Abnormal Evidence Accumulation Underlies the Positive Memory Deficit in Depression

Andrea M. Cataldo[1,2], Luke Scheuer[3], Arkadiy L. Maksimovskiy[1,2,4], Laura T. Germine[2,3,5], Daniel G. Dillon[1,2]

[1]Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, MA

[2]Department of Psychiatry, Harvard Medical School, Boston, MA

[3]Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA

[4]Brain Imaging Center, McLean Hospital, Belmont, MA

[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

Author Note

Contact: Andrea M. Cataldo, 225 de Marneffe Building, McLean Hospital, 115 Mill St., Belmont, MA, 02478, or by email at amcataldo@mclean.harvard.edu.

The data have been made publicly available at https://osf.io/w9asj/.

Disclosures: Dr. Dillon served as a consultant to Pfizer, Inc., for unrelated projects in 2016 and 2017. Laura Germine is on the Scientific Advisory Board of the nonprofit Sage Bionetworks, for which she receives a small honorarium.

This work was supported by grants awarded by the National Institute of Mental Health to Drs. Germine (R01 MH121617) and Dillon (R01 MH111676). Dr. Maksimovskiy was supported by a T32 training grant awarded by the National Institute on Drug Abuse to Dr. Scott Lukas (T32 DA015036).

Abstract

Healthy adults show better memory for low-arousing positive versus negative stimuli, but depression compromises this positive memory advantage. Existing studies are limited by small samples or analyses that provide limited insight into underlying mechanisms. Our study addresses these concerns by using a multi-staged analysis, including diffusion modeling, to identify precise psychological processes underlying the positive memory advantage and its disruption by depression in a large sample. A total of 1,358 participants completed the BDI-II (Beck, Steer, Ball, & Ranieri, 1996) and an emotional memory task. At encoding, participants judged whether positive and negative words were positive or self-descriptive. After a free recall test, participants viewed an equal mix of studied and unstudied words, and judged whether each one was "old" or "new"; if judged "old", they indicated the study source ("Positive?" or "Describes?" question). We replicate the positive memory advantage and its decrease in depression in recall, recognition, and source accuracy. The Hierarchical Drift Diffusion Model (HDDM; Wiecki et al, 2013) revealed that higher BDI scores are associated with more efficient evidence accumulation for negative words in the recognition and source memory tasks. By contrast, evidence accumulation for positive words is unaffected by BDI during the recognition task, but becomes less efficient with increased BDI during the source memory task. In conclusion, in a well-controlled design with a large sample, we find that depression reduces the positive memory advantage. HDDM analyses suggest that this reflects differential effects of depression on the speed of evidence accumulation during the retrieval of positive vs. negative memories.

Keywords: depression, emotion, memory, modeling, evidence accumulation

Abnormal Evidence Accumulation Underlies the Positive Memory Deficit in Depression

Major Depressive Disorder (MDD) is both a serious and common mental illness, with 7% of US adults reporting at least one episode in the past year (NIMH, 2017). In addition to its hallmark affective and physiological symptoms, MDD has long been associated with cognitive deficits such as impaired executive functioning (Snyder, 2013) and reduced sensitivity to rewards (Meehl, 1975; Pizzagalli, 2014; Proudfit, 2015; Treadway & Zald, 2011). Further, depression is associated with episodic memory deficits (MacQueen, Galway, Hay, Young, & Joffe, 2002; Rock, Roiser, Riedel, & Blackwell, 2014; Zakzanis, Leach, & Kaplan, 1998), with striking effects on memory for emotional material: Whereas healthy adults tend to show improved memory for low-arousing positive vs. negative stimuli, this advantage is reduced or even reversed in depression (Burt, Zembar, & Niederehe, 1995; Matt, Vazquez, & Campbell, 1992). This effect suggests that, in daily life, depressed adults likely retrieve proportionately fewer positive memories than do their non-depressed peers. The downstream consequences are substantial; for instance, broad memory deficits predict a longer course of illness (Hallford, Rusanov, Yeow, & Barry, 2021; Sumner, Griffith, & Mineka, 2010), and poorer specificity when recalling positive memories has been associated with increased vulnerability to depression among adolescents (Askelund, Schweizer, Goodyer, & van Harmelen, 2019). Encouragingly, the central role of emotional memory in depression is increasingly recognized, and therapies that aim to correct negative memory biases have yielded promising results in depressed patients (see Dalgleish & Werner-Seidler, 2014 for review).

Therapeutic interventions would likely be more effective if more were known about the upstream mechanisms that drive these memory deficits in the first place. Neuroimaging research has linked poor memory for positive material and enhanced memory for negative material to abnormal encoding responses in the dopaminergic midbrain (Dillon, Dobbins, & Pizzagalli, 2014) and amygdala (Hamilton & Gotlib, 2008), respectively, but these studies used relatively small samples and the specific cognitive processes affected by these neural abnormalities remain underspecified. In this context, computational models such as the Drift Diffusion Model (DDM; Ratcliff & McKoon, 2008) can provide insight into the specific psychological mechanisms that support the positive memory advantage in healthy adults, and that

are disrupted by depression. The DDM, depicted in Figure 1, assumes that simple binary decisions (such as judging whether or not a stimulus has been seen before) are the result of evidence accumulating over time from an initial starting point towards one of two boundaries that represent possible responses (such as "yes" or "no"). The starting point *z* ranges from 0-1, with values below and above 0.5 reflecting initial "starting" biases towards the lower ("no" in Figure 1) and upper ("yes" in Figure 1) response boundaries, respectively. The distance *a* between the two boundaries measures the amount of evidence needed to respond. Higher values of *a* correspond to greater distance between the boundaries and indicate greater response caution. Though evidence is assumed to accumulate noisily on a trial-by-trial basis (e.g., the orange line in Figure 1), the average rate of accumulation across trials (the "drift rate", e.g., the solid black line in Figure 1) is represented by the parameter *v*, with negative and positive values indicating that evidence generally accumulates towards the lower ("no" in Figure 1) and upper ("yes" in Figure 1) response boundaries, respectively. Regardless of sign, higher magnitude drift rates reflect more efficient evidence accumulation. Lastly, the time spent for all "non-decision" processes (e.g., encoding and motor response) is represented by the parameter t_0. By modeling the decision process dynamically, the DDM jointly estimates both choice proportions (i.e., the proportion of trials in which the upper vs. lower boundary was reached) and response times (i.e., the time taken to reach a boundary), thus providing a well-constrained model of decision-making as it unfolds over time.

The power of the DDM in uncovering subtle processing differences was demonstrated by White et al (2009) in a study of emotional memory in dysphoric college students. Though no significant behavioral effects of dysphoria or stimulus valence were found for memory accuracy or response times (RTs), the authors did find effects for drift rate. Specifically, whereas non-dysphoric students demonstrated faster drift rates for positive versus negative words, no such difference was found among dysphoric students, suggesting that the positive memory deficit in depression reflected slow evidence accumulation during the retrieval of positive versus negative memories. This demonstrates that computational models can uncover processing differences that may be difficult to detect with standard

behavioral measures such as percent correct or mean RT, as these are the net outcome of multiple component processes that the model can disentangle.

The findings of White et al (2009) thus demonstrate the utility of the DDM in clinical research, but several questions remain regarding the nature of the positive memory deficit in depression. First, it remains unclear whether the altered emotional memory bias seen in depressed adults mainly reflects altered processing of positive stimuli, negative stimuli, or both. Whereas some studies point to poorer memory for positive stimuli in depression (Auerbach, Stanton, Proudfit, & Pizzagalli, 2015; Burt et al., 1995), others suggest improved memory for negative stimuli (e.g., Hamilton & Gotlib, 2008), and still others find effects in both directions (Dillon et al., 2014; Dunbar & Lishman, 1984; Rouhani & Niv, 2019). Second, many studies of memory in depression rely on self-referential encoding tasks (Dainer-Best, Lee, Shumake, Yeager, & Beevers, 2018), typically finding that, relative to controls, depressed adults endorse more negative but fewer positive words as self-descriptive, and then go on to remember more negative but fewer positive words. This approach is robust but confounds valence and self-referential processing: One cannot tell if better memory for negative words in depressed individuals is due to exaggerated emotional responses or instead simply reflects greater elaboration at encoding, since judging material in relation to the self is a classic "deep processing" manipulation that reliably enhances memory (Rogers, Kuiper, & Kirker, 1977). To adjudicate between these possibilities, the self-referential task could be paired with a second encoding task that uses similarly valenced stimuli but that does not require self-referential processing. Third, many prior studies in this area, including our own, used relatively small samples and thus had limited power.

The goal of the present work is to address these limitations by applying the DDM to a large-scale online study of emotional memory in order to isolate specific psychological mechanisms that drive the positive memory advantage and its reduction in depression. Specifically, 1,358 participants studied a list of normatively positive and negative words by providing either a self-reference ("Does this word describe you?") or valence ("Is this word positive?") judgment for each word. Consistent with prior work, we expect higher BDI scores to be associated with a tendency to endorse more negative vs. positive in the

self-reference task. Due to the robustness of the positive memory deficit, we predict that increased BDI scores will be associated with worse memory for positive vs. negative words from both tasks, consistent with a task-general response to emotional material. Moreover, given the diversity of previous findings (e.g., Dillon et al., 2014; Dunbar & Lishman, 1984; Hamilton & Gotlib, 2008; Rouhani & Niv, 2019), we predict that such effects will most likely be driven by both poorer memory for positive material and enhanced memory for negative material, but potentially via distinct mechanisms. In an effort to dissociate these effects, we probed memory in three ways. We first examined free recall. We expected strong effects of depression on recall based on prior work (Matt et al., 1992), but one concern is that recall is particularly effortful and dependent on executive functioning, which is also affected by depression (Snyder, 2013). To address this potential confound, and following White et al (2009), we next tested recognition memory to reduce demands on executive functioning and collect data better suited for modeling. Lastly, we tested source memory by asking participants to identify the encoding task for words reported as "old". Probing source memory is useful because the task offers better experimental control than free recall but may be more sensitive to the negative effect of depression on hippocampal-dependent processes than recognition memory (MacQueen et al., 2003, 2002; Ramponi, Barnard, & Nimmo-Smith, 2004). Here, we make the additional prediction that the well-established tendency of depressed adults to endorse more negative than positive words as self-referential may carry over into source judgments. Specifically, as BDI increases, participants may be more likely to attribute negative words to the self-reference task.

In the modeling stage, recognition and source accuracy data were fit with the Bayesian Hierarchical Drift Diffusion Model (HDDM; Wiecki et al., 2013). Here, we expect the results of White et al (2009) to replicate for recognition and extend to source memory, with higher BDI scores leading to less efficient evidence accumulation during the retrieval of positive memories and more efficient evidence accumulation during the retrieval of negative memories. Lastly, because it is a more widely-used framework, we further fit the data to an equal variance Signal Detection model (SDT; Macmillan & Creelman, 2005) for generalizability. Because drift rates are traditionally associated with memory

strength (Ratcliff & McKoon, 2008), we expect the SDT results for both recognition and source accuracy to demonstrate decreased discriminability for positive vs. negative memories with increased BDI scores.

