Title

Self processing in relation to emotion and reward processing in depression

Authors

Catherine Hobbs[a], Jie Sui[b], David Kessler[c], Marcus R Munafò[d e f], Katherine S Button[a]

[a] Department of Psychology, University of Bath

[b] School of Psychology, University of Aberdeen

[c] Population Health Sciences, University of Bristol

[d] School of Psychological Science, University of Bristol

[e] MRC Integrative Epidemiology Unit at the University of Bristol

[f] National Institute of Health Research Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol

Declarations

Funding: This study was funded by a GW4 BioMed MRC Doctoral Training Partnership award to Catherine Hobbs.

Acknowledgements: We would like to thank the participants who took part in this study.

Conflicts of Interest: None.

Author Contribution: CH and KSB conceptualised the study aims and design. JS and KSB provided materials for the study, and CH created the remaining materials. All authors obtained funding for the study. JS, DK, MRM and KSB provided supervisory support. CH recruited participants and collected data. CH cleaned, analysed and archived the data. CH wrote the original draft of this manuscript, which was reviewed and edited by JS, DK, MRM and KSB.

Ethical Declaration: The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects were approved by the University of Bath Department of Psychology Ethics Committee (18-167). Written informed consent was obtained from all participants.

Data Availability: The data that support the findings of this study are openly available in the University of Bath Research Data Archive (https://doi.org/10.15125/BATH-00924).

Abstract Word Count: 238 / 250

Text Word Count: 6232 / 4500

Abstract

Background: Depression is characterised by a heightened self-focus, which is believed to be associated with differences in emotion and reward processing. However, the precise relationship between these cognitive domains is not well understood. We examined the role of self-reference in emotion and reward processing, separately and in combination, in relation to depression.

Methods: Adults experiencing varying levels of depression (n = 144) completed self-report depression measures (PHQ-9, BDI-II). We measured self, emotion, and reward processing, separately and in combination, using three cognitive tasks.

Results: When self processing was measured independently of emotion and reward, in a simple associative learning task, there was little association with depression. However, when self and emotion processing occurred in combination in a self-esteem go/no-go task, depression was associated with an increased positive other bias (*b* = 3.51, 95% CI: 1.24, 5.79). When the self was processed in relation to emotion and reward, in a social evaluation learning task, depression was associated with reduced positive self biases (*b* = 0.11, 95% CI: 0.05, 0.17).

Conclusions: Depression was associated with enhanced positive implicit associations with others, and reduced positive learning about the self, culminating in reduced self-favouring biases. However, when self, emotion and reward processing occurred independently there was little evidence of an association with depression. Treatments targeting reduced positive self-biases may provide more sensitive targets for therapeutic intervention and potential biomarkers of treatment responses, allowing the development of more effective interventions.

Introduction

Depression is a highly prevalent mental health problem worldwide (World Health Organization, 2017), and is projected to be the leading cause of disease burden globally by 2030 (World Health Organization, 2011). Treatments for depression are moderately effective (Cipriani et al., 2018; Cuijpers, Andersson, Donker, & Van Straten, 2011), but individual response varies (Maslej, Furukawa, Cipriani, Andrews, & Mulsant, 2020). Understanding the cognitive processes maintaining depression may allow us to develop sensitive targets for therapeutic intervention. In this study we explored the role of self processing in depression, in relation to emotion and reward processing.

Self Processing

Across the general population, people show greater attention, recall and learning of self-related stimuli, often referred to as the self-prioritisation effect (Cunningham & Turk, 2017; Sui, He, & Humphreys, 2012; Sui & Humphreys, 2015a). However, individuals experiencing depression exhibit a heighted focus on the self, and difficulty disengaging from an internal self-referential focus (Northoff, 2007; Sheline et al., 2009). Paradoxically, this heightened internal self-referential focus may prevent individuals from associating novel stimuli with internal representations of the self (Sui, Ohrling, & Humphreys, 2016). This concept has previously been demonstrated in a study where following a negative mood induction participants were worse at associating arbitrarily assigned neutral shapes with the self (Sui et al., 2016). Individuals experiencing depression may subsequently be limited in their ability to update their self-concept from environmental feedback, perpetuating maladaptive views of the self.

The strength and consistency of self-prioritisation effects has led to proposals of the self being an integrative hub through which incoming stimuli is processed (Sui & Humphreys, 2015a). Targeting abnormalities in self-referential processing in depression may have wider implications for other cognitive domains implicated in depression. This is likely to include emotion (Ma & Han, 2010) and reward (Northoff & Hayes, 2011), as they are fundamental behavioural drivers and neurally overlap in the medial prefrontal cortex.

Self and Emotional Processing

Negative perceptions of the self are believed to play a causal role in the development of depression. According to Beck's cognitive theory, individuals experiencing depression develop negative views of the self as an internalised reaction to repeated adverse social experiences. When activated by stressful life events these negative self-schema dominate information processing, promoting automatic processing of negative information about the self (Beck, 2008). Supportive of this theory,

emotional biases are more likely to be observed in depression when stimuli is processed in reference to the self (Gaddy & Ingram, 2014; Hertel & El-Messidi, 2006; Ji, Grafton, & MacLeod, 2017). Altering negative information processing in relation to the self is therefore a key target for therapeutic interventions for depression.

Self and Reward Processing

Depression is also associated with a hyposensitivity to reward and hypersensitivity to punishment (Eshel & Roiser, 2010). There is evidence to suggest that this is linked to self processing. Self-relevant information induces activity in areas of the brain also activated during reward processing, such as the ventral medial prefrontal cortex, ventral striatum and ventral tegmental area (Northoff & Hayes, 2011). Differences in self-processing in depression may be at least partially driven differences in reward processing. In keeping with this theory, individuals with greater depression were found to continue to selectively engage in negative thoughts about the self even when this resulted in economic loss (Takano, Van Grieken, & Raes, 2019). Targeting reward processing in depression may have wider effects on self processing and vice versa.

Self, Emotion and Reward Processing

The interaction between self, emotion and reward processing may be a key combination of cognitive processes maintaining depression. Patients with depression show reduced activation of both reward and self-related areas of the brain when processing positive stimuli (Northoff, 2007; Northoff & Hayes, 2011). Reduced self-referential processing of positive information has also been identified as the most robust predictor of low approach motivation and reward responsivity (Hsu et al., 2020). Increased sensitivity to punishing feedback may sustain preferential processing of negative information about the self, reinforcing negative self-schema. Likewise, reduced sensitivity to positive feedback may reduce the ability to learn positive information about the self. The intersection between self, emotion and reward may therefore be the most effective target for cognitive treatments for depression.

Aims and Hypotheses

We explored the role of the self in relation to emotion and reward processing associated with varying levels of depressive symptoms. In contrast to previous studies that focused on either of these independent cognitive process (based on self, reward, or emotion) or interactions between any two components, we used three cognitive tasks to examine relationships between these processes and depressive symptoms, not only as distinct cognitive processes but also how they functionally interact.

To examine self, reward and emotion processing occurring independently we used associative learning tasks where participants paired neutral shapes with self-relevant, emotionally valenced, and varying degrees of reward, in three separate tasks. Based on previous research (Sui et al., 2016), we predicted that increased depression severity would be associated with worse performance when associating shapes with the self. Similarly based on evidence of impaired affective processing in depression (Dalgleish & Watts, 1990; Dalili, Penton-Voak, Harmer, & Munafò, 2015; Eshel & Roiser, 2010), we predicted that depression would be associated with worse performance when associating shapes with positive and rewarding stimuli.

To examine self, reward and emotion processing occurring in interaction we used a social reinforcement learning task where participants learnt when the computer liked themselves and others. Based on previous evidence (Hobbs et al., 2019), we hypothesised that increasing depression severity would be associated with worse learning of the self being 'liked'.

We also included a self-esteem go/no-go task due to its ability to integrate self and emotion processing. Participants rapidly categorised emotional and referential words, with greater discriminative accuracy believed to reflect existing implicit associations. An implicit negative self-esteem would therefore be reflected by greater discriminative accuracy when categorising self-referential and negative stimuli. However, due to mixed findings regarding the role of response inhibition in depression (Lewis, Button, Pearson, Munafò, & Lewis, 2020), and no previous use of this task within our research group we made no hypotheses regarding this task.

