., Rubin, L.H., Daros, A.R., Gershon, E.S., Tamminga, C.A., Pearlson, G.D., Hill, S.K., Keshavan, M.S., Gur, R.C., Sweeney, J.A., 2014. Emotion recognition deficits in schizophrenia-spectrum disorders and psychotic bipolar disorder: Findings from the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) study. Schizophr. Res. 158, 105-112. DOI: 10.1016/j.schres.2014.07.001.

Samame, C., Martino, D.J., Strejilevich, S.A., 2015. An individual task meta-analysis of social cognition in euthymic bipolar disorders. J. Affect. Disord. 173, 146-153. DOI: 10.1016/j.jad.2014.10.055.

Schaefer, K.L., Baumann, J., Rich, B.A., Luckenbaugh, D.A., Zarate, C.A., Jr., 2010. Perception of facial emotion in adults with bipolar or unipolar depression and controls. J. Psychiatr. Res. 44, 1229-1235. DOI: 10.1016/j.jpsychires.2010.04.024.

Seidel, E.M., Habel, U., Finkelmeyer, A., Hasmann, A., Dobmeier, M., Derntl, B., 2012. Risk or resilience? Empathic abilities in patients with bipolar disorders and their first-degree relatives. J. Psychiatr. Res. 46, 382-388. DOI: 10.1016/j.jpsychires.2011.11.006.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59 Suppl 20, 22-33;quiz 34-57.

Soeiro-de-Souza, M.G., Bio, D.S., David, D.P., Rodrigues dos Santos, D., Jr., Kerr, D.S., Gattaz, W.F., Machado-Vieira, R., Moreno, R.A., 2012. COMT Met (158) modulates facial emotion recognition in bipolar I disorder mood episodes. J. Affect. Disord. 136, 370-376. DOI: 10.1016/j.jad.2011.11.021.

Tabachnick, B.G., Fidell, L.S., 2013. Using multivariate statistics. Pearson, United Kingdom.

Van Rheenen, T.E., Ganella, E., Bauer, I., Bartholemeusz, C., 2019. Characterisation of social cognitive deficits on the schizophrenia -bipolar spectrum: an overview of current evidence, in: Lewandowski, K., Moustafa, A.M. (Eds.), Social cognition in psychosis (1-36). Elsevier Inc, New York.

Van Rheenen, T.E., Joshua, N., Castle, D.J., Rossell, S.L., 2017. Configural and Featural Face Processing Influences on Emotion Recognition in Schizophrenia and Bipolar Disorder. J. Int. Neuropsychol. Soc. 23, 287-291. DOI: 10.1017/S1355617716001211.

Van Rheenen, T.E., Rossell, S.L., 2014a. Let's face it: facial emotion processing is impaired in bipolar disorder. J. Int. Neuropsychol. Soc. 20, 200-208. DOI: 10.1017/S1355617713001367.

Van Rheenen, T.E., Rossell, S.L., 2014b. Objective and subjective psychosocial functioning in bipolar disorder: an investigation of the relative importance of neurocognition, social cognition and emotion regulation. J. Affect. Disord. 162, 134-141. DOI: 10.1016/j.jad.2014.03.043.

Van Rheenen, T.E., Rossell, S.L., 2013. Is the non-verbal behavioural emotion-processing profile of bipolar disorder impaired? A critical review. Acta Psychiatr. Scand. 128, 163-178. DOI: 10.1111/acps.12125.

Vederman, A.C., Weisenbach, S.L., Rapport, L.J., Leon, H.M., Haase, B.D., Franti, L.M., Schallmo, M.P., Saunders, E.F., Kamali, M.M., Zubieta, J.K., Langenecker, S.A., McInnis, M.G., 2012. Modality-specific alterations in the perception of emotional stimuli in Bipolar Disorder compared to Healthy Controls and Major Depressive Disorder. Cortex 48, 1027-1034. DOI: 10.1016/j.cortex.2011.03.017.