Method

Participants

The study was reviewed by the Harvard University Committee on the Use of Human Subjects, protocol 15795 “Web-based studies of individual differences in cognition". A total of 1,945 participants were recruited through the online platform TestMyBrain.org (Germine et al., 2012). Prior to analyses, we excluded 450 participants for being younger than 18 years old and 115 for not completing all self-report measures. We removed an additional 22 subjects for negative recognition *d'* scores, which reflect below chance performance and strongly suggest a task misunderstanding. We did not, however, remove subjects for negative source *d'* scores, which could easily emerge due to low trial counts in participants who performed poorly at recognition. The final sample of 1,358 participants was diverse: 56% were female and 54% were of European descent, with a mean (SD) age of 32.14 (13.6) years. Participants were not compensated but received feedback on their performance upon completion of the study.

Self-Report Measures

Participants reported their age, gender, and ethnicity and completed the Beck Depression Inventory II (BDI; Beck, Steer, Ball, & Ranieri, 1996), a 21-item scale that assesses depressive symptoms for the prior two weeks. Age and sex data were collected prior to the task; all other measures were completed after.

Memory Task

Stimuli

Participants were randomly assigned to one of two 50-item word lists at encoding, A or B, each drawn from lists of 100 negative and 100 positive English words. Normative valence (1 = negative, 9 =

positive) and arousal (1 = calm, 9 = excited) ratings (Warriner, Kuperman, & Brysbaert, 2013) were available for each word, along with frequency (Brysbaert & New, 2009) and concreteness (Brysbaert, Warriner, & Kuperman, 2014) data. Participants viewed all words from both lists at recognition. Within each list, the negative and positive words differed significantly on valence (A: t = -41.49, p < .001; B: t = -42.23, p < 0.001) but not on other properties (arousal, number of letters, frequency, concreteness, imageability; all ps < .05). Given its particular relevance to studies of emotional memory, we note that arousal ratings were consistently moderate across lists and valence conditions (A_{pos}: median=4.67, min=3.14, max=6.65; A_{neg}: median=4.69, min=1.67, max=6.60; B_{pos}: median=4.57, min=2.38, max=6.65; B_{neg}: median=4.72, min=2.70, max=6.60).

Both lists were submitted to Latent Semantic Analysis (LSA; Landauer, 2006; see also Dillon, Cooper, Grent-'t-Jong, Woldorff, & LaBar, 2006). LSA is a machine learning tool that analyzes the semantic similarity of text; it indicated that iter-item associativity is equal across the 100 negative (0.11±0.10) and 100 positive (0.11±0.10) words, p = 0.13. This is important, because in many studies stimulus differences in valence are confounded with differences in semantic cohesion, which has independent effects on memory (Talmi & Moscovitch, 2004). Lastly, there were no differences on any property (all ps < .05) for negative or positive words assigned to list A vs. B.

Procedure

The memory task included encoding, recall, and recognition stages (Figure 2). On each of 50 encoding trials (Figure 2a), participants viewed a normatively positive or negative word and made one of two yes/no judgments: "Describes you?" (self-reference task) or "Positive word?" (valence task). All participants completed an equal number of self-reference and valence tasks, which were presented in a fixed random order. Words remained on screen until the participant responded or 10 seconds had elapsed. If no response was made after 10 seconds, participants saw a screen that read "Time's up! Next time try to respond more quickly."

After encoding, participants completed free recall and recognition tests. In the free recall test (Figure 2b), participants typed as many words from encoding as they could remember—there was no time limit, and participants could stop whenever they wished. The recognition test (Figure 2c) included all 50 "old" words from encoding plus 50 "new" positive and negative words drawn from the list that was not shown at encoding. Old and new words were presented in a fixed random order. The task was to indicate whether each word was from encoding (i.e., was old) or not (i.e., was new). If participants indicated that a word was new, they proceeded to the next trial. If they judged a word old, they were asked to report whether it was encoded in the self-reference or valence task. This final question tests source memory, as participants had to retrieve the context—the encoding task—in which each word was previously encountered.

Analyses

Behavior

Encoding behavior is measured by the proportion of "yes" responses. Recall performance is measured by the number of correctly recalled words. Recognition performance is measured by the proportion of "old" responses to words studied in each encoding task. Lastly, source accuracy is measured by the proportion of "self-reference" source attributions for correctly recognized words. Because the HDDM cannot handle especially long RTs, trials with RT greater than 3 SD above the subject's mean were excluded from all recognition and source analyses. This resulted in the exclusion of 1.6% of recognition trials and 2.0% of source trials.

HDDM

Diffusion modeling of recognition and source data was conducted with the Python package hddm (Wiecki et al., 2013). The HDDM is a Bayesian hierarchical model in which subject-level parameters are drawn from group-level distributions. We assumed uninformed priors in each analysis. In both the recognition and source models, threshold a, starting point bias z, and non-decision time t_0 were all fixed

within subjects[1]. In the recognition model, drift rate *v* was allowed to vary by valence and study status. Positive and negative drift rates indicated evidence accumulation towards "old" and "new" responses, respectively. In the source accuracy model, drift rate was allowed to vary by valence and the source (encoding) task. Here, positive and negative values indicated evidence accumulation towards "self-reference" and "valence" responses, respectively. Though the recognition and source memory behavioral analyses focus on accuracy for studied items, both studied and unstudied items were included in model fits.

Each model run included 2000 burn-in steps and 8000 saved steps. All models converged, and all were validated with posterior predictive checks of choice and RTs. Inferences are made by calculating the 95% highest density interval (HDI) around the mean of the posterior estimate of a given parameter. We consider a meaningful effect to be indicated by an HDI that excludes zero or, for between-group comparisons, a pair of intervals that do not overlap. We caution, however, that this is merely a guide to ease exposition; with Bayesian estimation, intervals that do not meet this threshold are not necessarily qualitatively distinct from those that do (see Kruschke, 2014).

SDT

To facilitate comparisons with broader research, we also describe the recognition and source accuracy data with parameters from an equal variance signal detection model (Macmillan & Creelman, 2005). Prior to analyses, the log-linear correction was applied to account for response proportions of 0 and 1 (Snodgrass & Corwin, 1988). In the recognition model, memory strength was measured by *d'*, calculated as z(H)-z(F), and bias was measured by *c*, calculated as -.5[z(H)+z(F)], where H and F refer to hit and false alarm rates, respectively. For source memory, we restricted the analysis to those words correctly recognized as old. Here, discriminability between words from the self-reference and valence

[1] Alternative models in which (1) threshold and (2) threshold and starting point bias were allowed to vary by valence were also tried, however these models yielded worse fits to the data. Given this, and given that it is not standard practice to allow these parameters to vary within-subjects, we proceed with the simpler model.

task was measured by *d'*, calculated as z(S)-z(V), and bias was measured by *c*, calculated as -.5[z(S)+z(V)], where S and V refer to the proportion of "self-reference" responses for words actually from the self-reference and valence tasks, respectively. Again, although the recognition and source memory behavioral analyses focus on accuracy for studied items, both studied and unstudied items were included in model fits.

Regression Models

To examine the impact of BDI scores on performance, we analyzed the behavioral data and the HDDM[2] and SDT model parameters with Bayesian regression models conducted with the R package rstanarm. Default priors and sampling parameters were used unless otherwise indicated (Goodrich, Gabry, Ali, & Brilleman, 2020). All models converged and were validated with posterior predictive checks. Inferences are made by calculating the 95% HDI around the mean of the posterior estimate of a given regression coefficient, as above. When included as a predictor, BDI score was always treated continuously rather than categorically (i.e., groups based on predetermined cutoffs).

Results

In each section below, we first show and describe the data, then perform the analyses. Because the analyses are extensive, only the most relevant results are presented in the text; additional details are reported in tables or the Supplemental Material as indicated. The data are provided at https://osf.io/w9asj/.

BDI Scores

BDI scores covered nearly the entire range (see Figure S1 in the Supplemental Material), but clustered at the lower end with a median of 13. Mean BDI scores were higher for females (M=15.46,

[2] Though inferences regarding categorical predictors can be drawn from the posteriors of HDDM fits alone, BDI scores are instead analyzed via subsequent regressions on subject-level parameters. An alternative approach incorporating BDI score into the model fits is possible the HDDMRegressor; however, given the size of our data set, this would incur a high computational cost with little expected quantitative benefit.

SD=11.74) than males (M=14.12, SD=10.89) and correlated negatively with age (r=-0.18). These observations were supported by a Bayesian negative binomial regression model: $BDI \sim age + gender$. The 95% HDIs for the estimated coefficients of male gender (-0.22:-0.03) and age (-0.02:-0.01) were entirely below zero, so age and gender were included as co-variates in all subsequent analyses. For brevity, we do not consider interactions of age and gender with the effects discussed below; those could be pursued in future work.

Encoding

As shown in Figure 3a (left panel), on average participants responded "yes" to more positive than negative words in the self-reference task. They also responded "yes" (indicating that the word was positive) to more positive than negative words in the valence task (right panel), and to a greater degree than in the self-reference task; this is unsurprising given that, in the valence task, participants were asked to indicate whether words were positive. As expected, Figure 3b shows that as BDI scores increased, participants endorsed more negative but fewer positive words as self-descriptive (left panel). Unexpectedly, participants with higher BDI scores were also more likely to disagree with normative valence ratings; that is, they were less likely to judge normatively positive words as positive but more likely to judge normatively negative words as positive (right panel).

As illustrated in Table 1, these patterns were all statistically supported by a Bayesian mixed effects logistic regression model: $p("yes") \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$. Most critically, the model estimates that as BDI score increases, participants are more likely to respond "yes" to negative words but are less likely to repond "yes" to positive words shown in the self-reference task, with similar but reduced effects in the valence task.

Recall

As shown in Figure 4a, participants recalled more positive than negative words as well as more words from the self-reference task versus the valence task. There does not appear to be an interaction

between valence and task: The recall advantage for positive vs. negative words was about the same for words recalled from both tasks. As shown in Figure 4b, higher BDI scores were associated with poorer recall of positive words but slightly better recall of negative words, with this effect appearing for words from both the self-reference (left panel) and valence (right panel) tasks.

As illustrated in Table 2, these observations were confirmed by a Bayesian mixed effects Poisson regression model: $\#\,recalled \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject)$. Notably, the model estimates that as BDI score increases, the number of correctly recalled positive words decreases but the number of correctly recalled negative words does not reliably change, with no additional interactions of BDI with encoding task for either positive or negative words.

Recognition

Recognition Accuracy for Old Words

Because the effects of encoding task on recognition accuracy cannot be easily incorporated into the HDDM and SDT analyses, we begin by briefly describing the hit rate data separately for words from each task before drawing inferences from the models. A full analysis is provided in the Supplemental Material. As shown in Figure S2a, participants correctly recognized more old positive versus old negative words, and more words from the self-reference versus valence task. These effects do not appear to interact: The recognition advantage for positive vs. negative words appears similar for words from both tasks. As shown in Figure S2b, higher BDI scores were associated with better recognition of negative words from both tasks, with perhaps slightly poorer recognition of positive words from the self-reference task. As described in the Supplemental Material, all observations were confirmed by a Bayesian mixed effects logistic regression model: $p(\text{"}old\text{"}) \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$.

HDDM

Figure 5a presents the estimated mean drift rates by condition, with error bars that depict 95% HDIs from the Bayesian model fits. Recall that positive and negative values indicate evidence accumulation towards "old" and "new" responses, respectively. As shown in the left panel, evidence accumulated more efficiently towards an "old" response for old positive versus old negative words. In contrast, the right panel shows that evidence accumulated slightly less efficiently towards a "new" response for new positive versus new negative words.