Methods

This study was pre-registered on the Open Science Framework (https://osf.io/34ma2), where study materials are also available. Study data are available in the University of Bath Research Data Archive (https://doi.org/10.15125/BATH-00924 ; Hobbs, Sui, Kessler, Munafò, & Button, 2020).

Participants

We recruited participants aged 18 to 65, fluent in English, with normal or corrected-to-normal vision, through campus advertising at the University of Bath. As depression severity is positively skewed (Tomitaka, Kawasaki, & Furukawa, 2015), to ensure balanced levels of depression we screened participants using the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001). We recruited an equal number of participants with no depression (PHQ-9 < 5), mild depression (PHQ-9 5-9) and moderate to severe depression (PHQ-9 $\geq$ 10).

Procedure

Participants completed two testing sessions, on average eight days apart (SD 3). At each session participants completed a social evaluation learning task, allowing measurement of test-retest reliability. To reduce fatigue effects associated with reaction time tasks, participants completed a go/no-go task at session one and an associative learning task at session two. At each session participants completed self-report measures of mood after the cognitive tasks.

Materials

Cognitive Tasks

To personalise tasks, prior to testing participants provided the first names of themselves, a friend, and a stranger.

Associative Learning Task

We used three simple associative learning tasks to measure how self, emotion and reward processing are independently associated with depression (Stolte, Humphreys, Yankouskaya, & Sui, 2017; Sui & Humphreys, 2015b). In each task, participants learnt to associate stimuli related to the relevant area of processing (e.g. Self: names of the self, a friend and a stranger; Emotion: happy, neutral and sad faces; Reward: £9, £3, £1), with abstract shapes. These tasks were completed sequentially in a counterbalanced order.

At the start of each task participants were instructed to learn randomly-assigned stimuli-shape pairings. Two blocks of 60 trials were completed per task. In each trial a fixation point was displayed

for 200 ms, followed by a stimuli-shape pairing presented for 100 ms (self, reward) or 150 ms (valence task only due to greater visual stimuli complexity). Participants pressed the 'n' or 'm' keys to indicate whether the presented pairings matched with the learnt association (Figure 1a). Key assignment to 'matching' or 'non-matching' responses was randomised for each participant but consistent across tasks. A response limit of 1100 ms was applied. Feedback was presented for each trial for 500 ms ("correct" / "incorrect" / "too slow"). At the end of each block participants were informed of their accuracy. For the reward task only, participants received a monetary reward based on the proportion of correct trials per reward stimuli.

Accuracy and reaction times (ms) were recorded. Prioritisation of stimuli is indicated by faster reaction times and/or higher accuracy.

Self-Esteem Go/No-Go Task

To measure how self and emotion processing occurring in interaction are associated with depression, we used a self-esteem go/no-go task. This task is proposed to measure implicit self-esteem (Gregg & Sedikides, 2010).

Participants were asked to categorise characteristics as positive (e.g. 'charming', 'smart') or negative (e.g. 'cruel', 'boring'), and referential worlds as related to the self, specified to participants as 'me' (e.g. participants' first name, 'me', 'I'), or others, specified to participants as 'not-me' (e.g. 'they', 'them', 'others'). In the training phase participants categorised words according to single categories (e.g. positive, negative, me, not-me), with 20 trials per condition. In the test phase, participants categorised words belonging to paired categories (e.g. positive OR me, positive OR not-me, negative OR me, negative OR not-me). There were 16 practice trials and 48 test trials for each paired combination of categories. An equal number of trials for stimuli relating to each condition was presented per block. A response timeout of 600 ms was applied. Block order was randomised.

At the beginning of each block the condition(s) by which words should be categorised was presented at the top of the screen and remained in place throughout the block. In each trial a word belonging to any of the conditions (e.g. positive, negative, me or not-me) was presented at the centre of the screen for 600 ms. Participants were asked to press the spacebar if the presented word related to the specified category (a 'go' response) or to refrain from pressing the spacebar if the word did not relate to the specified category (a 'no-go' response) (Figure 1b).

We categorised responses in test trials according to hits (a 'go' response when the stimuli belonged to the specified categories) and false alarms (a 'go' response when the stimuli did not belong to the specified categories). Responses to both referential and valence stimuli were included. For example,

if the specified categories were 'positive OR me' a trial was considered a hit if a 'go' response was given upon presentation of a positive characteristic *or* a self-referential word.

Discriminative accuracy (d') for each referential-emotion block was calculated by applying z-score transformations and subtracting hits from false alarms. Greater d' values indicate greater accuracy, suggesting stronger associations between paired-categories.

Social Evaluation Learning Task

To measure self, emotion and reward learning occurring simultaneously we used a reinforcement learning task within a social context (Button, Karwatowska, Kounali, Munafò, & Attwood, 2016; Button, Browning, Munafò, & Lewis, 2012; Button et al., 2015). Participants learnt how much the computer 'liked' the self, a friend and a stranger based on feedback to a forced choice selection between positive and negative social evaluation pairings (Figure 1c). A response time limit was not imposed. Participants learnt two rules based on the probability of the positive evaluations being 'correct' ('Like' 60-80%, 'Dislike' 20-40%). The number of errors made before reaching the criterion of eight consecutive rule-congruent responses were recorded. Bias scores were calculated by subtracting errors to criterion made when learning the dislike rule from the like rule. A positive value indicates a negative bias, as fewer errors were made learning the dislike rule compared to the like rule. We also calculated participants cumulative accuracy across trials in each condition-rule block to visualise learning curves.

After completing each rule block participants were also asked to provide a global rating of how much the computer liked the person, ranging from 'Complete Dislike' (0) to 'Complete Like' (10).

Participants completed all referential-conditions and rules. Order of referential-condition, and nested within this rule, was randomised. All participants completed 24 trials per referential-condition rule block.

[Figure 1 here]

Self-Report Measures

We measured depression severity using the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001) and Beck Depression Inventory (BDI-II) (Beck, Steer, & Brown, 1996). The PHQ-9 and BDI-II are self-administered questionnaires of the experience of depression within the previous two weeks. The PHQ-9 consists of nine items relating to the DSM-IV diagnostic criteria with scores ranging from 0-27, whereas the BDI-II consists of 21 items with scores ranging from 0-63 and has a greater focus on cognitive symptoms. Both measures demonstrate good psychometric properties (Cameron,

Crawford, Lawton, & Reid, 2008; Wang & Gorenstein, 2013), and are widely used in clinical practice (Kendrick et al., 2009).

We also identified whether participants met ICD-10 criteria for a primary diagnosis of a Major Depressive Episode (MDE) using the Clinical Interview Schedule-Revised (CIS-R; Lewis, Pelosi, Araya, & Dunn, 1992). The CIS-R is a fully structured self-administered computerised assessment that provides ICD-10 diagnoses of common mental health disorders. It has previously been used in large scale epidemiological studies within the general population.

As social anxiety has previously been associated with performance on the Social Evaluation Learning task (Button et al., 2015), we also measured social anxiety using the Brief Fear of Negative Evaluation Scale (Leary, 1983). To characterise the clinical profile of our sample we collected additional self-report measures of mental health and cognition. We measured anxiety using the Generalised Anxiety Disorder Scale (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006), anxiety relating to positive social feedback using the Fear of Positive Evaluation Scale (Weeks, Heimberg, & Rodebaugh, 2008), self-esteem using the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), and self-schema using the Dysfunctional Attitude Scale (DAS; Weissman & Beck, 1978). Finally, we measured change in state mood before and after completion of the cognitive tasks using the Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1988).

Statistical Analyses

Sample Size Calculation

A priori power calculations indicated that 144 participants would be required to provide greater than 80% power at an alpha level of 0.05 to detect previously observed effect sizes ($\eta^2 = 0.05$) for the relationship between bias scores in the self condition in the Social Evaluation Learning task and depression severity (Button et al., 2016, 2012, 2015; Hobbs et al., 2019), and greater than 99% power to detect previously observed effect sizes for the relationship between reaction times when matching shapes with the 'self' on the Associative Learning Task and depression severity ($\eta^2 = 0.17$) (Sui & Button, 2017).

Data Exclusion

Data was excluded according to a priori criteria as specified in our pre-registration.

For the associative learning task, trials with reaction times less than 200 ms (0.8%) and trials with no response (8%) were excluded. We included matching and non-matching trials in our analyses. For reaction time data we used both correct and incorrect responses.