Wechsler, D., 2001. Wechsler Test of Adult Reading: WTAR. Psychological Corporation.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. Retrieved from: https://ggplot2.tidyverse.org.

Yalcin-Siedentopf, N., Hoertnagl, C.M., Biedermann, F., Baumgartner, S., Deisenhammer, E.A., Hausmann, A., Kaufmann, A., Kemmler, G., Muhlbacher, M., Rauch, A.S., Fleischhacker, W.W., Hofer, A., 2014. Facial affect recognition in symptomatically remitted patients with schizophrenia and bipolar disorder. Schizophr. Res. 152, 440-445. DOI: 10.1016/j.schres.2013.11.024.

Young, R.C., Biggs, J.T., Ziegler, V.E., Meyer, D.A., 1978. A rating scale for mania: reliability, validity and sensitivity. Br. J. Psychiatry 133, 429-435. DOI: 10.1192/bjp.133.5.429.

Table 1. Demographic and clinical characteristics of the sample

	BIPOLAR DISORDER N=34		HEALTHY CONTROLS N=32		COMPARISONS		
	%	*M (SD)*	*%*	*M (SD)*	*Test Statistic (t/χ^2)*	*p*	*Post-hoc*
Age (years)		38.00 (12.08)		34.94 (11.12)		.289	
Sex (%M)	64.71		43.75		2.92	.087	
Estimated premorbid IQ		107.79 (12.71)		115.19 (8.83)	-2.72	.008*	HC > BD
MADRS		9.09 (8.03)		1.69 (2.04)	5.06	.001*	BD > HC
YMRS		4.15 (3.57)		.59 (1.04)	5.42	.001*	BD > HC
% Affectively stable[†]	70.6		n/a				
Age of Illness Onset (years)	21.4 (8.8)		n/a				
% on Antipsychotics	50.0% (17)		n/a				
% on Antidepressants	35.3% (12)		n/a				
% on Mood stabilisers	41.2% (14)		n/a				

Note. BD = bipolar disorder; HC = healthy controls; MADRS = Montgomery-Asberg Depression Rating Scale; YMRS = Young Mania Rating Scale.

[†] Affectively stable patients had MADRS ≤ 12 and YMRS ≤ 10.

*Significant at $p < .01$.

Table 2. Emotion mislabelling for all emotions in bipolar disorder and healthy control groups

		False Alarm (%)				
		BIPOLAR DISORDER		**HEALTHY CONTROLS**		
Emotion foil	**Emotion target**	***M***	**SD**	***M***	**SD**	**Cohen's *d***
Neutral	Fear	4.04	7.03	5.27	9.13	-.15
	Angry	12.87	9.85	16.80	15.91	-.30
	Sad	13.79	15.67	10.35	13.15	.24
Fear	Neutral	11.40	11.96	5.66	7.34	.58
	Angry	15.99	15.01	21.48	25.25	-.26
	Sad	15.63	14.21	23.44	22.39	-.42
Angry	Neutral	14.52	15.13	11.52	14.12	.21
	Fear	14.71	20.23	15.82	20.58	-.05
	Sad	13.60	14.55	17.38	25.34	-.18
Sad	Neutral	29.60	22.73	24.80	22.02	.21
	Fear	16.91	17.98	22.66	23.10	-.28
	Angry	10.66	9.80	10.35	7.71	.04

Note. BD = bipolar disorder; HC = healthy controls.

False alarm relates to the proportion of emotional foils mislabeled as the emotional target; No between-groups false alarm comparisons were significant. Refer to Figure 2 for graphical representation of this table, and Supplementary Table 5 for within-group pairwise-comparison outcomes.