Figure 5b shows that, as BDI score increases, drift rate increases for negative words—especially old negative words—whereas drift rate for positive words is weakly decreased. That is, as depressive symptoms increase, participants accumulate more evidence towards an "old" response for negative words and a "new" response for positive words, whether or not words were actually old or new (i.e., regardless of study status). The observed effects of BDI on drift rates were supported by a Bayesian mixed effects regression model: $v \sim age + gender + valence \cdot study \cdot BDI + (1 \mid subject)$ (see Table 3). We note that this systematic pattern suggests a potential role of bias not accounted for in the model. More detailed consideration of this point is provided in the Discussion.

Bayesian regression models were also conducted to evaluate the effect of BDI score on threshold *a*, starting point bias *z*, and non-decision time t_0: $parameter \sim age + gender + BDI$. As BDI scores increased, participants exhibited weakly decreased thresholds (95% HDI: -0.003:0.000), a weakly increased bias towards "old" responses (95% HDI: 0.000:0.001), and weakly increased non-decision times (95% HDI: 0.000:0.001).

SDT

Figure 6a presents the SDT parameters for memory strength (*d'*, left column) and bias (*c*, right column). As can be seen, participants exhibited greater memory strength and a weaker tendency to respond "new" to positive than negative words, with a conservative response bias across valences. Figure 6b reveals that, unexpectedly, memory strength for positive and negative words was unaffected by BDI score, as was response bias for positive words. By contrast, *c* decreased with BDI score for negative

words, indicating greater willingness to endorse both old and new negative words as "old" as the severity of depression increases. This selective effect of BDI on response bias for negative words likely contributes to the positive association observed between BDI scores and hit rates for negative words, but not positive words (see Figure S2), and further supports a role of bias not captured by the HDDM.

Each parameter was analyzed with a Bayesian mixed effects regression model: $parameter \sim age + gender + valence \cdot BDI + (1 \mid subject)$ (see Table 4). Supporting the above impressions, the model does not estimate a relationship between BDI score and *d'* for either positive or negative words, or between BDI score and *c* for positive words[3]; however there was a modest negative association between BDI scores and *c* for negative words.

Re-analysis Excluding Recalled Items

It is possible that by testing free recall before recognition memory, item effects were introduced for recalled words. We therefore repeated all the above analyses—analysis of recognition hit rates, HDDM, and SDT—with recalled words excluded; see the Supplemental Materials for details. Results were essentially unchanged, indicating that the recognition findings were not strongly influenced by free recall performance.

Source Accuracy

Source accuracy for old words

As for the recognition analyses, because the effects of encoding task cannot be easily incorporated into the HDDM and SDT analyses of source accuracy, we begin by briefly describing the response proportions from each task separately before drawing inferences from the models. A full analysis is provided in the Supplemental Material. As predicted, Figure S6a reveals greater source

[3] Though the 95% HDI for the interaction of positive valence and BDI scores on *c* (positive:BDI) is entirely above zero, this merely serves to counteract the negative effect of BDI on *c* for negative words, thereby estimating the same null effect of BDI on positive words observed in Figure 6b.

accuracy for positive than negative words from the self-reference task. Surprisingly, however, source accuracy was higher for negative than positive words from the valence task. Figure S6b (left panel) shows that as BDI scores increased, source accuracy improved for negative words but decreased for positive words from the self-reference task. However, the opposite pattern was found for words from the valence task (right panel): here, increased BDI scores were associated with lower source accuracy for negative words but higher source accuracy for positive words. As described in the Supplemental Material, all observations were confirmed by a Bayesian mixed effects logistic regression model: $p(correct) \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$.

HDDM

Figure 7a presents the estimated mean drift rates by condition, with error bars that depict 95% HDIs from the Bayesian model fits. Recall that positive and negative values indicated evidence accumulation towards "self-reference" and "valence" responses, respectively. Appropriately, participants had positive drift rates for words from the self-reference task (left panel) and negative drift rates for words from the valence task (right panel). For words from the self-reference task, evidence accumulated more efficiently towards a "self-reference" response for positive words relative to negative words. For words from the valence task, evidence accumulated less efficiently towards a "valence" response for positive vs. negative words.

As shown in Figure 7b, as BDI score increased participants tended to accumulate evidence in favor of a "self-reference" response for negative words and a "valence" response for positive words, and this pattern did not vary strongly depending on which task the words were actually from (i.e., both panels show the same basic pattern). The observed effects of BDI on drift rates were supported by a Bayesian mixed effects regression model: $v \sim age + gender + valence \cdot source \cdot BDI + (1 \mid subject)$ (see Table 3). As for the recognition results, this pattern again suggests a potential role of bias not currently accounted for in the model, which we detail in the Discussion section.

Bayesian regression models were also conducted to evaluate the effect of BDI score on threshold *a*, starting point bias *z*, and non-decision time t_0: $parameter \sim age + gender + BDI$. As BDI score increased, participants exhibited decreased thresholds (95% HDI: -0.005:-0.001), no discernable change in starting point bias (95% HDI: 0.000:0.000), and weakly increased non-decision times (95% HDI: 0.000:0.002).

SDT

Figure 8a presents source discriminability between the self-reference and valence tasks (*d'*, left panel) as well as bias towards a "valence" source attribution (*c*, right panel). As shown on the left, participants exhibited greater source discriminability for positive versus negative words. The right panel reveals that response bias was liberal for positive words but conservative for negative words, indicating that participants tended to attribute positive words to the self-reference task but negative words to the valence task. Figure 8b presents the SDT parameters (*d'*, left column; *c*, right column) as a function of BDI score. Higher BDI scores were associated with weakly decreased source discriminability for positive words, with no effect on negative words. Furthermore, as BDI score increases, response bias decreases for negative words but increases for positive words; in the context of source judgments, this indicates that participants with low BDI scores tended to attribute negative words to the valence task and positive words to the self-reference task, but participants with high BDI scores tended to attribute negative words to the self-reference task and positive words to the valence task. This crossover effect of BDI and valence on response bias likely contributes to the startlingly symmetric effects of BDI on source accuracy for positive vs. negative items (see Figure S6 in the Supplemental Material).

The above impressions were supported by a Bayesian mixed effects regression model for each parameter: $parameter \sim age + gender + valence \cdot BDI + (1 \mid subject)$ (see Table 4). As seen in Figure 8b, higher BDI scores are weakly associated with lower values of *d'* for positive words but not negative words. Much stronger associations are indicated between BDI score and *c*, with higher BDI

scores associated with an increased tendency to attribute positive words to the valence task but negative words to the self-reference task.

Source Attributions for Recognition False Alarms

The analyses of recognition and source memory consistently indicate a role for bias. To further explore this possibility, recall that at recognition, participants were asked to report the encoding task for each word that they endorsed as "old". Though source attributions for recognition hits can be driven by either memory strength or response bias, such attributions for recognition false alarms can only reflect bias. Thus, an effect of BDI on source attributions for recognition false alarms would suggest a key role for bias. As detailed in the Supplemental Material, we find such an effect. Specifically, low BDI scores are associated with a greater probability of attributing positive versus negative false alarms to the self-reference task, whereas the opposite pattern is found for high BDI scores (see Figure S7 in the Supplemental Material). HDDM fits suggest that this is driven by faster accumulation towards a "self-reference" response for negative, but not positive, items as BDI score increases (see Figure S8 in the Supplemental Material). We consider the possibility that this reflects a systematic sampling bias in the Discussion.

Discussion

Previous research has found that while healthy adults tend to show better memory for low-arousing positive vs. negative stimuli, this advantage is reduced in depressed adults (Burt et al., 1995; Matt et al., 1992). Although the clinical importance of memory deficits in depression is increasingly recognized (Askelund et al., 2019; Dalgleish & Werner-Seidler, 2014; Sumner et al., 2010), the underlying mechanisms driving these deficits remain unclear. Computational modeling offers a powerful method of isolating the specific psychological mechanisms underlying memory, and previous applications of the Drift Diffusion Model (DDM; Ratcliff & McKoon, 2008) have suggested that the positive memory advantage—and its compromise in depression—is the result of differential evidence accumulation for

positive versus negative material (White et al., 2009). In the present study, we replicated the positive memory advantage and its reduction in depression in a large online sample. Specifically, although participants overall demonstrated better recall, recognition, and source accuracy for positive versus negative stimuli, the positive memory advantage decreased with increased depressive symptoms reported on the BDI. Fitting the recognition and source accuracy data with the HDDM revealed more efficient evidence accumulation for positive versus negative stimuli, but the size of this effect decreased with increasing BDI scores due to increased drift rate for negative words (in the recognition and source tests) and reduced drift rate for positive words (in the source test). Overall, this indicates that the positive memory advantage and its reduction in depression stems largely from effects on the evidence accumulation process.

Our multi-staged behavioral and modeling analyses yielded several additional insights. First, at encoding, higher BDI scores were associated with a decreased probability of endorsing positive words and a correspondingly increased probability of endorsing negative words as self-referential. Surprisingly, a weaker but similar effect was seen for the valence task, where higher BDI scores were associated with a decreased probability of judging positive words as positive, but an increased probability of judging negative words as positive. Two considerations are worth noting here. First, because the BDI is related to greater disagreement with normative valence, an interesting question is whether the recognition results differ when analyzed according to subjective valence. This is an important consideration, but because the effect of BDI on valence judgments is relatively small, and because such an analysis would require quartering the number of trials (to include only studied words from the valence task; old words from the self-referential task and all new words would be excluded), the current paradigm is underpowered for this line of questioning. We therefore leave this issue as a target for future work. Second, interpretation of these results is limited by the lack of attention checks in the current study, which leaves open the possibility that disagreement with normative valence reflects reduced engagement from participants with greater depressive symptoms (see Zorowitz, Niv, & Bennett, in prep, for detailed discussion). A holistic view of the data, however, suggests that the impact of BDI on encoding behavior may truly be task-

general, because this pattern extends to the recall and recognition stages. Specifically, the effect of BDI on recall and recognition memory was similar for words from both encoding tasks, despite better overall memory for words from the self-reference task. By pairing the self-referential encoding task with a valence judgement task, the current study disentangles two separate effects that are otherwise confounded: a levels-of-processing effect that globally enhances encoding and subsequent memory for words from the self-referential encoding task (Craik & Tulving, 1975; Rogers et al., 1977), plus a depression effect that depends heavily on stimulus valence but that can be seen whether or not self-referential processing is manipulated.

A second insight is that although the positive memory advantage decreased with higher BDI score in each memory test, whether this occurred due to poorer memory for positive stimuli versus improved memory for negative stimuli varied. Specifically, whereas poorer memory for positive stimuli appears to have driven the effect of BDI at recall, performance at recognition appears to be driven by improved evidence accumulation for negative stimuli, and performance at source accuracy appears to result from both. This variability across tests may help explain similarly diverse findings in previous literature. Specifically, consistent with the present findings, depression has been associated with stronger effects on memory for positive versus negative stimuli for recall (Auerbach et al., 2015; Burt et al., 1995), and with effects in both directions for source memory (Dillon et al., 2014). The recognition literature is less consistent; although some work has found specific effects on memory for negative material (Hamilton & Gotlib, 2008), other have found effects in both directions (Dunbar & Lishman, 1984; Rouhani & Niv, 2019). This pattern of results indicates an interesting target for future work on potential neural correlates. Specifically, depression is thought to exert many of its negative effects by impairing hippocampal function (MacQueen & Frodl, 2011). Free recall is strongly hippocampal-dependent, whereas recognition draws more heavily on familiarity, with source memory falling between the two. Thus, these results raise the interesting possibility that asymmetric memory for positive vs. negative material may be driven by differential connectivity between regions traditionally involved in the processing of positive and negative

material (e.g., ventral striatum and amygdala, respectively) and the hippocampus, which is itself differentially activated across tasks. Targeting such mechanisms remains a goal for future work.