We excluded 36 (25%) participants from the Go/No-Go Self-Esteem analyses due to a pattern of response indicating non-compliance (discrimination scores lower than 5 and/or bias scores less than 12 or greater than 36). As the exclusion rate was high, we repeated the main analyses for this task with all participants included as a sensitivity analysis.

Due to a technical error, data for the social evaluation learning task was unavailable in the second session for one participant.

Statistical Models

All analyses were conducted in R 3.6.

To aid interpretation we have provided both standardised (β) and unstandardised (b) regression coefficients.

We first assessed whether task performance differed across conditions using mixed-effects linear regression models. Separate models were used for each task, and for each measure of performance. Subject was entered as a random effect to account for within-subject effects. Task performance measures were entered as the outcome, and conditions as predictors.

Whilst the associative learning task and go/no-go task have previously been evidenced to have acceptable levels of reliability (Stolte, Humphreys, Yankouskaya, & Sui, 2016; Williams & Kaufmann, 2012), the reliability of the social evaluation learning task is yet to be tested. We calculated intraclass correlation coefficients for bias scores in the social evaluation learning task, using two-way mixed-effects models to calculate absolute agreement and consistency as recommended for cognitive-behavioural measures (Parsons, Kruijt, & Fox, 2019).

We used linear regression models to assess the relationship between task performance and depression. In all models, task outcomes were entered as separate predictors according to condition (e.g. in the self associative learning task accuracy model, accuracy in the self, friend and stranger condition were entered as separate predictors). We used depression as the outcome in these models, rather than a predictor as is typical in psychiatric experimental models, in preparation for future work using the cognitive task outcomes as predictors of change in depression severity. Separate models were conducted for each task outcome with PHQ-9 or BDI-II scores used as continuous outcomes. As the social evaluation learning task was completed in two sessions, we used mixed-effects linear regression models with session included as an additional predictor and subject as a random effect.

To examine the reliability of our findings for individuals meeting diagnostic criteria for depression, we repeated the primary analyses for each task using logistic regression models. Primary diagnosis of major depressive episode derived from the CIS-R was used as a binary outcome (diagnostic criteria met/not met). As the CIS-R was only completed at session 1, for tasks with multiple timepoints data from session 1 was used.

Full details of models are provided in the supplementary materials.

Results

Participant Characteristics

We recruited 144 participants, all of whom provided data for analysis. To demonstrate variability across depression severity, participant characteristics grouped according to PHQ-9 clinical cut-offs are presented in Table 1. The PHQ-9 and BDI-II showed excellent test-retest reliability between sessions (PHQ-9: ICC 0.94 (95% CI 0.89, 0.96), BDI-II: ICC: 0.96 (95% CI: 0.94, 0.97)), and strongly correlated (r = 0.90, 95% CI: 0.88, 0.92).

[Table 1 here]

Associative Learning Task

Hypothesis: Depression will be associated with reduced learning of self, highly rewarding and positive stimuli as indicated by reduced accuracy and greater reaction times.

Self

Consistent with prior evidence of self-prioritisation (Sui et al., 2012), participants on average showed the highest level of accuracy and fastest reaction times when matching shapes with the name of the self versus a friend or stranger (Supplementary Tables S1 and S2). We found no evidence to support our hypothesis; ability to associate shapes with the self, a friend or a stranger was not associated with depression (Table 2).

Reward

Likewise consistent with previous evidence of prioritisation of higher levels of reward (Sui & Humphreys, 2015b), participants on average were more accurate and faster when matching shapes with the highest level of reward (Supplementary Tables S1 and S2).

We found some evidence that increased accuracy when matching shapes with the medium level of reward was associated with greater depression. For every 1% increase in accuracy when matching shapes with '£3', PHQ-9 and BDI-II scores increased by 0.10 (b 95% CI: 0.02, 0.19, p = 0.021) and 0.24 (b 95% CI: 0.05, 0.43, p = 0.012) points respectively.

There was also weak evidence that decreasing accuracy when matching shapes with the high level of reward was associated with increased BDI-II scores (b = -0.19, b 95% CI: -0.37, 0.00, p = 0.051). However, confidence intervals overlapped with the null and there was little evidence of a similar relationship for PHQ-9 scores. We therefore found only weak support for our hypothesis.

No association was observed between accuracy when matching shapes with the low level of reward (£1) and depression severity (Table 2). We also found no relationship between reaction times and depression for this task (Table 2).

Emotion

Consistent with previous evidence of prioritisation of positive stimuli (Stolte et al., 2017), participants on average were more accurate and faster at matching shapes with happy faces (Supplementary Tables S1 and S2). However, in contrast with our hypothesis, accuracy or reaction times were not associated with depression (Table 2).

[Table 2 here]

Self-Esteem Go/No-Go Task

Due to previous mixed findings for the role of response inhibition in depression we made no hypothesis regarding this task, our findings should therefore be considered exploratory.

We found strong evidence of an interaction between referential condition and emotion on discriminative accuracy in the self-esteem Go/No-Go Task (b = 0.79, b 95% CI: 0.61, 0.97, β = 1.31, β 95% CI: 1.01, 1.61, p < .001; Supplementary Table S3). On average, participants showed a positive bias towards the self with greater discriminative accuracy for positive (M 1.40, SD 0.56) versus negative (M 1.0, SD 0.52) associations with the self. The opposite pattern was observed when associating words with the other (positive: M 0.71, SD 0.48, negative: M 1.12, SD 0.62).

We found consistent evidence that discriminative accuracy in the 'other' condition was associated with depression severity. Increased discriminative accuracy when associating positive words with others was associated with greater depression severity using both the PHQ-9 (b = 3.51, b 95% CI: 1.24, 5.79, β = 0.30, β 95% CI: 1.24, 5.79, p = 0.003) and BDI-II (b = 6.78, b 95% CI: 1.93, 11.64, β = 0.28, β 95% CI: 0.08, 0.47, p = 0.007). Conversely, increased discriminative accuracy when associating negative words with others was associated with lower PHQ-9 (b = -2.46, b 95% CI: -4.24, -0.68, β = -0.27, β 95% CI: -0.46, - 0.07, p = 0.007), and BDI-II scores (b = -5.13, b 95% CI: -8.92, -1.34, β = -0.27, β 95% CI: -0.46, - 0.07, p = 0.008). Individuals with greater depression therefore showed both a greater positive bias, and a reduced negative bias, when processing information about others.

Evidence for an association between discriminative accuracy in the self condition and depression was less consistent. Increased discriminative accuracy when associating positive words with the self was associated with a decrease in PHQ-9 scores (b = -2.47, b 95% CI: -4.54, -0.39, β = -0.24, β 95% CI: -0.44, -0.04, p = 0.020). Although an effect in the same direction was observed for BDI-II scores,

confidence intervals overlapped substantially with the null (b = -3.20, b 95% CI: -7.62, 1.23, β = -0.15, β 95% CI: -0.36, 0.06, p = 0.155). There was little evidence of an association between discriminative accuracy when associating negative words with the self with either the PHQ-9 (b = -0.59, b 95% CI: -2.57, 1.39, β = -0.05, β 95% CI: -0.24, 0.13, p = 0.553) or BDI-II (b = 0.81, b 95% CI: -5.03, 3.41, β = -0.04, β 95% CI: -0.22, 0.15, p = 0.704).

As we excluded a large proportion of participants (25%) in these analyses due to a priori criteria indicating non-compliance with the task, we repeated these analyses including all participants as a sensitivity analysis. We no longer found evidence for an association between discriminative accuracy in the other-negative condition and PHQ-9 severity, as confidence intervals overlapped with the null. However, the results described above persisted for all other associations (Supplementary Table S4).

Social Evaluation Learning

Hypothesis: Depression will be associated with reduced positive biases when learning about the self, driven by a greater number of errors before learning the positive 'like' rule.

Bias Scores

Participants on average were most positively biased when learning about the friend, making 2.07 fewer errors learning positive relative to negative evaluations (b 95% CI: -2.93, -1.21, β = -0.35, β 95% CI: -0.49, -0.20, p < .001), compared to when learning about the self. Participants displayed similar levels of bias when learning about the self and stranger (b = -0.44, b 95% CI: -1.31, 0.42, β = -0.07, β 95% CI: -0.22, 0.07, p = 0.318). The estimated agreement and consistency for bias scores across test sessions was ICC = 0.41 (95% CI: 0.29, 0.52).