Table 3. Signal detection analysis for emotion mislabelling in bipolar disorder and healthy control groups, with mean values for discrimination and criterion location

		Signal Detection			
		BIPOLAR DISORDER		**HEALTHY CONTROLS**	
Emotion foil	**Emotion target**	**Mean (SE) d’**	**Mean (SE) c**	**Mean (SE) d’**	**Mean (SE) c**
Neutral	Fear	-.19 (.26)	.21 (.07)	.20 (.37)	-.22 (.05)
	Angry	-.10 (.22)	.18 (.11)	.11 (.35)	-.19 (.07)
	Sad	-.27 (.24)	.14 (.11)	.29 (.35)	-.15 (.06)
Fear	Neutral	-.27 (.21)	.15 (.10)	.28 (.36)	-.16 (.07)
	Angry	-.10 (.22)	.18 (.11)	.11 (.35)	-.19 (.07)
	Sad	-.07 (.22)	.24 (.11)	.08 (.34)	-.25 (.08)
Angry	Neutral	-.29 (.23)	.14 (.12)	.31 (.33)	-.15 (.06)
	Fear	-.26 (.23)	.18 (.11)	.27 (.34)	-.19 (.06)
	Sad	-.18 (.25)	.18 (.08)	.19 (.36)	-.19 (.07)
Sad	Neutral	-.39 (.18)	.09 (.15)	.41 (.27)	-.10 (.09)
	Fear	-.17 (.25)	.22 (.09)	.18 (.33)	-.23 (.08)
	Angry	-.13 (.24)	.12 (.10)	.14 (.35)	-.17 (.06)

Note. BD = bipolar disorder; HC = healthy controls; d’ = discriminability/sensitivity; c = criterion location (a measure of bias).

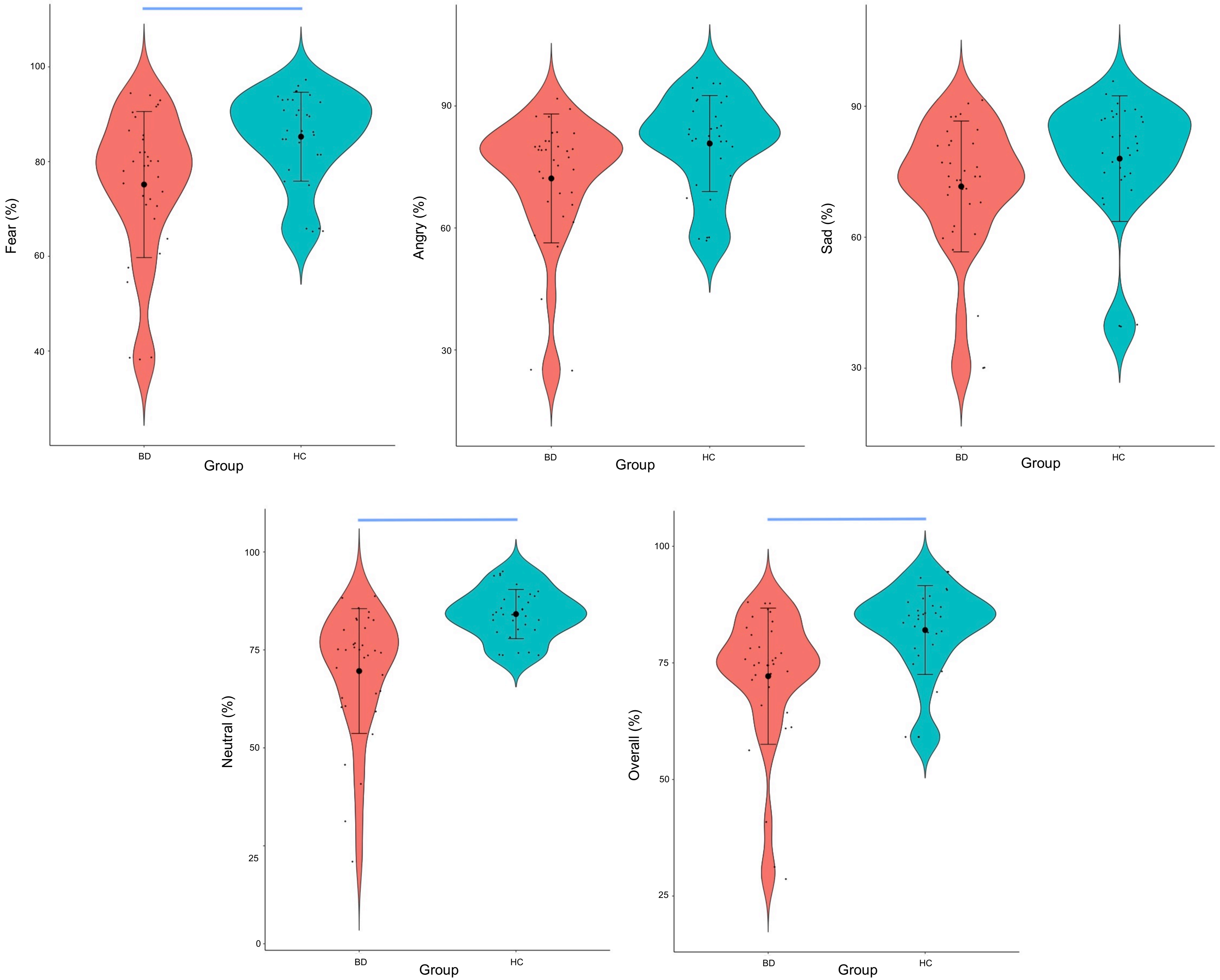
Fear (%)
100
80
60
40
BD
HC
Group
Angry (%)
90
60
30
BD
HC
Group
Sad (%)
90
60
30
BD
HC
Group
Neutral (%)
100
75
50
25
0
BD
HC
Group
Overall (%)
100
75
50
25
BD
HC
Group