From a behavioral perspective, a leading candidate for this pattern of results is a valence-dependent retrieval bias. Indeed, in the recognition test higher BDI scores were clearly associated with an increased bias to endorse negative words as old. This account would seem to fail with respect to the free recall data, where—if a negative bias were critical—one might expect increased BDI scores to be associated with enhanced recall of negative words, and perhaps a large number of negative intrusions, rather than the decrease in positive recall that was observed. One interpretation is that such a bias may be moderated by task effort, which is generally higher in recall vs. recognition. But a limitation of the study is also pertinent here: Participants could quit the recall task whenever they wanted, and it is well-known that people tend to stop recalling items long before their memory is exhausted (Wixted & Rohrer, 1994). Given that low motivation is a common aspect of depressive illness (see Grahek et al., 2019 for review), it is plausible that participants with high BDI scores quit the recall task too quickly, before an increase in the number of negative words recalled could be seen. A related possibility is that in adults with high BDI scores, negative words come to mind relatively easily but dredging up positive words is difficult, such that these individuals recalled a few negative words but quit the test before recalling a similar number of positive words; this would give rise to the observed pattern of recall findings.

A particularly fascinating aspect of these data concerns performance on the source memory test; here, higher BDI scores were associated with a bias to assign negative words to the self-reference source while assigning positive words to the valence source. In contrast to the recall and recognition results, which did not depend heavily on self-referential processing, the tendency of more depressed adults to misremember negative words as having been encoded in relation to the self is striking and speaks to the overly negative self-concept in depressive illness (Strauman, 2002, 2017). Moreover, the near-perfect symmetry between the two valence conditions—in which the effect on self-reference attributions for negative words is mirrored by the effect on valence attributions for positive words—suggests a remarkably complementary relationship that is not inherent to the model structure itself. Though it is

possible that bias plays a role in producing these results, the exact nature of this relationship is an open question for future work, as we discuss below.

A third insight, consistent with results from White et al (2009), is that the modeling indicates that the positive memory advantage and its reduction in depression reflect differences in the speed of evidence accumulation for positive versus negative material. We believe this is a valuable foundation for future research seeking to characterize memory deficits in depression because evidence accumulation is a more narrowly defined construct than response bias or discriminability and its neural correlates are at least partially understood (Gold & Shadlen, 2007; Shadlen & Shohamy, 2016). However, our concurrent analyses of behavioral and signal detection measures suggest that future research on this topic ought to carefully consider the nature of the effect on evidence accumulation. Again, both the recognition and source accuracy SDT results indicate strong effects of BDI score on bias, with only subtle effects on discriminability. This would seem slightly at odds with the HDDM results, which pointed to effects on evidence accumulation that are typically associated with differences in discriminability, not bias (Ratcliff & McKoon, 2008).

One possible exaplanation for these apparent discrepancies is that the effect on evidence accumulation observed here in fact reflects biased *sampling*, measured by drift criterion (Ratcliff, 1978; Scimeca, Katzman, & Badre, 2016; Starns, Ratcliff, & White, 2012). Whereas drift rates correspond to evidence strength, drift criterion determines the probability that each bit of accumulated evidence should count in favor of one response boundary or the other, much like the bias parameter *c* in SDT. This process is in fact directly analogous to imposing a basic signal detection model at each step in the accumulation process. That is, at each step, a piece of evidence is sampled. If it exceeds the criterion, it is judged in favor of an “old” response and evidence accumulates upward, otherwise it is judged in favor of a “new” response and evidence accumulates downward. Thus, a drift criterion that is low relative to the old and new drift rates will bias sampled evidence towards an “old” response. It is plausible that, as BDI scores increase, the drift criterion for responding “old” might decrease so as to be lower for negative but not positive items, reflecting a disturbance in how adults with depression act upon negative memories.

Alternatively, the drift criterion may be unaffected while drift rates increase for both old and new negative items, reflecting a false sense of familiarity for all negative material in depression (e.g., Bowen, Kark, & Kensinger, 2018).

A drift criterion that is lower relative to both old and new drift rates for negative (but not positive) words would be consistent with our SDT results, as we found a decrease in *c* with increased BDI score for negative material. Unfortunately, because the HDDM can only capture the relative value of drift rate vs. drift criterion, accounts of biased sampling vs. false familiarity cannot be disentangled by model fits alone, and so it is not possible to decide between these two possibilities in the current dataset. Instead, studies employing experimental manipulations designed to target each process (e.g., Starns, Dubé, & Frelinger, 2018; Starns et al., 2012) are needed, and this may be profitable because recent work indicates that striatal prediction errors—a topic of perennial interest to depression researchers—affect the drift criterion (Scimeca et al., 2016). Further, though the models are accounting for behavior at retrieval, they remain agnostic as to whether increased memory strength or biased responding are ultimately driven by differences in processing at encoding or retrieval. Neuroimaging research could thus target this important issue, and could provide additional insight into the neural basis of the effects observed here.

Lastly, this work takes a step towards increased reproducibility of memory research in depression. Online data collection yielded a substantial sample of 1,358 subjects who are racially diverse and span a wide age range. Further, the Bayesian approach to both our behavioral and modeling analyses shifts the focus of the results from statistical significance to effect size estimation. Together, these components provide a solid foundation for future work to consider when generating novel predictions. Indeed, the reported task and valence effects were initially discovered in two additional samples, who completed the task but not the BDI (see Supplemental Material).

Our multi-stage analyses and Bayesian approach offer a well-informed starting point for future work, with results that indicate several avenues for targeted studies. There are also several limitations to the current work for future studies to address. First, the effect of recall performance on subsequent recognition and source accuracy behavior is unclear. Although no feedback was given at recall, it is fair to

assume that correctly recalled words are also more likely to be (1) correctly recognized and (2) attributed to the correct source. Encouragingly, a re-analysis of the current recognition data excluding recalled words found that the main results were essentially unchanged (see Supplemental Material). Nevertheless, future research may wish to more closely the examine effects of recall on subsequent tests of recognition and source memory in depressed adults. Second, collecting data online allowed us to reach a large and diverse sample of participants, but it prohibited clinical testing for depression, and a desire to not overburden participants led us to collect the BDI as our only clinical measure. Consequently, while we can make claims about depressive severity, we cannot rule out possible contributions from conditions that are typically comorbid with depression, such as generalized anxiety, and we cannot assume that these results will generalize to adults diagnosed with Major Depressive Disorder, although we expect that that would be the case. Follow-up studies with in-person data collection and diagnostic interviews are needed to address these issues. Third, the present study relied exclusively on behavioral data. By combinining behavioral data with SDT and HDDM analyses, we were able to begin characterizing the psychological processes responsible for altered emotional memory in depressed adults. However, the power of this approach would be greatly amplified by future work linking model parameters with EEG or fMRI data, as this will let us tie evidence accumulation (and the other parameters) to their neural underpinnings. This kind of model-informed neuroscientific analysis will ultimately give us a detailed understanding of the impact of depressive illness on the encoding and retrieval of emotional memories.

Context

Our group is interested in the neurocognitive mechanisms underlying psychiatric disorders. Members of the Motivated Learning & Memory Lab—Drs. Cataldo, Maksimovskiy, and Dillon—maintain a focus on identifying the subprocesses of memory and decision-making that are compromised in depression (e.g., Dillon, 2015; Dillon et al., 2014; Lawlor et al., 2020). Members of the Laboratory for Brain and Cognitive Health Technology—Mr. Scheuer and Dr. Germine—focus on individual differences in cognition and mental health, and developing technology to study cognition and behavior in large and

diverse online samples (see TestMyBrain.org). This collaborative study was motivated by previous research demonstrating that while healthy adults tend to show better memory for low-arousing positive vs. negative stimuli, this advantage is reduced in depressed adults (Burt et al., 1995; Matt et al., 1992). Although the clinical importance of memory deficits in depression is well established (Askelund et al., 2019; Dalgleish & Werner-Seidler, 2014; Sumner et al., 2010), our group is interested in isolating the upstream mechanisms driving these deficits. Thus, in the present study, we applied a computational modeling approach to data from a large and diverse online sample to provide a detailed account of the mechanisms driving the positive memory reduction in depression.

References

Askelund, A. D., Schweizer, S., Goodyer, I. M., & van Harmelen, A. L. (2019, March 1). Positive memory specificity is associated with reduced vulnerability to depression. *Nature Human Behaviour*, Vol. 3, pp. 265–273. https://doi.org/10.1038/s41562-018-0504-3

Auerbach, R. P., Stanton, C. H., Proudfit, G. H., & Pizzagalli, D. A. (2015). Self-referential processing in depressed adolescents: A high-density event-related potential study. *Journal of Abnormal Psychology*, *124*(2), 233–245. https://doi.org/10.1037/abn0000023

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. *Journal of Personality Assessment*, *67*(3), 588–597. https://doi.org/10.1207/s15327752jpa6703_13

Bowen, H. J., Kark, S. M., & Kensinger, E. A. (2018). NEVER forget: negative emotional valence enhances recapitulation. *Psychonomic Bulletin & Review*, *25*(3), 870–891. https://doi.org/10.3758/s13423-017-1313-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Burt, D. B., Zembar, M. J., & Niederehe, G. (1995). Depression and memory impairment: a meta-analysis of the association, its pattern, and specificity. *Psychological Bulletin*, *117*(2), 285–305. https://doi.org/10.1037/0033-2909.117.2.285

Cataldo, A. M., Dillon, D. G., Maksimovskiy, A., & Germine, L. (2022, February 23). Abnormal evidence accumulation underlies the positive memory deficit in depression. https://doi.org/10.17605/OSF.IO/W9ASJ

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, Vol. 104, pp. 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Dainer-Best, J., Lee, H. Y., Shumake, J. D., Yeager, D. S., & Beevers, C. G. (2018). Determining optimal parameters of the self-referent encoding task: A large-scale examination of self-referent cognition and depression. *Psychological Assessment*, *30*(11), 1527–1540. https://doi.org/10.1037/pas0000602

Dalgleish, T., & Werner-Seidler, A. (2014, November 1). Disruptions in autobiographical memory processing in depression and the emergence of memory therapeutics. *Trends in Cognitive Sciences*, Vol. 18, pp. 596–604. https://doi.org/10.1016/j.tics.2014.06.010

Dillon, D. G. (2015). The neuroscience of positive memory deficits in depression. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01295

Dillon, D. G., Cooper, J. J., Grent-'t-Jong, T., Woldorff, M. G., & LaBar, K. S. (2006). Dissociation of event-related potentials indexing arousal and semantic cohesion during emotional word encoding. *Brain and Cognition*, *62*(1), 43–57. https://doi.org/10.1016/j.bandc.2006.03.008

Dillon, D. G., Dobbins, I. G., & Pizzagalli, D. A. (2014). Weak reward source memory in depression reflects blunted activation of VTA/SN and parahippocampus. *Social Cognitive and Affective Neuroscience*, *9*(10), 1576–1583. https://doi.org/10.1093/scan/nst155