In support of our hypothesis, bias scores when learning about the self were associated with depression severity. For every additional error learning the positive relative to the negative rule, PHQ-9 scores increased by 0.11 points (b 95% CI: 0.05, 0.17, p < .001) and BDI-II scores increased by 0.23 points (b 95% CI: 0.12, 0.34, p < .001). Effects were specific to learning about the self; bias scores when learning about the friend or a stranger were not associated with depression (Figure 2a; Table 3).

We also conducted additional exploratory analyses to examine whether the relationship between self bias scores and depression symptoms was consistent across sessions. We found little evidence of an interaction suggesting that the relationship did not vary over the two sessions (PHQ-9 b = 0.04, b 95% CI: -0.04, 0.11, β = 0.02, β 95% CI: -0.02, 0.06, p = 0.377; BDI-II b = 0.07, b 95% CI: -0.07, 0.21, β = 0.02, β 95% CI = -0.02, 0.06, p = 0.315).

[Figure 2 here]

Errors to Criterion

To investigate whether the relationship between bias scores and depression severity was driven by worse learning of the positive rule, or better learning of the negative rule, we examined the relationship between errors to criterion in each referential-rule condition and depression.

Participants overall were positively biased, making greater errors learning the negative versus positive rules (b = 1.45, b 95% CI: 0.82, 2.07, p < .001; Supplementary Table S5) The greater bias scores in the friend condition, as outlined above, was driven by participants making both fewer errors learning the positive rule (M 5.39, SD 3.76) and greater errors learning the negative rule (M 8.90, SD 4.24), compared to the self (positive M 6.50, SD 4.22; negative M 7.95 SD 4.28) and stranger (positive M 6.34 SD 3.90, negative M 8.23 SD 3.97) conditions.

We found consistent evidence to support our hypothesis that depression would be associated with a greater number of errors when learning the self-positive rule. For every additional error before learning the self-positive rule, PHQ-9 scores increased by 0.17 points (b 95% CI: 0.08, 0.26, p < .001) and BDI-II scores increased by 0.31 points (b 95% CI: 0.15, 0.47, p < .001).

We also found weak evidence that worse learning of the friend being disliked was associated with greater PHQ-9 scores, and better learning of the self being disliked was associated with reduced BDI-II scores (Table 3). However, confidence intervals were relatively wide, and these effects were not observed in the alternative depression measure for each, suggesting unreliable effects.

Errors to criterion when learning that a friend was liked, or either rule about the stranger, were not associated with PHQ-9 or BDI-II scores (Table 3).

Cumulative Accuracy

Figure 2b demonstrates the cumulative mean accuracy over the 24 learning trials for the positive 'like' and negative 'dislike' rules about the self in participants grouped according to none, mild, and moderate to severe levels of depression on the PHQ-9 and BDI-II. In keeping with our findings for errors to criterion, participants with moderate to severe levels of depression demonstrated impaired learning of the self-like rule as indicated by lower levels of mean accuracy both initially and cumulatively across trials.

Global Ratings

After each rule we asked participants to provide a global rating of how much the computer 'liked' the person.

Demonstrating understanding of each rule, participants gave lower global ratings following completion of the negative versus positive rules (b = -2.67, 95% CI: -2.85, -2.49, p < .001).

Additionally, participants showed slightly increased perceptions of the friend being liked compared to the self (b = 0.32, 95% CI: 0.14, 0.50, p = 0.001), but gave similar global ratings in the self and stranger conditions (b = 0.09, b 95% CI: -0.10, 0.27, p = 0.354). Full results are available in supplementary Table S5.

Consistent with our findings for errors to criterion, increased perceptions of being liked after completing the self-positive rule were associated with lower depression severity (Table 3). We also found weak evidence that greater global ratings in the stranger-positive condition was associated with greater PHQ-9 scores, however there was little evidence of this association with BDI-II scores (Table 3).

Social Anxiety

The effects outlined above persisted when social anxiety was taken into account, suggesting an independent relationship between social evaluation learning and depression (Supplementary Table S6).

[Table 3 here]

Reliability of findings with clinical diagnosis of depression

To examine whether our findings were valid for participants meeting clinical diagnostic criteria for depression, we repeated the primary analyses for each task using logistic regression models with primary diagnosis of major depressive episode, derived from the CIS-R, as a binary outcome. The primary effects of each task were replicated; increased positive biases towards others in the self-esteem go/no-go task and reduced positive biases towards the self in the social evaluation learning task, were associated with an increased odds of meeting diagnostic criteria for a major depressive episode. Full details are available in supplementary materials (Supplementary Tables S7-S9).

Adjusting for Age and Gender

The results of our primary analyses were consistent when we adjusted for age and gender (Supplementary Table S10).

Discussion

Depression is characterised by differences in processing self-related information, which are believed to be related to emotion and reward cognition. However, the precise relationship between these areas of processing is not yet well understood. In this study we examined the role of the self in emotion and reward processing, separately and in interaction, in individuals experiencing varying levels of depression. Healthy individuals typically show enhanced positive perceptions of the self, relative to others (De Jong, 2002). We found that when the self was processed in relation to emotion and reward, this self-favouring bias was reduced in individuals with greater depression severity. However, when self, emotion and reward processing occurred independently there was little evidence of an association with depression.

Using a social evaluation learning task, we found evidence of interaction between self, emotion, and reward processing with depression. During social interactions, healthy individuals preferentially incorporate positive evaluations into their self-concept (Korn, Prehn, Park, Walter, & Heekeren, 2012). In support of our pre-registered hypothesis, we found that participants with greater depression showed a reduced positive self-bias when learning social evaluations. Participants with greater depression made a greater number of errors before learning that they were 'liked' and gave lower global ratings of being liked. Depression was therefore consistently associated with a reduced ability to learn positive, socially rewarding information about the self.

Using a go/no-go task, we found that individuals with greater depression severity showed increased sensitivity to positive words in relation to others, and decreased sensitivity to negative words. However, in keeping with previous research using response inhibition tasks we found only weak evidence of an association between implicit self-esteem and depression (De Jong, Sportel, De Hullu, & Nauta, 2012; Franck, De Raedt, & De Houwer, 2008; Van Tuijl, De Jong, Sportel, De Hullu, & Nauta, 2014). Depression was therefore characterised by increased positive 'other-esteem', but not by an increased negative self-esteem. Our research adds to evidence suggesting that individuals with depression tend to perceive others more positively (Kuiper, Derry, & MacDonald, 1982). Depression has previously been theorised to originate from discrepancies between internal self-representations, and representations of the ideal self (Higgins, 1987). Enhanced positive perceptions of others may increase discrepancies between views of the actual and idealised self, perpetuating depressive symptoms. Alternatively, our findings of a weak association between implicit self-esteem and depression may reflect debate over the construct validity of implicit association tests (Hahn, Judd, Hirsh, & Blair, 2014), or questions over the extent to which affective response inhibition are associated with depression severity (Lewis et al., 2020).

When the self was processed independently of emotion or reward, within an associative learning task, we did not find evidence of changes in self-prioritisation with greater depression severity. This contrasts with previous findings of reduced self-prioritisation following negative mood induction (Sui et al., 2016). Whilst temporary, sudden changes in state mood may inhibit self-prioritisation in the absence of emotional processing, this does not seem to apply to low trait mood. We also found no evidence that depression was associated with differences in learning emotional associations when processed independently of the self. There were some indications of differences in reward learning associated with depression. Although, in contrast to our expectations this was only observed for medium levels of reward. It is possible that depression alters sensitivity to reward, with greater value being placed on lower levels of reward. However, confidence intervals were relatively wide for this effect. Further research replicating these results is therefore required in order to understand their importance.

A substantial body of research suggests that healthy individuals hold relatively enhanced perceptions of the self versus others (Kuiper et al., 1982), and typically rate their abilities as better-than-average (Zell, Strickhouser, Sedikides, & Alicke, 2019). These positive self-biases are believed to be beneficial for mental health in increasing self-esteem and confidence (Button et al., 2015). Our results indicate that when processed independently of emotion, at least at a 'cold' perceptual level as in the associative learning task, self-referential processing is similar irrespective of depression severity. However, differences were observed when integrating positive and negative information with the self and others. Overall, depression was characterised by a reduction in self-favouring biases. Individuals with greater depression showed both greater implicit positive perceptions of others, and impaired learning of positive associations with the self. Depression may be driven by other-favouring biases strengthened by reduced learning of positive information about the self. In combination, reduced positive perceptions of the self and enhanced positive perceptions of others are likely to maintain negative views of the self.