Figure 1. *Violin plots showing means, standard deviations and probability distributions for task accuracy for fear, angry, sad, and neutral conditions, and overall performance, in bipolar disorder (BD) and healthy control (HC) groups.* Comparisons represented by a blue line are significant at p < 0.013 (Bonferroni corrected).

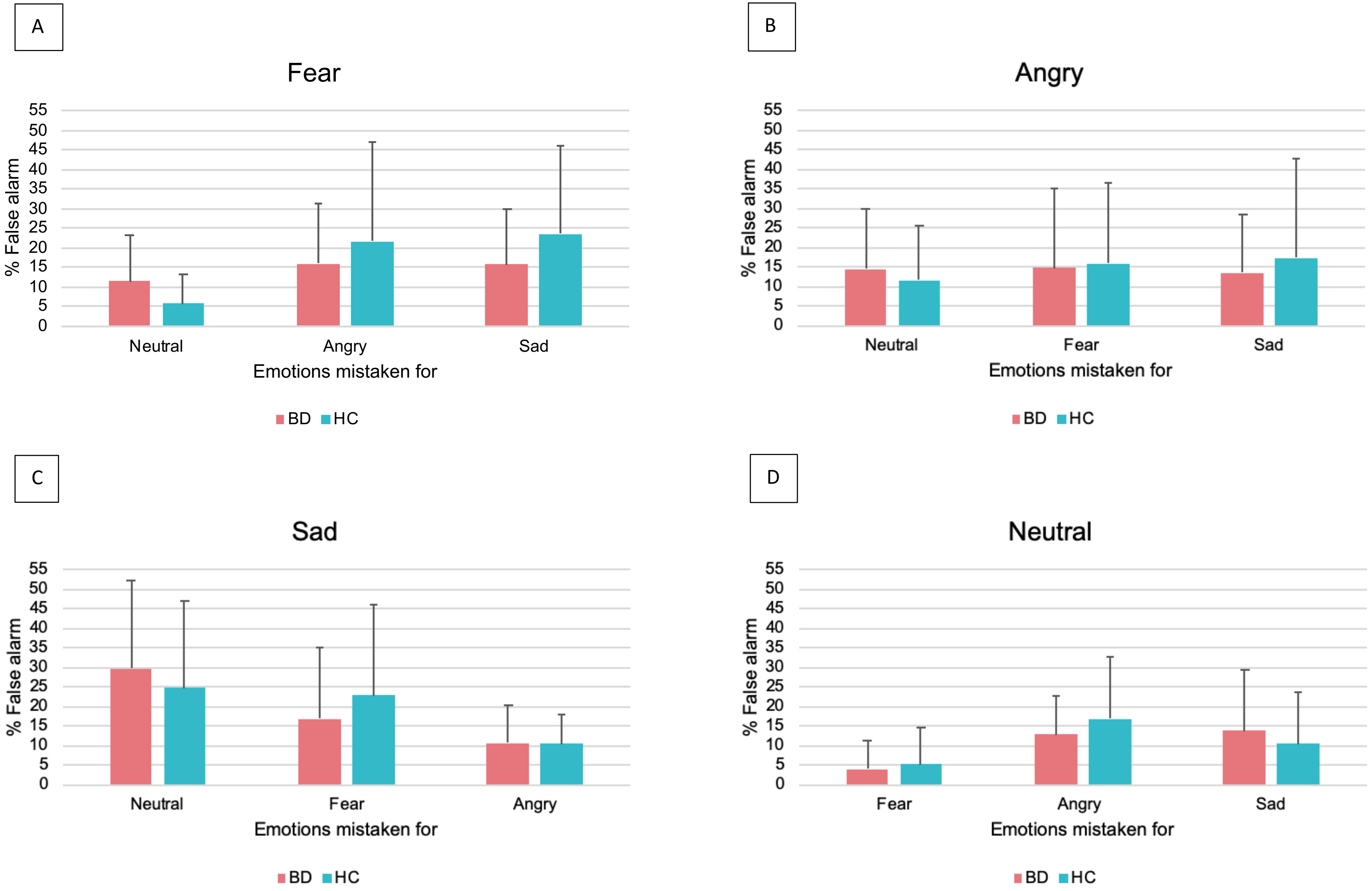
A
Fear
% False alarm
55
50
45
40
35
30
25
20
15
10
5
0
Neutral
Angry
Sad
Emotions mistaken for
BD
HC
B
Angry
% False alarm
Neutral
Fear
Sad
Emotions mistaken for
BD
HC
C
Sad
% False alarm
Neutral
Fear
Angry
Emotions mistaken for
BD
HC
D
Neutral
% False alarm
Fear
Angry
Sad
Emotions mistaken for
BD
HC

Figure 2. *Bar graphs showing the emotion mislabelling in bipolar disorder (BD) and healthy control (HC) groups.* A) emotion foil is fear, b) emotion foil is anger, c) emotion foil is sadness, d) emotion foil is neutral. Note that there were no significant differences in the false-alarm rates between-groups. Within-group pairwise comparison outcomes are displayed in Supplementary Table 6.

Supplementary material for "Characterisation of facial emotion recognition in bipolar disorder: focus on emotion mislabelling and neutral expressions"

Reasoning behind reporting only behavioural data in this study

The facial emotion recognition task was completed in the context of a functional magnetic resonance imaging scan. We elected not to report the functional neuroimaging data here, in order to focus the scope of this paper and permit a more thorough and detailed analysis of behavioural task performance than would have been possible if neuroimaging data was also reported. In future studies we plan to report the neuroimaging task findings.