Dunbar, G. C., & Lishman, W. A. (1984). Depression, recognition-memory and hedonic tone: a signal detection analysis. *British Journal of Psychiatry*, *144*(4), 376–382. https://doi.org/10.1192/bjp.144.4.376

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin and Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via stan*. Retrieved from httpsL//mc-stan.org/rstanarm

Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., & Koster, E. H. W. (2019, July 1). Motivation and cognitive control in depression. *Neuroscience and Biobehavioral Reviews*, Vol. 102, pp. 371–381. https://doi.org/10.1016/j.neubiorev.2019.04.011

Hallford, D. J., Rusanov, D., Yeow, J. J. E., & Barry, T. J. (2021, April 1). Overgeneral and specific autobiographical memory predict the course of depression: An updated meta-analysis. *Psychological Medicine*, Vol. 51, pp. 909–926. https://doi.org/10.1017/S0033291721001343

Hamilton, J. P., & Gotlib, I. H. (2008). Neural Substrates of Increased Memory Sensitivity for Negative Stimuli in Major Depression. *Biological Psychiatry*, *63*(12), 1155–1162. https://doi.org/10.1016/j.biopsych.2007.12.015

Kruschke, J. K. (2014). Doing Bayesian Data Anlaysis. In *Doing Bayesian Data Analysis*. https://doi.org/10.1016/B978-0-12-405888-0.09996-7

Landauer, T. K. (2006). Latent Semantic Analysis. In *Encyclopedia of Cognitive Science*. https://doi.org/10.1002/0470018860.s00561

Lawlor, V. M., Webb, C. A., Wiecki, T. V., Frank, M. J., Trivedi, M., Pizzagalli, D. A., & Dillon, D. G. (2020). Dissecting the impact of depression on decision-making. *Psychological Medicine*, *50*, 1613–1622. https://doi.org/10.1017/S0033291719001570

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates.

MacQueen, G. M., Campbell, S., McEwen, B. S., Macdonald, K., Amano, S., Joffe, R. T., … Trevor Young, L. (2003). Course of illness, hippocampal function, and hippocampal volume in major depression. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 1387–1392. https://doi.org/10.1073/pnas.0337481100

MacQueen, G. M., & Frodl, T. (2011). The hippocampus in major depression: Evidence for the convergence of the bench and bedside in psychiatric research. *Molecular Psychiatry*, *16*(3), 252–

264. https://doi.org/10.1038/mp.2010.80

MacQueen, G. M., Galway, T. M., Hay, J., Young, L. T., & Joffe, R. T. (2002). Recollection memory deficits in patients with major depressive disorder predicted by past depressions but not current mood state or treatment status. *Psychological Medicine*, *32*, 251–258. https://doi.org/10.1017/S0033291701004834

Matt, G. E., Vazquez, C., & Campbell, W. K. (1992). Mood-congruent recall of affectively toned stimuli: A meta-analytic review. *Clinical Psychology Review*, *12*, 227–255. https://doi.org/10.1016/0272-7358(92)90116-P

Meehl, P. E. (1975). Hedonic capacity: some conjectures. *Bulletin of the Menninger Clinic*, *39*(1975), 295–307.

Passell, E., Dillon, D. G., Baker, J. T., Vogel, S. C., Scheuer, L. S., Mirin, N. L., Rutter, L. A., Pizzagalli, D. A., & Germine, L. (2019). *Digital Cognitive Assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) Field Test Battery Report*. https://psyarxiv.com/dcszr/

Pizzagalli, D. A. (2014). Depression, Stress, and Anhedonia: Toward a Synthesis and Integrated Model. *Annual Review of Clinical Psychology*, *10*(1), 393–423. https://doi.org/10.1146/annurev-clinpsy-050212-185606

Proudfit, G. H. (2015). The reward positivity: from basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459. https://doi.org/10.1111/psyp.12370

Ramponi, C., Barnard, P. J., & Nimmo-Smith, I. (2004). Recollection deficits in dysphoric mood: An effect of schematic models and executive mode? *Memory*, *12*(5), 655–670. https://doi.org/10.1080/09658210344000189

Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037//0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Rock, P. L., Roiser, J. P., Riedel, W. J., & Blackwell, A. D. (2014). Cognitive impairment in depression:

a systematic review and meta-analysis. *Psychological Medicine*, *44*(10), 2029–2040. https://doi.org/10.1017/S0033291713002535

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*. https://doi.org/10.1037//0022-3514.35.9.677

Rouhani, N., & Niv, Y. (2019). Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, *236*(8), 2425–2435. https://doi.org/10.1007/s00213-019-05322-z

Scimeca, J. M., Katzman, P. L., & Badre, D. (2016). Striatal prediction errors support dynamic control of declarative memory decisions. *Nature Communications*. https://doi.org/10.1038/ncomms13061

Shadlen, M. N., & Shohamy, D. (2016). *Neuron Perspective Decision Making and Sequential Sampling from Memory*. https://doi.org/10.1016/j.neuron.2016.04.036

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*(1), 34–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2966230

Snyder, H. R. (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review. *Psychological Bulletin*, *139*(1), 81–132. https://doi.org/10.1037/a0028727

Starns, J. J., Dubé, C., & Frelinger, M. E. (2018). The speed of memory errors shows the influence of misleading information: Testing the diffusion model and discrete-state models. *Cognitive Psychology*, *102*, 21–40. https://doi.org/10.1016/j.cogpsych.2018.01.001

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: an analysis of the strength-based mirror effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(5), 1137–1151. https://doi.org/10.1037/a0028151

Strauman, T. J. (2002). Self-Regulation and Depression. *Self and Identity*. https://doi.org/10.1080/152988602317319339

Strauman, T. J. (2017). Self-Regulation and Psychopathology: Toward an Integrative Translational Research Paradigm. *Annual Review of Clinical Psychology*. https://doi.org/10.1146/annurev-clinpsy-032816-045012

Sumner, J. A., Griffith, J. W., & Mineka, S. (2010). Overgeneral autobiographical memory as a predictor of the course of depression: A meta-analysis. *Behaviour Research and Therapy*, *48*(7), 614–625. https://doi.org/10.1016/j.brat.2010.03.013

Talmi, D., & Moscovitch, M. (2004). Can semantic relatedness explain the enhancement of memory for emotional words? *Memory and Cognition*, *32*(5), 742–751. https://doi.org/10.3758/BF03195864

Treadway, M. T., & Zald, D. H. (2011, January 1). Reconsidering anhedonia in depression: Lessons from translational neuroscience. *Neuroscience and Biobehavioral Reviews*, Vol. 35, pp. 537–555. https://doi.org/10.1016/j.neubiorev.2010.06.006

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., … Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (September 2017). *Zenodo*, p. https://zenodo.org/record/883859#.XwOHPyhKg2w. https://doi.org/10.5281/zenodo.883859

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion-model analysis. *Cognition & Emotion*, *23*(1), 181–205. https://doi.org/10.1080/02699930801976770

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89–106. https://doi.org/10.3758/BF03200763

Zakzanis, K. K., Leach, L., & Kaplan, E. (1998). On the nature and pattern of neurocognitive function in major depressive disorder. *Neuropsychiatry, Neuropsychology and Behavioral Neurology*, *11*(3), 111–119.

Zorowitz, S., Niv, Y., & Bennett, D. (n.d.). Inattentive responding can induce spurious associations between task behavior and symptom measures. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/RYNHK

Table 1. *Estimated effects of age, gender, word valence, encoding task, and BDI on encoding response*

	Mean	Lower HDI	Upper HDI	R-hat
(Intercept)	−2.481	−2.741	−2.215	1.005
age	−0.003	−0.006	0.000	1.000
genderqueer	0.187	−0.067	0.441	1.001
male	0.093	0.016	0.173	1.000
positive word	4.103	3.783	4.427	1.003
valence task	−0.562	−0.683	−0.439	1.000
BDI	0.056	0.052	0.060	1.000
positive word:valence task	1.790	1.639	1.947	1.000
positive word:BDI	−0.095	−0.099	−0.090	1.000
valence task:BDI	−0.018	−0.023	−0.012	1.000
positive word:valence task:BDI	0.024	0.017	0.032	1.000

Note: The HDI represents the 95% highest density interval from the posterior distribution.
The reference level is a female participant responding to a negative word in the self-reference task.

Table 2. *Estimated effects of age, gender, word valence, encoding task, and BDI on number of correctly recalled words*

	Mean	Lower HDI	Upper HDI	R-hat
(Intercept)	0.236	0.120	0.353	1.003
age	−0.012	−0.014	−0.009	1.004
genderqueer	0.178	−0.019	0.375	1.001
male	−0.087	−0.147	−0.024	1.001
positive word	0.307	0.221	0.394	1.000
valence task	−0.176	−0.272	−0.077	0.999
BDI	−0.001	−0.005	0.003	1.001
positive word:valence task	0.036	−0.091	0.169	0.999
positive word:BDI	−0.008	−0.012	−0.003	0.999
valence task:BDI	0.000	−0.005	0.005	0.999
positive word:valence task:BDI	0.001	−0.006	0.008	0.999

Note: The HDI represents the 95% highest density interval from the posterior distribution.
The reference level is a female participant recalling negative words from the self-reference task.

Table 3. *Estimated effects of age, gender, relevant task conditions, and BDI on drift rates*

	Mean	Lower HDI	Upper HDI	R-hat
Recognition				
(Intercept)	−0.960	−1.004	−0.913	1.000
age	−0.001	−0.001	0.000	1.000
genderqueer	0.032	−0.043	0.111	1.000
male	0.018	−0.005	0.041	1.000
positive word	0.126	0.083	0.170	0.999
old word	1.739	1.695	1.782	1.000
BDI	0.001	0.000	0.003	0.999
positive word:old word	0.260	0.201	0.323	1.000
positive word:BDI	−0.003	−0.005	0.000	0.999
old word:BDI	0.003	0.000	0.005	0.999
positive word:old word:BDI	−0.003	−0.006	0.000	0.999
Source Memory				
(Intercept)	0.170	0.114	0.224	0.999
age	0.000	−0.001	0.001	1.000
genderqueer	0.007	−0.073	0.090	0.999
male	−0.005	−0.028	0.018	1.000
positive word	0.669	0.614	0.726	0.999
valence task	−0.863	−0.919	−0.807	0.999
BDI	0.007	0.005	0.009	0.999
positive word:valence task	−0.301	−0.380	−0.223	1.000
positive word:BDI	−0.013	−0.016	−0.010	1.000
valence task:BDI	−0.002	−0.005	0.001	0.999
positive word:valence task:BDI	0.004	0.000	0.008	0.999

Note: The HDI represents the 95% highest density interval from the posterior distribution.
For recognition, the reference level is a female participant responding to a new negative word.
For source memory, the reference level is a female participant responding to a negative word from the self-reference task.