Clinical Implications

Acknowledging that much of the work in therapy already implies self-reference, our findings suggest that it may be beneficial to explicitly manipulate referential focus and target biases in emotion and reward processing in relation to the self. Social evaluation learning in particular may be an important target for intervention. Depression is associated with poorer quality social interactions (Teo, Choi, & Valenstein, 2013), and social withdrawal (Hirschfeld et al., 2000). Our findings suggest that individuals with depression show a stable pattern of reduced learning of positive evaluations about the self. Reduced positive self-biases in social interactions are likely to maintain negative perceptions

of the self, reinforcing social withdrawal and increasing the likelihood of poor social relationships, subsequently maintaining depression symptoms (Lewinsohn, Mischel, Chaplin, & Barton, 1980). Social evaluation learning provides an important and potentially reversible target for therapeutic intervention that can address impairments in social functioning, negative perceptions of the self, and wider depressive symptoms. It is also possible that social evaluation learning may be a transdiagnostic mechanism. Future research examining latent mental health traits would allow us to understand the importance of social evaluation learning across mental health disorders.

Additionally, we found evidence that the relationship between biased learning about the self and depression was consistent across testing sessions. Change in social evaluation learning may therefore be a viable predictor of change in depressive symptoms. Individual treatment response for depression is varied (Maslej et al., 2020). It is currently difficult to predict which treatments are effective at an individual level (Simon & Perlis, 2010). Exacerbating these difficulties are the long time periods between commencing treatment and improvement in mood (Uher et al., 2011). Identifying markers of therapeutic change would be beneficial in allowing identification of effective treatments at an earlier timepoint. Further research examining changes in learning positive evaluations about the self as a potential predictor of treatment response would be beneficial.

Limitations

We recruited participants based on depression severity to gain a balanced range of depression. However, in the time between screening and testing, depression severity on average decreased potentially weakening our effects. In-depth analysis of larger samples representative of the spectrum of individuals with depression would be fruitful to further characterise changes in self-referential processing and to replicate the current findings. Although, our results were replicated for individuals meeting diagnostic criteria for depression, suggesting that our results are reliable for greater severities of depression.

Additionally, whilst our sample was representative of the range of depressive symptoms experienced in the general population it was limited in its demographic diversity. Participants were predominantly young, students and female. While this may be an ideal sample to investigate the role of self biases in depression, given the worrying increase of depression in this population at a developmentally sensitive time where self-identity and peer relations are evolving (Blakemore & Mills, 2014; Royal College of Psychiatrists, 2011), future studies should investigate whether these finding generalise across the wider population and test whether the strength of the associations alter across adulthood.

Whilst we found evidence of a consistent relationship between biased learning about the self and depression in the social evaluation learning task, bias scores themselves showed limited reliability between test sessions. Further development of this task to improve reliability would be beneficial.

Finally, this was a cross-sectional study examining the association between self, emotion and reward processing with depression. We are therefore unable to comment on the causal role of self processing in relation to emotion and reward. Future research examining the longitudinal relationship between self processing and depression would provide insight into the potential causal role of reduced positive self-biases. Additionally, manipulating self-referential affective processing through cognitive bias modification would help us understand the importance of this cognitive style in maintaining depression symptoms.

Conclusion

Overall, our findings suggest that depression is characterised by enhanced positive implicit associations with others, and reduced positive learning about the self, culminating in reduced self-favouring biases observed in healthy individuals. We also found some evidence of altered sensitivity to reward in individuals with greater depression severity using a simple associative learning task, although this effect requires further replication. Treatments targeting reduced positive self-biases may provide more sensitive targets for therapeutic intervention and potential biomarkers of treatment responses, allowing the development of more effective interventions.

564 **References**

565 Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological
566 correlates. *American Journal of Psychiatry*, *165*(8), 969–977.
567 https://doi.org/10.1176/appi.ajp.2008.08050721

568 Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck depression inventory-II. *San*
569 *Antonio, TX: Psychological Corporation*.

570 Blakemore, S. J., & Mills, K. L. (2014). Is adolescence a sensitive period for sociocultural processing?
571 *Annual Review of Psychology*, *65*, 187–207. https://doi.org/10.1146/annurev-psych-010213-
572 115202

573 Button, K. S., Browning, M., Munafò, M. R., & Lewis, G. (2012). Social inference and social anxiety:
574 Evidence of a fear-congruent self-referential learning bias. *Journal of Behavior Therapy and*
575 *Experimental Psychiatry*, *43*(4), 1082–1087. https://doi.org/10.1016/j.jbtep.2012.05.004

576 Button, K. S., Karwatowska, L., Kounali, D., Munafò, M. R., & Attwood, A. S. (2016). Acute anxiety and
577 social inference: An experimental manipulation with 7.5% carbon dioxide inhalation. *Journal of*
578 *Psychopharmacology*, *30*(10), 1036–1046. https://doi.org/10.1177/0269881116653105

579 Button, K. S., Kounali, D., Stapinski, L., Rapee, R. M., Lewis, G., & Munafò, M. R. (2015). Fear of
580 negative evaluation biases social evaluation inference: Evidence from a probabilistic learning
581 task. *PLoS ONE*, *10*(4), e0119456. https://doi.org/10.1371/journal.pone.0119456

582 Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9
583 and HADS for measuring depression severity in primary care. *British Journal of General Practice*,
584 *58*(546), 32–36. https://doi.org/10.3399/bjgp08X263794

585 Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., … Geddes, J. R.
586 (2018). Comparative Efficacy and Acceptability of 21 Antidepressant Drugs for the Acute
587 Treatment of Adults With Major Depressive Disorder: A Systematic Review and Network Meta-
588 Analysis. *FOCUS*, *16*(4), 420–429. https://doi.org/10.1176/appi.focus.16407

589 Cuijpers, P., Andersson, G., Donker, T., & Van Straten, A. (2011). Psychological treatment of
590 depression: Results of a series of meta-analyses. *Nordic Journal of Psychiatry*, *65*(6), 354–364.
591 https://doi.org/10.3109/08039488.2011.596570

592 Cunningham, S. J., & Turk, D. J. (2017). A review of self-processing biases in cognition. *Quarterly*
593 *Journal of Experimental Psychology*, *70*(6), 987–995.
594 https://doi.org/10.1080/17470218.2016.1276609

595 Dalgleish, T., & Watts, F. N. (1990). Biases of attention and memory in disorders of anxiety and
596 depression. *Clinical Psychology Review*, *10*(5), 589–604. https://doi.org/10.1016/0272-
597 7358(90)90098-U

598 Dalili, M. N., Penton-Voak, I. S., Harmer, C. J., & Munafò, M. R. (2015). Meta-analysis of emotion
599 recognition deficits in major depressive disorder. *Psychological Medicine*, *45*(6), 1135–1144.
600 https://doi.org/10.1017/S0033291714002591

601 De Jong, P. J. (2002). Implicit self-esteem and social anxiety: Differential self-favouring effects in high
602 and low anxious individuals. *Behaviour Research and Therapy*, *40*(5), 501–508.
603 https://doi.org/10.1016/S0005-7967(01)00022-5

604 De Jong, P. J., Sportel, B. E., De Hullu, E., & Nauta, M. H. (2012). Co-occurrence of social anxiety and
605 depression symptoms in adolescence: Differential links with implicit and explicit self-esteem?
606 *Psychological Medicine*, *42*(3), 475. https://doi.org/10.1017/S0033291711001358

607 Eshel, N., & Roiser, J. P. (2010). Reward and punishment processing in depression. *Biological*
608 *Psychiatry*, *68*(2), 118–124. https://doi.org/10.1016/j.biopsych.2010.01.027

609 Franck, E., De Raedt, R., & De Houwer, J. (2008). Activation of latent self-schemas as a cognitive
610 vulnerability factor for depression: The potential role of implicit self-esteem. *Cognition and*
611 *Emotion*, *22*(8), 1588–1599. https://doi.org/10.1080/02699930801921271

612 Gaddy, M. A., & Ingram, R. E. (2014). A meta-analytic review of mood-congruent implicit memory in
613 depressed mood. *Clinical Psychology Review*, *34*(5), 402–416.
614 https://doi.org/10.1016/j.cpr.2014.06.001

615 Gregg, A. P., & Sedikides, C. (2010). Narcissistic fragility: Rethinking its links to explicit and implicit
616 self-esteem. *Self and Identity*, *9*(2), 142–161. https://doi.org/10.1080/15298860902815451

617 Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of*
618 *Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

619 Hertel, P. T., & El-Messidi, L. (2006). Am I Blue? Depressed Mood and the Consequences of Self-
620 Focus for the Interpretation and Recall of Ambiguous Words. *Behavior Therapy*, *37*(3), 259–
621 268. https://doi.org/10.1016/j.beth.2006.01.003

622 Higgins, E. T. (1987). Self-Discrepancy: A Theory Relating Self and Affect. *Psychological Review*, *94*(3),
623 319. https://doi.org/10.1037/0033-295X.94.3.319

624 Hirschfeld, R., Montgomery, S., Keller, M., Kasper, S., Schatzberg, A., Moller, H.-J., … Bourgeois, M.

(2000). Social functioning in depression: a review. *Journal of Clinical Psychiatry*, *61*(4), 268–275.