Reasoning behind emotional stimuli selection for task

We selected fearful, angry and sad emotional face stimuli for use in our experimental task over other basic emotions such as happiness, surprise and disgust, because the task was designed for use within a neuroimaging study. Thus, we were unable to be comprehensive in our inclusion of all basic emotions as we had to balance the capacity to achieve an adequate number of trials for each stimulus condition for neuroimaging analysis with the length of the task. Further, we chose to limit our stimuli to those of a negative valence, because there is evidence that the complexity of positive and negative faces differ, as do the use of the face processing strategies used to process them (see Van Rheenen and Rossell, 2013 for a discussion of this issue). For example, happy faces may be processed more holistically, whereas the processing of negative faces may rely more heavily on the understanding the relationship between local facial information. Given that second order configural face processing (i.e., processing spatial configurations of faces) impairments have been demonstrated in BD (Joshua et al., 2016; Van Rheenen et al., 2017), we reasoned that negative emotions were most relevant to the aims of the study at hand. Further, we decided against including one positive emotion amongst multiple negative emotions, as is commonly done in the face processing tasks used in BD research, as this could produce a potentially artificial contrast that would confound facial emotion recognition performance on account of the task context (Van Rheenen and Rossell, 2013).

Effect of medication status on emotion recognition accuracies and response times

Two multivariate analyses of variance (MANOVAs) were conducted (one for accuracy and one for response time) to examine if emotion recognition performance differed by medication status in the BD group. There were four dependent variables for each MANOVA (accuracies or response times for fearful, angry, sad, and neutral emotion conditions) and three independent variables (mood stabilisers, antipsychotics, and antidepressants). Medication status in the BD group was coded dichotomously as either ‘taking’ or ‘not taking’ a given medication at the time of testing.

Accuracy: There were no significant differences between those taking and those not taking medication on the combined emotion recognition accuracy variables. This was the case for all medications examined; mood stabilisers (Wilks’ lambda = .76, *F* (4, 23) = 1.86, *p* =

.152, η_p^2 = .24), antipsychotics (Wilks' lambda = .85, *F* (4, 23) = .99, *p* = .433, η_p^2 = .15), and antidepressants (Wilks' lambda = .87, *F* (4, 23) = .87, *p* = .495, η_p^2 = .13). The means and standard deviations for this analysis is presented in Supplementary Table 4.

Response time: There were also no significant differences between those taking and those not taking medication on emotion recognition response times for mood stabilisers (Wilks' lambda = .92, *F* (4, 23) = .52, *p* = .719, η_p^2 = .08), antipsychotics (Wilks' lambda = .85, *F* (4, 23) = 1.02, *p* = .418, η_p^2 = .15), and antidepressants (Wilks' lambda = .93, *F* (4, 23) = .42, *p* = .792, η_p^2 = .07). The means and standard deviations for this analysis is presented in Supplementary Table 5.

Description of minor differences in emotion misattribution between BD and control groups

In the BD group, neutral faces were preferentially mistaken for sad *and* angry faces compared to fearful faces, whereas in the healthy control group neutral faces were preferentially mistaken for angry faces compared to the other emotions. Sad faces were also preferentially mistaken for neutral faces compared to fear and anger in the BD group, whereas in the healthy control group sad faces were preferentially mistaken for neutral *and* fearful faces over angry faces. In the healthy control group, fearful faces were also preferentially mistaken for angry or sad faces compared to neutral faces, but there were no differences in the pattern of preferentially misattributing fear for any of the other emotions in the BD group.

Table 1. Correlations between premorbid IQ and accuracy, response time and false alarm variables in bipolar disorder and healthy control groups

		Premorbid IQ			
		BIPOLAR DISORDER		**HEALTHY CONTROLS**	
Accuracy		***r***	***p***	***r***	***p***
	Fear	.10	.600	-.08	.682