Table 4. *Estimated effects of age, gender, word valence, and BDI on SDT parameters*

	d'				*c*			
	Mean	Low HDI	Upper HDI	R-hat	Mean	Lower HDI	Upper HDI	R-hat
Recognition								
(Intercept)	2.484	2.365	2.603	1.002	0.492	0.435	0.549	1.002
age	−0.011	−0.013	−0.008	1.001	0.001	0.000	0.003	1.003
genderqueer	0.001	−0.220	0.235	1.002	0.045	−0.061	0.156	1.001
male	−0.096	−0.165	−0.029	1.000	−0.004	−0.036	0.029	1.001
positive word	0.140	0.089	0.193	0.999	−0.150	−0.183	−0.117	0.999
BDI	−0.003	−0.006	0.001	1.000	−0.002	−0.004	0.000	1.001
positive:BDI	−0.001	−0.004	0.002	0.999	0.003	0.001	0.004	0.999
Source Memory								
(Intercept)	1.288	1.126	1.447	1.001	0.191	0.115	0.268	1.000
age	−0.009	−0.012	−0.005	1.001	0.001	0.000	0.003	1.000
genderqueer	0.218	−0.080	0.532	1.003	−0.005	−0.142	0.136	1.000
male	−0.107	−0.200	−0.009	1.004	−0.001	−0.044	0.042	1.000
positive word	0.249	0.175	0.326	0.999	−0.555	−0.619	−0.487	0.999
BDI	−0.002	−0.007	0.002	1.001	−0.009	−0.012	−0.007	1.000
positive:BDI	−0.004	−0.008	0.000	0.999	0.016	0.012	0.019	0.999

Note: The HDI represents the 95% highest density interval from the posterior distribution.
For both recognition and source memory, the reference level is a female participant responding to a negative word.

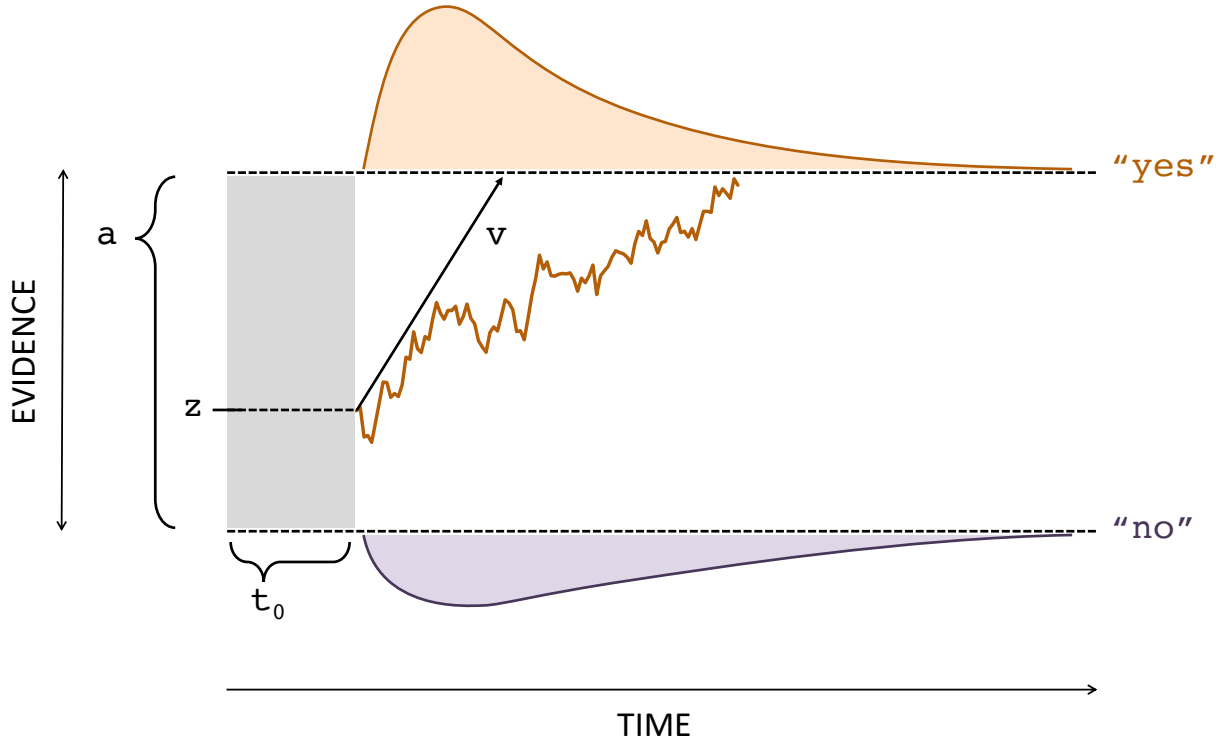

Figure 1. Diagram of the Drift Diffusion Model (DDM; Ratcliff & McKoon, 2008).

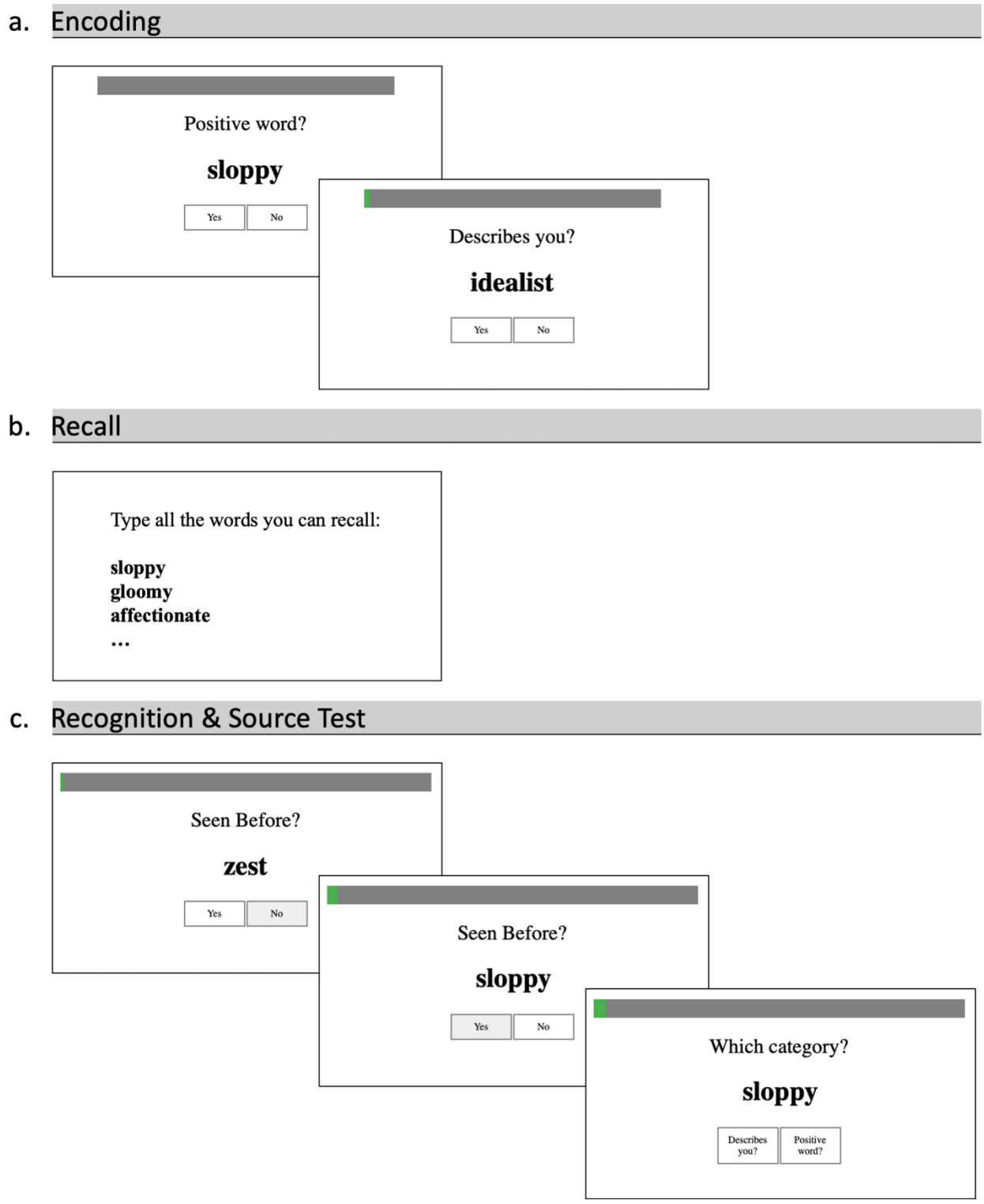

Figure 2. Diagram of the memory task, including (a) encoding, (b) free recall, and (c) the recognition memory test, which included a source memory test for words endorsed as "old" at recognition. A progress bar was presented at the top of the screen during each stage.

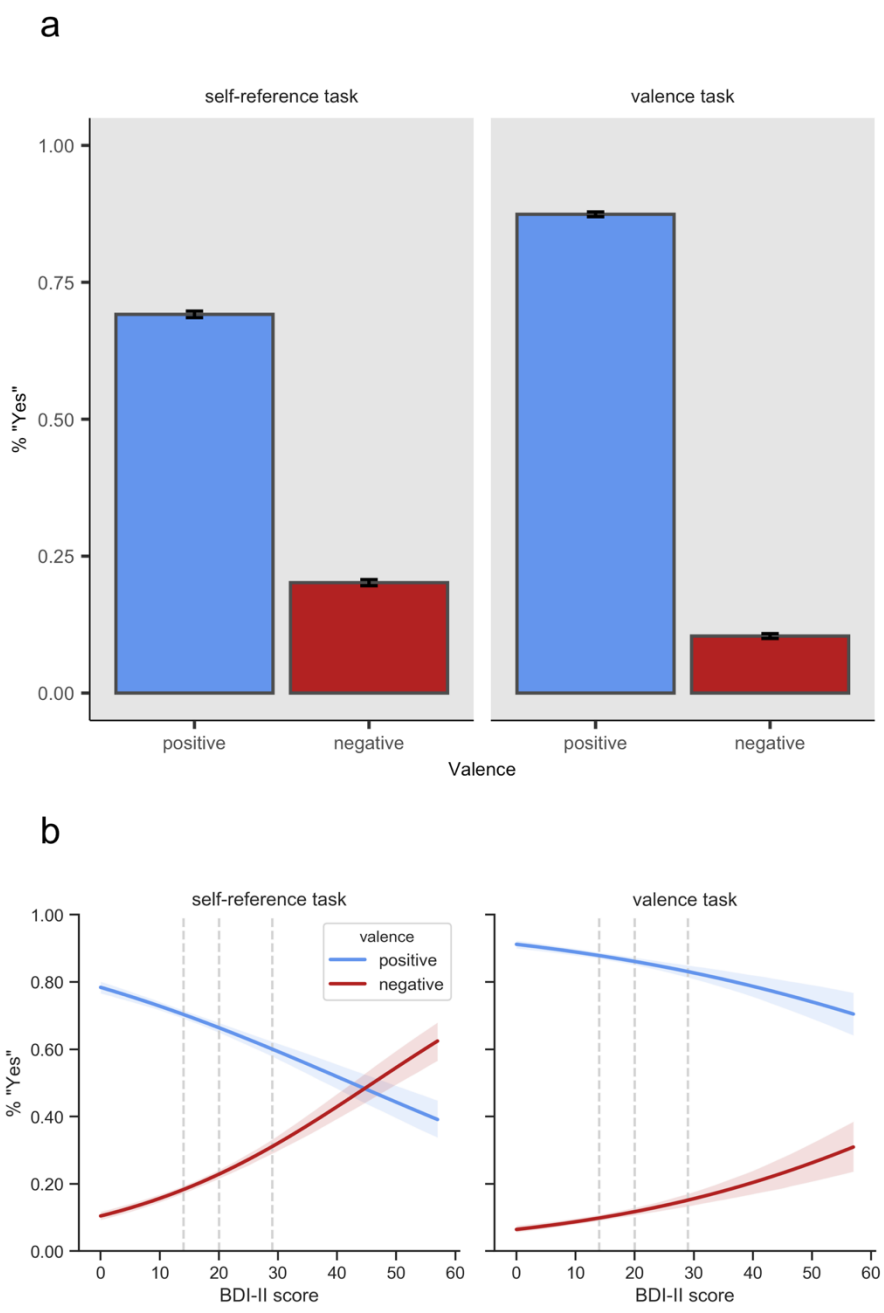

Figure 3. Proportion of "yes" responses at encoding. In each panel, columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). In Panel (b), dashed vertical lines denote standard cutoffs for mild, moderate, and severe depression. Panel (a) depicts a higher average proportion of "yes" responses to positive than negative words in both encoding tasks. Panel (b) demonstrates that as BDI scores increase, participants endorsed more negative but fewer positive words as self-descriptive (self-reference task) and positive (valence task).