Hobbs, C., Faraway, J., Kessler, D., Munafò, M. R., Sui, J., & Button, K. S. (2019). Social evaluation learning in social anxiety disorder and depression: a mega-analysis presented at the Summer Meeting of the Bristish Association for Psychopharmacology. *Journal of Psychopharmacology*, *Supplement*(8), A33.

Hobbs, C., Sui, J., Kessler, D., Munafò, M. R., & Button, K. S. (2020). *Dataset for 'Self processing in relation to emotion and reward processing in depression*. University of Bath Research Data. https://doi.org/https://doi.org/10.15125/BATH-00924

Hsu, K. J., McNamara, M. E., Shumake, J., Stewart, R. A., Labrada, J., Alario, A., … Beevers, C. G. (2020). Neurocognitive predictors of self-reported reward responsivity and approach motivation in depression: A data-driven approach. *Depression and Anxiety*, *37*, 682–697. https://doi.org/10.1002/da.23042

Ji, J. L., Grafton, B., & MacLeod, C. (2017). Referential focus moderates depression-linked attentional avoidance of positive information. *Behaviour Research and Therapy*, *93*, 47–54. https://doi.org/10.1016/j.brat.2017.03.004

Kendrick, T., Dowrick, C., McBride, A., Howe, A., Clarke, P., Maisey, S., … Smith, P. W. (2009). Management of depression in UK general practice in relation to scores on depression severity questionnaires: Analysis of medical record data. *BMJ (Online)*, *338*(7697), b750. https://doi.org/10.1136/bmj.b750

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Kuiper, N. A., Derry, P. A., & MacDonald, M. R. (1982). Self-reference and person perception in depression: A social cognition perspective. In *Integrations of clinical and social psychology* (pp. 79–2103).

Leary, M. R. (1983). A Brief Version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, *9*(3), 371–375. https://doi.org/10.1177/0146167283093007

Lewinsohn, P. M., Mischel, W., Chaplin, W., & Barton, R. (1980). Social competence and depression: The role of illusory self-perceptions. *Journal of Abnormal Psychology*, *89*(2), 203.

https://doi.org/10.1037/0021-843X.89.2.203

Lewis, G., Button, K. S., Pearson, R. M., Munafò, M. R., & Lewis, G. (2020). Inhibitory control of positive and negative information and adolescent depressive symptoms: a population-based cohort study. *Psychological Medicine*, 1–11. https://doi.org/https://doi.org/10.1017/S0033291720002469

Lewis, G., Pelosi, A., Araya, R., & Dunn, G. (1992). Measuring psychiatric disorder in the community: A standardized assessment for use by lay interviewers. *Psychological Medicine*, *22*(2), 465–486. https://doi.org/10.1017/S0033291700030415

Ma, Y., & Han, S. (2010). Why we respond faster to the self than to others? An implicit positive association theory of self-advantage during implicit face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 619–633. https://doi.org/10.1037/a0015797

Maslej, M. M., Furukawa, T. A., Cipriani, A., Andrews, P. W., & Mulsant, B. H. (2020). Individual differences in response to antidepressants: a meta-analysis of placebo-controlled randomized clinical trials. *JAMA Psychiatry*, *77*(6), 607–617. https://doi.org/10.1001/jamapsychiatry.2019.4815

Northoff, G. (2007). Psychopathology and pathophysiology of the self in depression - Neuropsychiatric hypothesis. *Journal of Affective Disorders*, *104*(1–3), 1–14. https://doi.org/10.1016/j.jad.2007.02.012

Northoff, G., & Hayes, D. J. (2011). Is our self nothing but reward? *Biological Psychiatry*, *69*(11), 1019–1025. https://doi.org/10.1016/j.biopsych.2010.12.014

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Rosenberg, M. (1965). Rosenberg self esteem scle. *Personality and Individual Differences*. https://doi.org/10.1007/s12671-015-0407-6

Royal College of Psychiatrists. (2011). *Mental health of students in higher education*. https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/college-report-cr166.pdf?sfvrsn=d5fa2c24_2

Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., … Raichle, M. E. (2009). The default mode network and self-referential processes in depression. *Proceedings of*

687 *the National Academy of Sciences of the United States of America*, *106*(6), 1942–1947.
688 https://doi.org/10.1073/pnas.0812686106

689 Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: Can we match patients
690 with treatments? *American Journal of Psychiatry*, *167*(12), 1445–1455.
691 https://doi.org/10.1176/appi.ajp.2010.09111680

692 Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing
693 Generalized Anxiety Disorder. *Archives of Internal Medicine*, *166*(10), 1092–1097.
694 https://doi.org/10.1001/archinte.166.10.1092

695 Stolte, M., Humphreys, G., Yankouskaya, A., & Sui, J. (2016). Dissociating biases towards the self and
696 positive emotion. *Quarterly Journal of Experimental Psychology*, *70*(6), 1011–1022.
697 https://doi.org/10.1080/17470218.2015.1101477

698 Stolte, M., Humphreys, G., Yankouskaya, A., & Sui, J. (2017). Dissociating biases towards the self and
699 positive emotion. *Quarterly Journal of Experimental Psychology*, *70*(6), 1011–1022.
700 https://doi.org/10.1080/17470218.2015.1101477

701 Sui, J., & Button, K. S. (2017). Associative Learning and Depression. *Unpublished Data*.

702 Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-
703 prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human*
704 *Perception and Performance*, *38*(5), 1105. https://doi.org/10.1037/a0029792

705 Sui, J., & Humphreys, G. W. (2015a). The Integrative Self: How Self-Reference Integrates Perception
706 and Memory. *Trends in Cognitive Sciences*, *19*(12), 719–728.
707 https://doi.org/10.1016/j.tics.2015.08.015

708 Sui, J., & Humphreys, G. W. (2015b). The interaction between self-bias and reward: Evidence for
709 common and distinct processes. *Quarterly Journal of Experimental Psychology*, *68*(10), 1952–
710 1964. https://doi.org/10.1080/17470218.2015.1023207

711 Sui, J., Ohrling, E., & Humphreys, G. W. (2016). Negative mood disrupts self- and reward-biases in
712 perceptual matching. *Quarterly Journal of Experimental Psychology*, *69*(7), 1438–1448.
713 https://doi.org/10.1080/17470218.2015.1122069

714 Takano, K., Van Grieken, J., & Raes, F. (2019). Difficulty in updating positive beliefs about negative
715 cognition is associated with increased depressed mood. *Journal of Behavior Therapy and*
716 *Experimental Psychiatry*, *64*, 22–30. https://doi.org/10.1016/j.jbtep.2019.02.001

717 Teo, A. R., Choi, H. J., & Valenstein, M. (2013). Social Relationships and Depression: Ten-Year Follow-
718 Up from a Nationally Representative Study. *PLoS ONE*, *8*(4), e62396.
719 https://doi.org/10.1371/journal.pone.0062396

720 Tomitaka, S., Kawasaki, Y., & Furukawa, T. (2015). A distribution model of the responses to each
721 depressive symptom item in a general population: A cross-sectional study. *BMJ Open*, *5*(9),
722 e008599. https://doi.org/10.1136/bmjopen-2015-008599

723 Uher, R., Mors, O., Rietschel, M., Rajewska-Rager, A., Petrovic, A., Zobel, A., … McGuffin, P. (2011).
724 Early and Delayed Onset of Response to Antidepressants in Individual Trajectories of Change
725 During Treatment of Major Depression. *The Journal of Clinical Psychiatry*, *72*(11), 1586–1592.
726 https://doi.org/10.4088/jcp.10m06419