	Angry	.14	.447	-.05	.773
	Sad	.06	.732	-.08	.678
	Neutral	.07	.721	-.09	.615
Response time					
	Fear	.01	.938	-.16	.393
	Angry	.06	.740	-.10	.571
	Sad	.15	.416	-.05	.780
	Neutral	-.02	.892	-.08	.684
False alarm					
Neutral foil	Fear target	-.05	.778	.07	.714
	Angry target	.04	.836	.05	.768
	Sad target	-.04	.822	.01	.939
Fear foil	Neutral target	-.13	.477	-.00	.994
	Angry target	.22	.225	-.04	.830
	Sad target	.03	.881	-.01	.967
Angry foil	Neutral target	.07	.701	-.02	.895
	Fear target	-.08	.659	.07	.723
	Sad target	-.11	.562	.03	.886
Sad foil	Neutral target	.06	.729	.03	.887
	Fear target	-.10	.585	-.09	.625
	Angry target	.05	.778	-.07	.700

Note. IQ = intelligence quotient.
Premorbid IQ was measured using the Wechsler Test of Adult Reading.
False alarm relates to the proportion of emotional foils mislabelled as the emotional target.
No correlations survived Bonferroni correction (adjusted $\alpha = .003$).

Table 2. Accuracy and response time means and standard deviations for facial emotion labelling

		BIPOLAR DISORDER		**HEALTHY CONTROLS**		**COMPARISONS**		
		M	***SD***	***M***	***SD***	**Cohen's d**	***F***	***p***
Accuracy (%)	Fear	75.15	15.43	85.25	9.38	-.79	10.17	**.002**
	Angry	72.18	15.85	80.76	11.80	-.61	6.16	.016
	Sad	71.63	14.99	77.99	14.39	-.43	3.09	.084
	Neutral	69.61	15.93	84.15	6.28	-1.20	23.25	**.000**
	Overall	72.14	14.60	82.04	9.52	-.80	8.23	**.000**
Response time (ms)†	Fear	1022.91	250.51	944.46	162.73	.37		
	Angry	960.46	196.24	912.25	160.80	.27		
	Sad	1021.33	211.13	964.71	168.64	.29		
	Neutral	990.32	198.44	912.12	189.28	.42		
	Overall	998.76	203.95	933.39	159.25	.35		

Note. Bolded values survive Bonferroni correction of *p* < .013.

†As there were no significant interaction effects observed for response time, between-group comparisons for separate emotions were not examined.

Table 3. Response time means and standard deviations across each run of the task

	BIPOLAR DISORDER		**HEALTHY CONTROLS**	
	M	***SD***	***M***	***SD***

Response time (ms)	Run 1	960.46	196.24	912.25	160.80
	Run 2	1021.33	211.13	964.71	168.64
	Run 3	1022.91	250.51	944.46	162.73
	Run 4	990.32	198.44	912.12	173.07

Note. The interaction between run*group was not significant, Wilks' lambda = .95, *F* (3, 62) = 1.01, *p* = .362.

Table 4. Means and standard deviations of facial emotion recognition accuracy in patients with bipolar disorder taking mood stabilisers, antipsychotics, and/or antidepressants

Medication type	**Emotion condition**	**On medication**			**Off medication**		
		n	**Accuracy (%)**		***n***	**% Accuracy**	
			M	***SD***		***M***	***SD***
Mood stabilisers	Fear	14	73.07	14.87	20	76.61	16.02
	Angry	14	74.63	16.62	20	70.47	15.50
	Sad	14	72.25	16.36	20	71.20	14.37
	Neutral	14	69.27	18.64	20	69.84	14.23
Antipsychotics	Fear	16	76.30	13.34	18	74.13	15.55
	Angry	16	72.66	16.000	18	71.76	16.18
	Sad	16	69.99	14.65	18	73.09	15.46
	Neutral	16	68.36	15.97	18	70.72	16.26
Antidepressants	Fear	12	74.83	14.14	22	75.33	16.41
	Angry	12	70.49	16.64	22	73.11	15.73
	Sad	12	69.70	17.68	22	72.68	13.63
	Neutral	12	70.05	17.60	22	69.37	15.37

Table 5. Means and standard deviations of facial emotion recognition response time in patients with bipolar disorder taking mood stabilisers, antipsychotics, and/or antidepressants