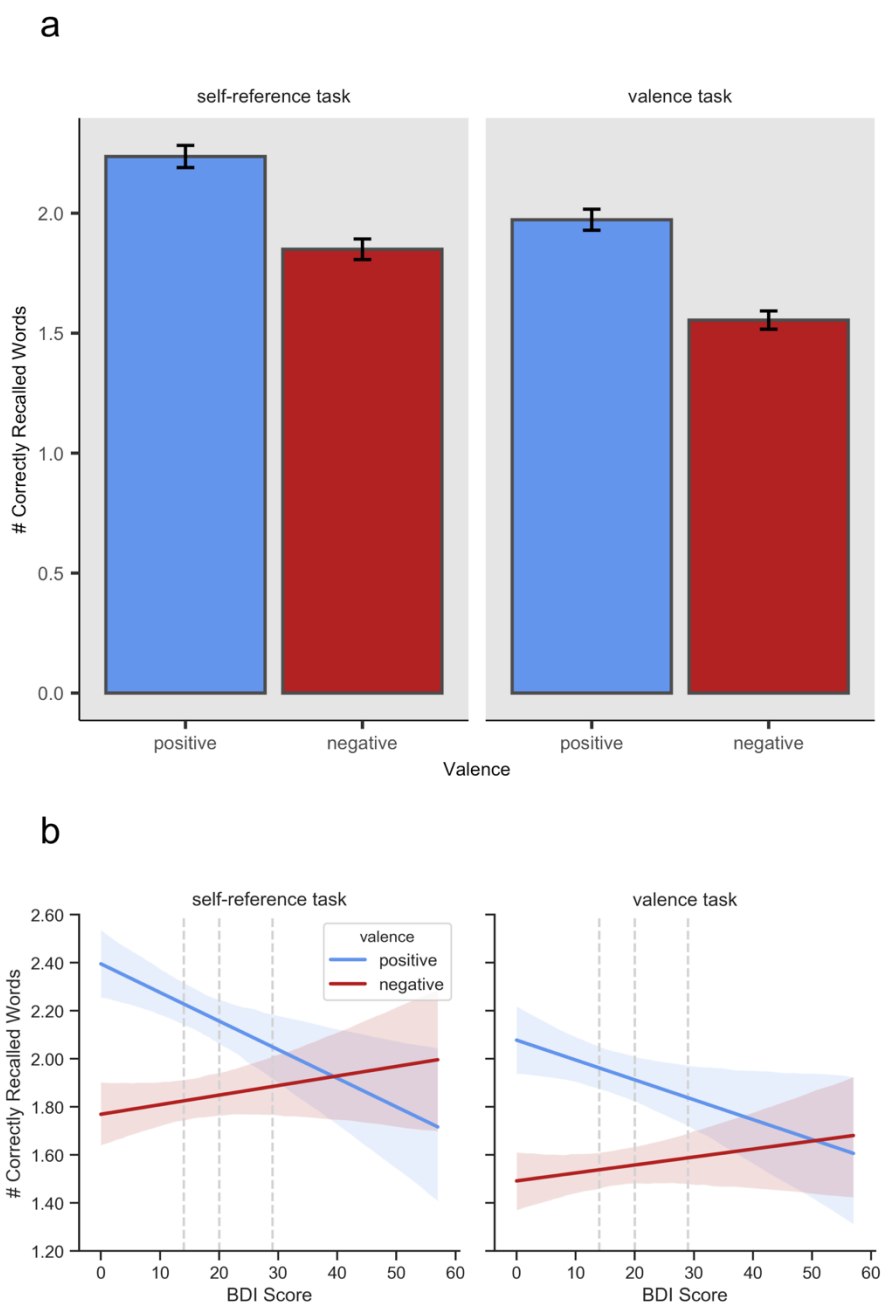

Figure 4. Number of correctly recalled words. In each panel, columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). In Panel (b), dashed vertical lines denote standard cutoffs for mild, moderate, and severe depression. Panel (a) depicts a higher average number of recalled positive vs. negative words from each encoding task, and a higher average number of words recalled from the self-reference task overall. Panel (b) demonstrates that higher BDI scores were associated with poorer recall of positive words but slightly better recall of negative words from both encoding tasks.

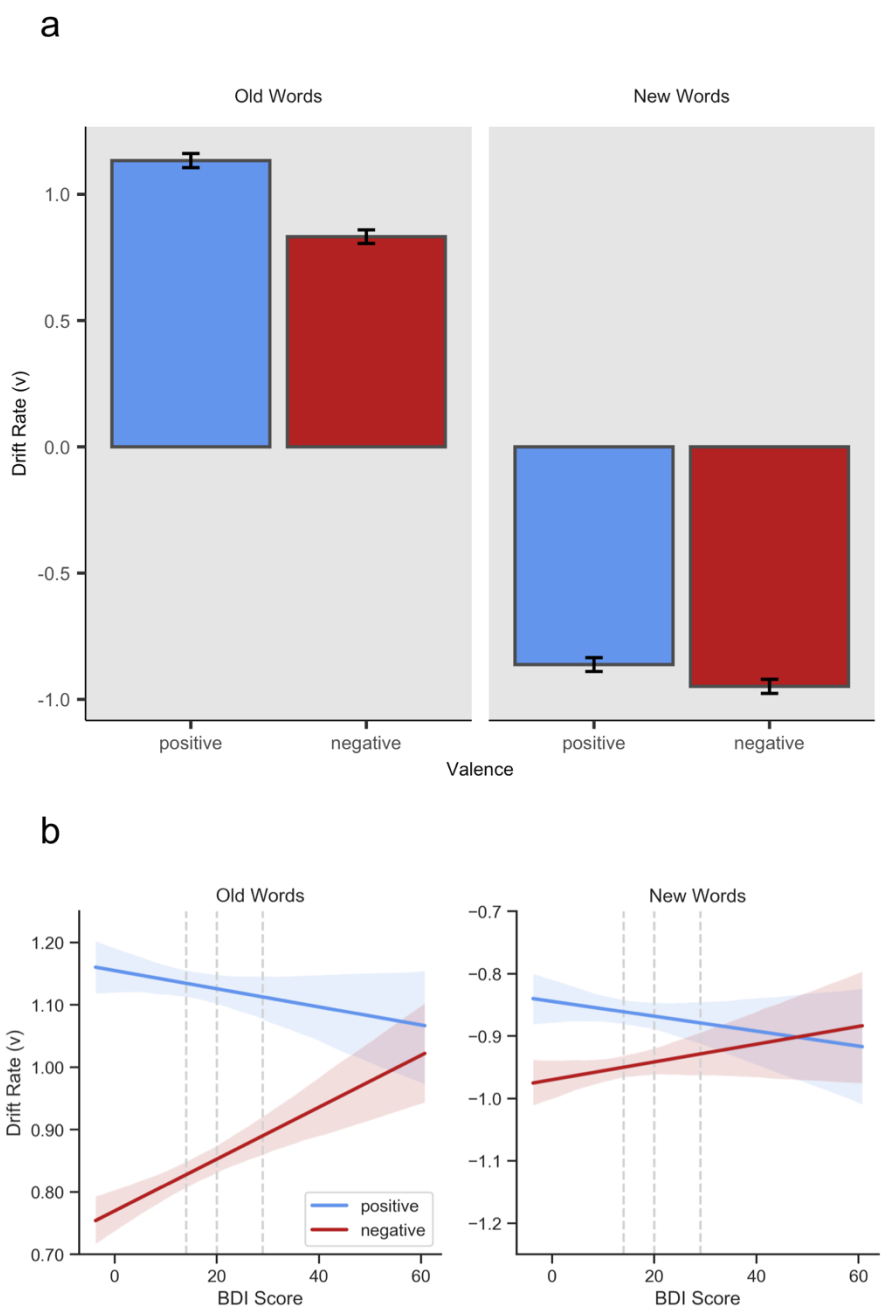

Figure 5. HDDM drift rate estimates for the recognition data. Positive values indicate evidence accumulation towards an "old" response, while negative values indicate evidence accumulation towards a "new" response. In each panel, columns denote study status (left: old/studied words, right: new/unstudied words), and colors denote normative valence (blue: positive, red: negative). Error bars in panel (a) represent 95% HDIs. Error bands in panel (b) represent 95% bootstrap confidence intervals (Waskom et al., 2017). In Panel (b), dashed vertical lines denote standard cutoffs for mild, moderate, and severe depression. Panel (a) depicts more efficient evidence accumulation towards an "old" response for old positive vs. old negative words, but less efficient accumulation towards a "new" response for new positive vs. new negative words. Panel (b) demonstrates that as BDI scores increase, drift rate increases (towards an "old" response) for negative words but weakly decreases (towards a "new" response) for positive words, regardless of whether the word was actually old or new.