727 Van Tuijl, L. A., De Jong, P. J., Sportel, B. E., De Hullu, E., & Nauta, M. H. (2014). Implicit and explicit
728 self-esteem and their reciprocal relationship with symptoms of depression and social anxiety: A
729 longitudinal study in adolescents. *Journal of Behavior Therapy and Experimental Psychiatry*,
730 *45*(1), 113–121. https://doi.org/10.1016/j.jbtep.2013.09.007

731 Wang, Y. P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A
732 comprehensive review. *Brazilian Journal of Psychiatry*, *35*(4), 416–431.
733 https://doi.org/10.1590/1516-4446-2012-1048

734 Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of
735 positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*,
736 *54*(6), 1063. https://doi.org/10.1037//0022-3514.54.6.1063

737 Weeks, J. W., Heimberg, R. G., & Rodebaugh, T. L. (2008). The Fear of Positive Evaluation Scale:
738 Assessing a proposed cognitive component of social anxiety. *Journal of Anxiety Disorders*, *22*(1),
739 44–55. https://doi.org/10.1016/j.janxdis.2007.08.002

740 Weissman, A. N., & Beck, A. T. (1978). Development and Validation of the Dysfunctional Attitude
741 Scale: A Preliminary Investigation. *Acceptance and Commitment Therapy*, *54*.

742 Williams, B. J., & Kaufmann, L. M. (2012). Reliability of the Go/No Go Association Task. *Journal of*
743 *Experimental Social Psychology*, *48*(4), 879–891. https://doi.org/10.1016/j.jesp.2012.03.001

744 World Health Organization. (2011). Global burden of mental disorders and the need for a
745 comprehensive, coordinated response from health and social sectors at the country level.
746 *EB130/9*. https://doi.org/https://apps.who.int/gb/ebwha/pdf_files/eb130/b130_9-en.pdf

747 World Health Organization. (2017). Depression and Other Common Mental Disorders: Global Health

748 Estimates. *No. WHO/MSD/MER/2017.2*.
749 https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf
750 Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2019). The Better-Than-Average Effect in
751 Comparative Self- Evaluation: A Comprehensive Review and Meta-Analysis. *Psychological*
752 *Bulletin*, *146*(2), 118. https://doi.org/10.1037/bul0000218

753

754

755 Table 1

756 *Participant Characteristics according to Depression Severity*

	PHQ-9 Depression Severity		
	None (<5)	**Mild (5-9)**	**Moderate to Severe (≥10)**
N (%)	48 (33)	56 (39)	40 (28)
Age, M (SD)	23.4 (7.3)	22.6 (7.9)	20.9 (3.1)
Gender, N (%)			
Male	11 (23)	11 (19)	9 (22)
Female	37 (77)	44 (79)	31 (78)
Other	0 (0)	1 (2)	0 (0)
Ethnicity, N (%)			
White	33 (69)	30 (54)	33 (85)
Black	0 (0)	2 (4)	1 (3)
Asian	11 (23)	18 (32)	3 (7)
Mixed	4 (8)	4 (7)	2 (5)
Other	0 (0)	2 (3)	0 (0)
Employment, N (%)			
Student	42 (88)	50 (89)	36 (90)
Employed	5 (10)	6 (11)	3 (8)
Other	1 (2)	0 (0)	1 (2)
CIS-R Primary Diagnosis Major Depressive Episode, N (%)	0 (0)	9 (16)	26 (65)[a]
Current Treatment, N (%)			
Psychological Therapy	0 (0)	3 (5)	5 (13)
Antidepressants	0 (0)	2 (4)	7 (18)
PHQ-9, M (SD)	2.5 (1.2)	6.9 (1.4)	15.0 (4.0)
BDI-II, M (SD)	4.6 (3.6)	13.1 (5.6)	27.2 (10.5)
BFNE, M (SD)	34.3 (10.2)	38.8 (9.1)	45.9 (8.2)
GAD-7, M (SD)	2.1 (2.1)	5.4 (3.0)	10.9 (4.1)
FPE, M (SD)	23.2 (11.1)	26.8 (13.5)	36.5 (14.2)
DAS-24, M (SD)	90.3 (17.8)	94.9 (18.5)	108.3 (15.5)

RSES, M (SD)	13.6 (1.9)	12.9 (2.5)	12.7 (2.1)
PANAS Positive Change, M (SD)	-1.5 (3.2)	-1.9 (3.3)	-1.9 (4.2)
PANAS Negative Change, M (SD)	-0.7 (2.1)	-0.7 (2.2)	-1.1 (4.0)

[a] Participants who met criteria for a primary diagnosis of a MDE within this group had higher PHQ-9 (M 16.21, SD 4.35) and BDI-II scores (M 31.88, SD 10.42), compared to those that did not have a primary diagnosis of a MDE (PHQ-9: M 12.00, SD 1.83, BDI-II: 19.57, SD 5.95).

CIS-R = Clinical Interview Schedule Revised, PHQ-9 = Patient Health Questionnaire, BDI-II = Beck Depression Inventory, BFNE = Brief Fear of Negative Evaluation, GAD-7 = Generalised Anxiety Questionnaire, BFNE = Brief Fear of Negative Evaluation Scale, FPE = Fear of Positive Evaluation Scale, DAS-24 = Dysfunctional Attitudes Scale, RSES = Rosenberg Self-Esteem Scale, PANAS = Positive and Negative Affect Schedule.

Note: All data presented in this table were collected at the first testing session. PANAS change scores reflect differences in scores from pre- to post-completion of the cognitive tasks.

766 Table 2

767 *Results from linear regression models examining the association between accuracy and reaction times for each task condition (predictors) in the associative*

768 *learning task with depression (Outcome: PHQ-9/BDI-II)*

Task	Stimuli	PHQ-9					BDI-II				
		b	*b* 95% CI	β	β 95% CI	p	*b*	*b* 95% CI	β	β 95% CI	p
Accuracy (%)											
Self	Intercept	11.44	3.38, 19.49	0.00	-0.16, 0.16	0.006	14.51	-2.78, 31.79	0.00	-0.17, 0.17	0.099
	Self	-0.06	-0.17, 0.05	-0.12	-0.33, 0.10	0.288	-0.15	-0.39, 0.10	-0.13	-0.35, 0.08	0.231
	Friend	-0.04	-0.15, 0.06	-0.09	-0.32, 0.13	0.414	0.03	-0.20, 0.26	0.03	-0.20, 0.26	0.790
	Stranger	0.05	-0.04, 0.15	0.13	-0.10, 0.36	0.279	0.11	-0.09, 0.32	0.13	-0.10, 0.36	0.275
Reward	Intercept	6.07	1.08, 11.06	0.00	-0.16, 0.16	0.018	8.59	-2.02, 19.20	0.00	-0.16, 0.16	0.112
	High (£9)	-0.06	-0.15, 0.03	-0.18	-0.43, 0.07	0.166	-0.19	-0.37, 0.00	-0.25	-0.50, 0.00	0.051
	Medium (£3)	0.10	0.02, 0.19	0.30	0.05, 0.56	0.021	0.24	0.05, 0.43	0.33	-0.07, 0.58	0.012
	Low (£1)	-0.03	-0.10, 0.04	-0.10	-0.31, 0.11	0.366	0.02	-0.13, 0.16	0.02	-0.18, 0.23	0.814
Emotion	Intercept	6.05	1.05, 11.04	0.00	-0.17, 0.17	0.018	10.72	0.05, 21.39	0.00	-0.17, 0.17	0.049
	Happy	-0.02	-0.09, 0.05	-0.06	-0.27, 0.15	0.588	-0.05	-0.21, 0.11	-0.06	-0.28, 0.15	0.547
	Neutral	0.03	-0.05, 0.11	0.08	-0.15, 0.32	0.498	0.06	-0.11, 0.23	0.08	-0.15, 0.32	0.489
	Sad	0.01	-0.07, 0.08	0.02	-0.22, 0.25	0.881	0.04	-0.13, 0.20	0.05	-0.18, 0.28	0.668
Reaction Times (ms)											
Self	Intercept	11.50	2.64, 20.37	0.00	-0.17, 0.17	0.011	24.45	5.50, 43.40	0.00	-0.16, 0.16	0.012
	Self	0.00	-0.02, 0.03	0.05	-0.27, 0.37	0.755	0.00	-0.05, 0.06	0.01	-0.31, 0.34	0.929