Medication type	**Emotion condition**	**On medication**			**Off medication**		
		n	**Response time (ms)**		***n***	**Response time (ms)**	
			M	***SD***		***M***	***SD***
Mood stabilisers	Fear	14	1021.21	316.85	20	1024.10	200.76
	Angry	14	959.96	205.19	20	960.82	195.14
	Sad	14	1034.17	241.94	20	1012.34	192.77
	Neutral	14	984.01	207.73	20	994.72	197.02
Antipsychotics	Fear	16	997.93	250.05	18	1045.12	256.01
	Angry	16	936.44	175.78	18	981.82	215.56
	Sad	16	998.84	204.09	18	1041.31	221.08
	Neutral	16	987.41	200.09	18	992.92	202.72
Antidepressants	Fear	12	1034.74	267.43	22	1016.46	247.04
	Angry	12	980.89	167.38	22	949.33	213.25
	Sad	12	1033.25	192.64	22	1014.82	224.68
	Neutral	12	1025.90	187.59	22	970.92	205.74

Table 6. Pairwise comparisons for emotion mislabelling within bipolar disorder and healthy control groups

Emotion foil	Within-group pairwise comparisons
Neutral	BD: Sad = Angry > Fear HC: Angry > Sad > Fear
Fear	HC: Angry = Sad > Neutral
Angry	-
Sad	BD: Neutral = Fear; Neutral > Angry; Fear = Angry

	HC: Fear = Neutral > Angry

Note. BD = bipolar disorder; HC = healthy controls.
Only significant within-group pairwise comparisons are displayed.

Table 7. Correlations between mood symptom severity and accuracy, response time and false alarm variables in the bipolar disorder group

		MADRS		YMRS	
Accuracy		***r***	***p***	***r***	***p***
	Fear	.15	.406	.12	.507
	Angry	.28	.103	.07	.701
	Sad	.42	.014	.12	.497
	Neutral	.26	.134	.14	.437
Response time					
	Fear	-.09	.630	.13	.457
	Angry	-.13	.472	.19	.295
	Sad	-.11	.523	.20	.268
	Neutral	-.08	.642	.10	.562
False alarm					
Neutral foil	Fear target	-.15	.395	-.28	.107
	Angry target	.24	.169	-.43	.012
	Sad target	-.05	.776	-.24	.165
Fear foil	Neutral target	-.32	.067	-.08	.651
	Angry target	.15	.403	-.22	.207
	Sad target	.06	.720	.20	.266
Angry foil	Neutral target	-.17	.328	-.03	.850
	Fear target	.11	.546	-.09	.618
	Sad target	-.31	.077	-.25	.158
Sad foil	Neutral target	-.27	.121	.03	.882
	Fear target	.17	.339	-.02	.927
	Angry target	-.12	.488	-.35	.043

Note. MADRS = Montgomery-Asberg Depression Rating Scale; YMRS = Young Mania Rating Scale.
False alarm relates to the proportion of emotional foils mislabelled as the emotional target.
No correlations survived Bonferroni correction (adjusted $\alpha = .003$).

References

Joshua, N., Van Rheenen, T.E., Castle, D.J., Rossell, S.L., 2016. Taking It at "Face Value": The Use of Face Processing Strategies in Bipolar Disorder and Schizophrenia. J. Int. Neuropsychol. Soc. 22, 652-661. DOI: 10.1017/S1355617716000412.

Van Rheenen, T.E., Joshua, N., Castle, D.J., Rossell, S.L., 2017. Configural and Featural Face Processing Influences on Emotion Recognition in Schizophrenia and Bipolar Disorder. J. Int. Neuropsychol. Soc. 23, 287-291. DOI: 10.1017/S1355617716001211.

Van Rheenen, T.E., Rossell, S.L., 2013. Is the non-verbal behavioural emotion-processing profile of bipolar disorder impaired? A critical review. Acta Psychiatr. Scand. 128, 163-178. DOI: 10.1111/acps.12125.