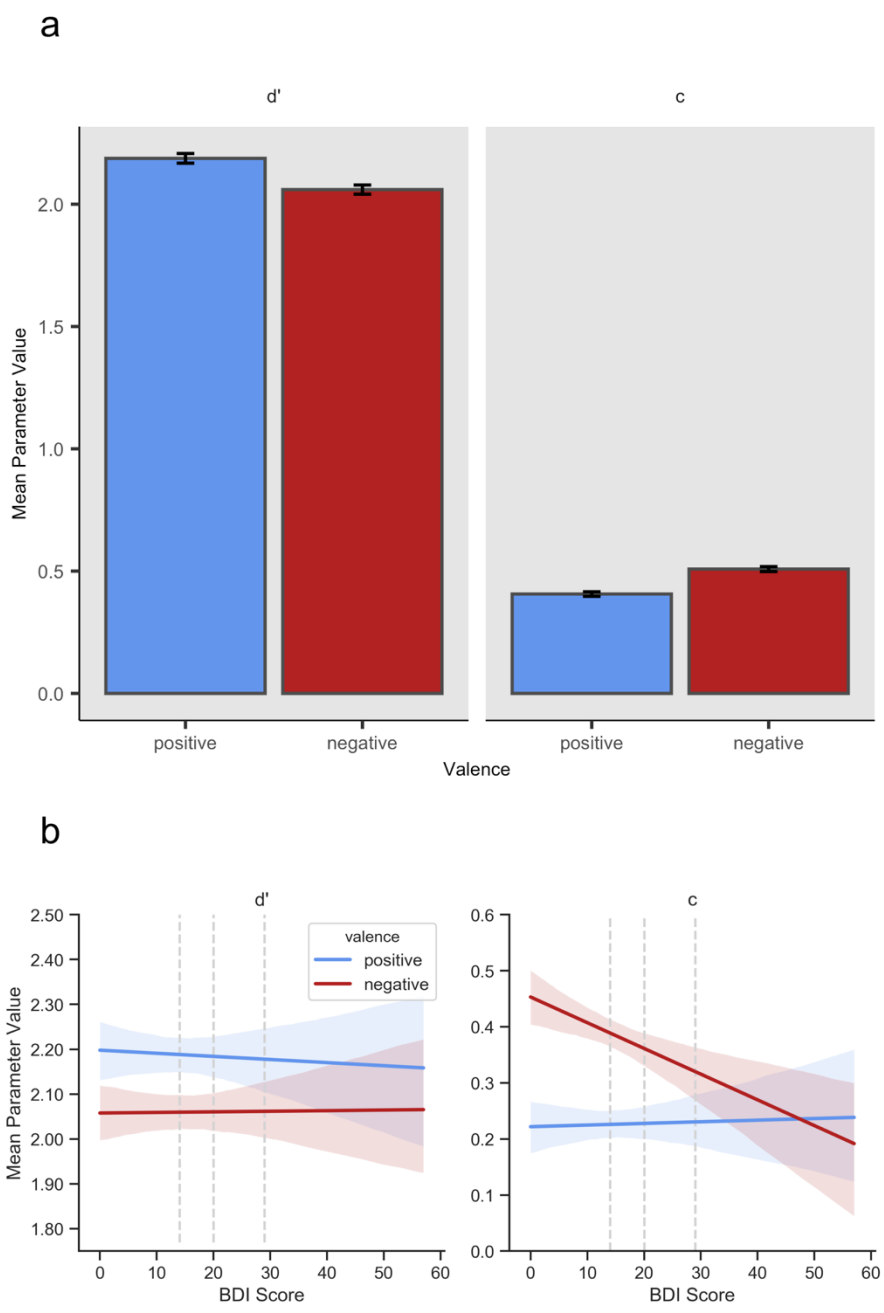

Figure 6. SDT parameter values for the recognition data. In each panel, columns denote parameter (left: *d'*, right: *c*), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). In Panel (b), dashed vertical lines denote standard cutoffs for mild, moderate, and severe depression. Panel (a) depicts higher average values of *d'* and lower (more liberal) average values of *c* for positive vs. negative words. Panel (b) demonstrates that as BDI score increases, *d'* is unaffected but *c* decreases (becomes more liberal) for negative words.

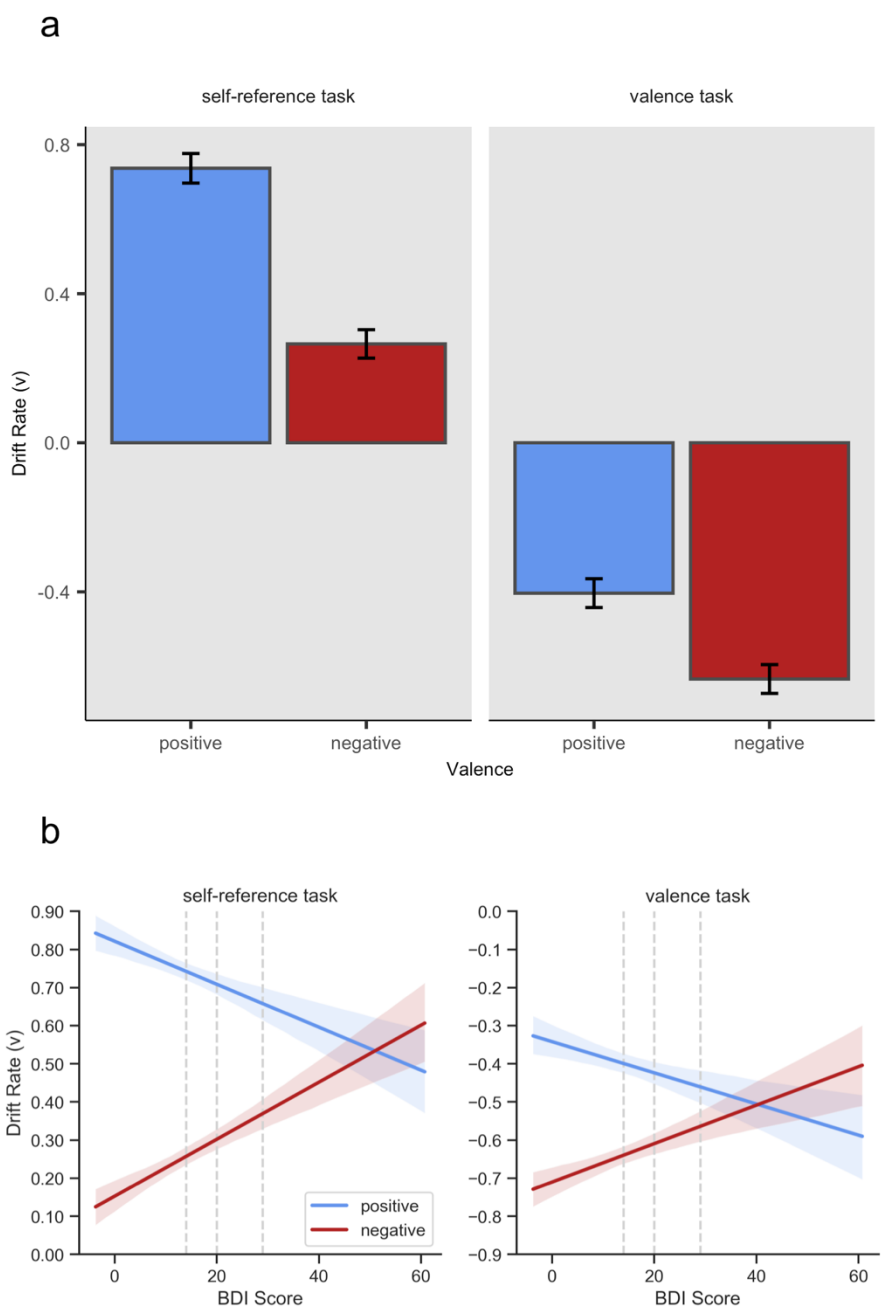

Figure 7. HDDM drift rate estimates for source memory. In each panel, columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars in panel (a) represent 95% HDIs. Error bands in panel (b) represent 95% bootstrap confidence intervals (Waskom et al., 2017). In Panel (b), dashed vertical lines denote standard cutoffs for mild, moderate, and severe depression. Panel (a) depicts more efficient evidence accumulation towards a "self-reference" response for positive vs. negative words from the self-reference task, but less efficient accumulation towards a "valence" response for positive vs. negative words from the valence task. Panel (b) demonstrates that as BDI scores increase, drift rate increases (towards a "self-reference" response) for negative words but decreases (towards a "valence" response) for positive words, regardless of task.

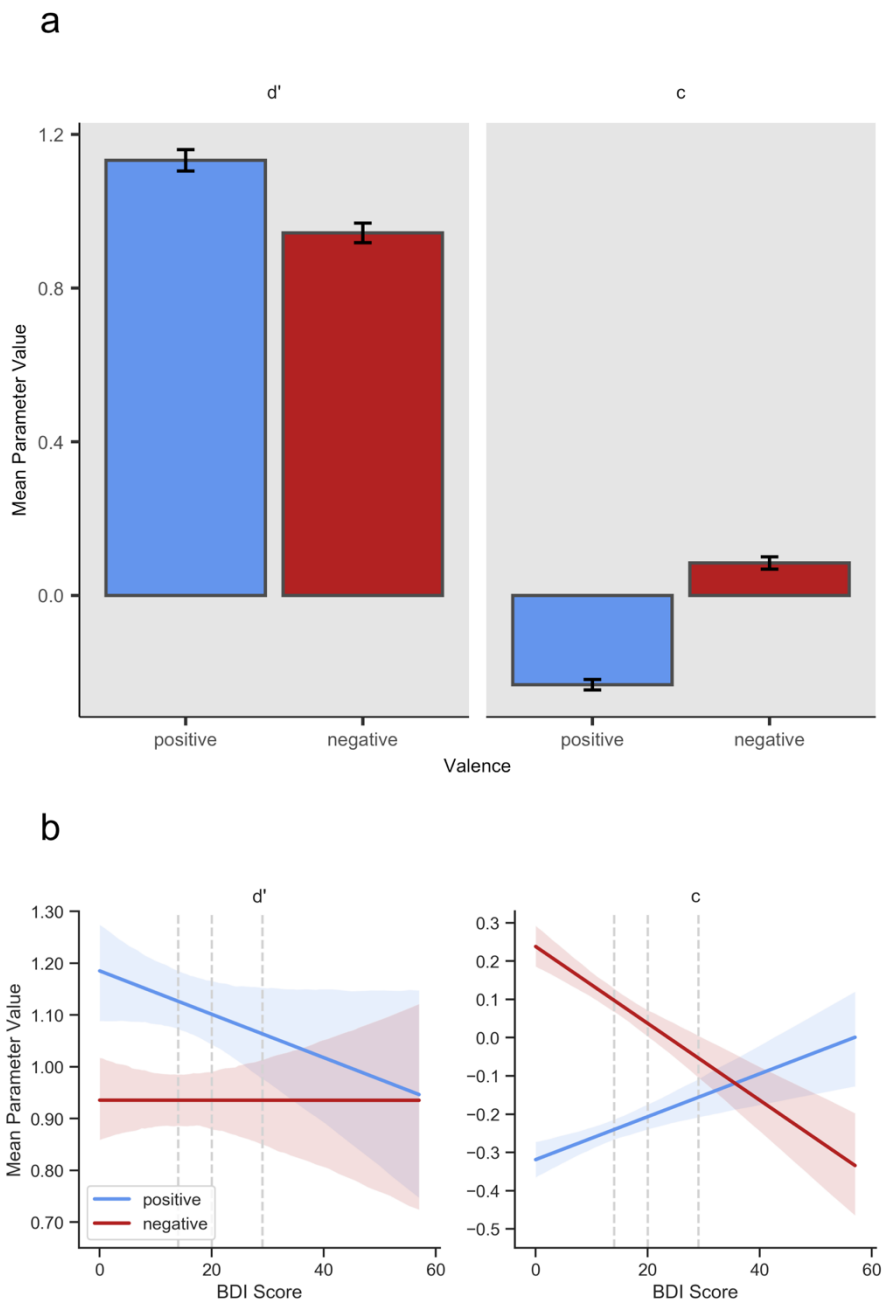

Figure 8. SDT parameter values for source memory. In each panel, columns denote parameter (left: *d'*, right: *c*), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). In Panel (b), dashed vertical lines denote standard cutoffs for mild, moderate, and severe depression. Panel (a) depicts higher average values of *d'* and lower (more liberal) average values of *c* for positive vs. negative words. Panel (b) demonstrates that as BDI score increases, *d'* decreases slightly for positive words but is unaffected for negative words, and that *c* decreases (becomes more liberal) for negative words but increases (becomes more conservative) for positive words. In the context of this model, the results for *c* indicate that higher BDI

scores are associated with a greater tendency to attribute old negative words to the self-referential encoding task and old positive words to the valence task.