	Friend	-0.01	-0.04, 0.01	-0.20	-0.60, 0.19	0.317	-0.04	-0.09, 0.02	-0.24	-0.64, 0.15	0.277
	Stranger	0.00	-0.03, 0.03	0.04	-0.39, 0.48	0.846	0.02	-0.05, 0.08	0.11	-0.32, 0.55	0.610
Reward	Intercept	4.53	-2.38, 11.44	0.00	-0.17, 0.17	0.197	7.89	-6.88, 22.65	0.00	-0.17, 0.17	0.293
	High (£9)	0.01	-0.01, 0.04	0.22	-0.15, 0.59	0.245	0.03	-0.01, 0.08	0.26	-0.11, 0.63	0.168
	Medium (£3)	-0.01	-0.04, 0.02	-0.19	-0.65, 0.27	0.422	-0.01	-0.07, 0.04	-0.09	-0.55, 0.36	0.685
	Low (£1)	0.00	-0.02, 0.02	0.02	-0.36, 0.40	0.933	-0.01	-0.06, 0.03	-0.11	-0.49, 0.26	0.549
Emotion	Intercept	7.51	1.64, 13.37	0.00	-0.17, 0.17	0.013	14.33	1.75, 26.91	0.00	-0.17, 0.17	0.026
	Happy	0.01	-0.01, 0.02	0.10	-0.28, 0.48	0.614	0.00	-0.04, 0.04	0.02	-0.36, 0.41	0.898
	Neutral	0.00	-0.03, 0.02	-0.06	-0.55, 0.44	0.824	0.00	-0.05, 0.05	0.00	-0.49, 0.50	0.990
	Sad	0.00	-0.03, 0.02	-0.07	-0.57, 0.44	0.793	0.00	-0.05, 0.04	-0.04	-0.55, 0.46	0.867

769 *b* = unstandardised regression coefficients, β = standardised regression coefficients

770

Table 3

Results from mixed-effect linear regression models examining the relationship between social evaluation learning task outcomes (predictors) and depression (Outcome: PHQ-9/BDI-II)

	PHQ-9					BDI-II				
	b	***b* 95% CI**	**β**	**95% CI**	**p**	***b***	***b* 95% CI**	**β**	**95% CI**	**p**
Bias Scores										
Intercept	8.54	7.47, 9.60	0.00	-0.15, 0.15	< .001	15.18	13.06, 17.30	0.00	-0.16, 0.15	< .001
Self	0.11	0.05, 0.17	0.13	0.06, 0.20	< .001	0.23	0.12, 0.34	0.13	0.07, 0.19	< .001
Friend	-0.03	-0.09, 0.03	-0.04	-0.11, 0.01	0.259	0.01	-0.10, 0.11	0.00	-0.05, 0.06	0.898
Stranger	-0.01	-0.08, 0.05	-0.01	-0.08, 0.05	0.731	0.00	-0.12, 0.11	0.00	-0.06, 0.05	0.943
Session	-0.88	-1.29, -0.46	-0.08	-0.12, -0.04	< .001	-0.73	-1.47, 0.02	-0.03	-0.06, 0.00	0.057
Errors to Criterion										
Intercept	7.45	5.91, 8.99	0.00	-0.15, 0.15	< .001	13.79	10.84, 16.73	0.00	-0.15, 0.15	< .001
Self-Positive	0.17	0.08, 0.26	0.13	0.06, 0.20	< .001	0.31	0.15, 0.47	0.12	0.06, 0.18	< .001
Self-Negative	-0.05	-0.13, 0.04	-0.04	-0.10, 0.03	0.264	-0.17	-0.32, -0.02	-0.06	-0.12, -0.01	0.031
Friend-Positive	0.03	-0.05, 0.16	0.02	-0.04, 0.08	0.492	0.01	-0.14, 0.16	0.00	-0.05, 0.05	0.916
Friend-Negative	0.08	0.05, 0.16	0.06	0.00, 0.12	0.038	-0.01	-0.15, 0.13	0.00	-0.06, 0.05	0.867
Stranger-Positive	-0.05	-0.13, 0.04	-0.03	-0.09, 0.03	0.294	0.02	-0.14, 0.17	0.01	-0.05, 0.06	0.840
Stranger-Negative	-0.03	-0.12, 0.06	-0.02	-0.09, 0.04	0.475	0.04	-0.13, 0.20	0.01	-0.04, 0.07	0.659
Session	-0.87	-1.29, -0.45	-0.08	-0.12, -0.04	< .001	-0.73	-1.49, 0.03	-0.03	-0.07, 0.00	0.062
Global Ratings										

Intercept	9.24	6.45, 12.02	0.00	-0.16, 0.16	< .001	17.77	12.48, 23.06	0.00	-0.16, 0.15	<.001
Self-Positive	-0.52	-0.82, -0.22	-0.12	-0.19, -0.05	0.001	-0.73	-1.29, -0.17	-0.08	-0.14, -0.02	0.012
Self-Negative	0.13	-0.17, 0.44	0.03	-0.04, 0.10	0.398	0.03	-0.54, 0.60	0.00	-0.06, 0.07	0.925
Friend-Positive	-0.04	-0.35, 0.28	-0.01	-0.07, 0.06	0.806	0.08	-0.51, 0.67	0.01	-0.05, 0.07	0.796
Friend-Negative	0.23	-0.04, 0.51	0.05	-0.01, 0.12	0.094	0.34	-0.16, 0.85	0.04	-0.02, 0.09	0.186
Stranger-Positive	0.32	0.03, 0.62	0.07	0.01, 0.14	0.033	-0.17	-0.72, 0.39	-0.02	-0.08, 0.04	0.554
Stranger-Negative	-0.16	-0.45, 0.13	-0.04	-0.10, 0.03	0.272	0.20	-0.34, 0.74	0.02	-0.04, 0.08	0.465
Session	-0.91	-1.33, -0.49	-0.08	-0.12, -0.04	< .001	-0.78	-1.57, 0.00	-0.03	-0.07, 0.00	0.052

774 *b* = unstandardised regression coefficients, β = standardised regression coefficients

Figure 1

Cognitive Task Procedures

(a) Associative Learning Tasks: Example of an introduction, trial and feedback for each type for each type of task (self, reward, emotion). In the introduction of each task participants were instructed to associate specified randomly-assigned shape and stimuli pairings. In each trial participants were presented with a random combination of these shape-stimuli pairings and were asked to use the 'n' and 'm' keys to indicate whether these matched with the pairings they had previously learnt. In these examples, the 'm' key indicates a 'matching' responses and the 'n' key indicates a 'non-matching' response, however key assignment was randomised for each participant. Following each trial, feedback was given indicating if the participant was correct, incorrect, or too slow (> 1100 ms). Each of these examples demonstrate a 'matching' trial, where the presented shape-stimuli match with the pairings specified in the introduction. A 'matching' response would therefore be correct, in this example the 'm' key, whereas an 'non-matching' response would be incorrect, in this example the 'n' key.

(b) Go/No-Go Self-Esteem Association Task: Example of a trial and feedback for the Self-Positive condition. The conditions that words should be categorised according to (in this instance Me or Nice) were presented at the top of the screen throughout the block. In each trial a word was presented at the centre of the screen. Participants were asked to press the spacebar if the word belonged to a specified category (a 'go' response) or to refrain from pressing the spacebar if the word did not belong to the specified category (a 'no-go' response). Feedback (correct indicated by a green circle, or incorrect indicated by a red cross) was given for each response. In this example, a 'no-go' response would be considered a correct rejection and a 'go' response would be considered a false alarm, as the stimuli ('those') does not belong to the Me or Positive categories.

(c) Social Evaluation Learning Task: Example of a trial and feedback. Participants were asked to select the word that they felt reflected the computers' opinion of the person being learnt about (self, friend or stranger), and were given feedback on their response. The proportion of trials deemed correct upon selection of the positive word was manipulated to reflect learning of two different rules: positive 'like' 60-80%, negative 'dislike' 20-40%.

Figure 2

(a) Relationship between bias scores in the self, friend and stranger conditions in the social evaluation learning task with (i) PHQ-9 and (ii) BDI-II scores.

(b) Learning curves in the self condition in the social evaluation learning task based on cumulative accuracy with depression severity grouped according to (i) PHQ-9 clinical cut-offs and (ii) BDI-II clinical cut-offs.