Supplementary Materials for:

Abnormal Evidence Accumulation Underlies the Positive Memory Deficit in Depression

Andrea M. Cataldo, Luke Scheuer, Arkadiy L. Maksimovskiy, Laura T. Germine, Daniel G. Dillon
McLean Hospital / Harvard Medical School

Correspondence to: amcataldo@mclean.harvard.edu

This file includes:

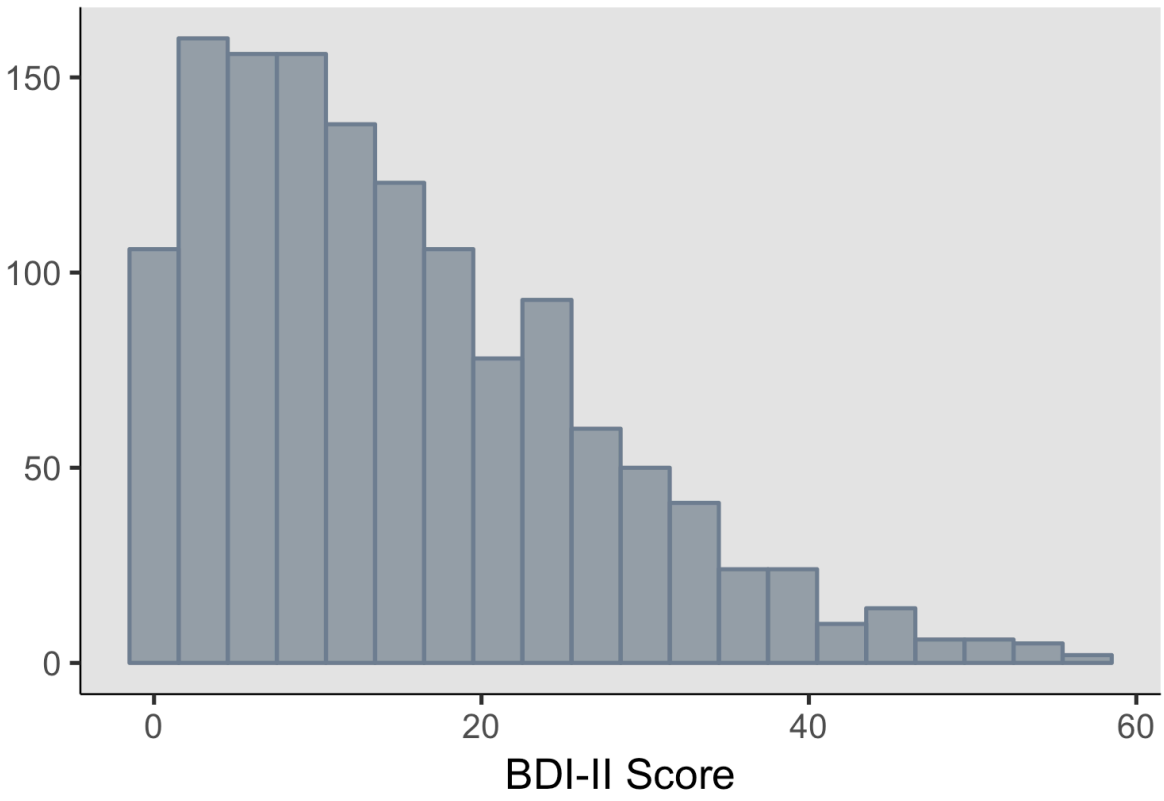

Figure S1. Distribution of BDI scores.

Recognition Hit Rates

Because the effects of encoding task on recognition accuracy cannot be easily incorporated into the HDDM and SDT analyses, we present here an analysis of the hit rate data from each task separately to supplement the models. 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 versus 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.

The above observations were confirmed by a Bayesian mixed effects logistic regression model: $p(\text{"}old\text{"}) \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$. Participants are more likely to correctly recognize positive vs. negative words from the self-reference task (95% HDI: 0.432:0.804). Participants are also estimated to correctly recognize fewer negative words from the valence task than the self-reference task (95% HDI: -0.462:-0.288), with no difference for positive words (95% HDI: -0.183:0.071). As BDI score increases, the number of correctly recognized positive words from the self-reference task decreases (95% HDI: -0.018:-0.008), but the number of correctly recognized negative words increases (95% HDI: 0.002:0.012). There were no additional interactions of BDI with encoding task for either positive (95% HDI: -0.002:0.011) or negative (95% HDI: -0.004:0.005) words.

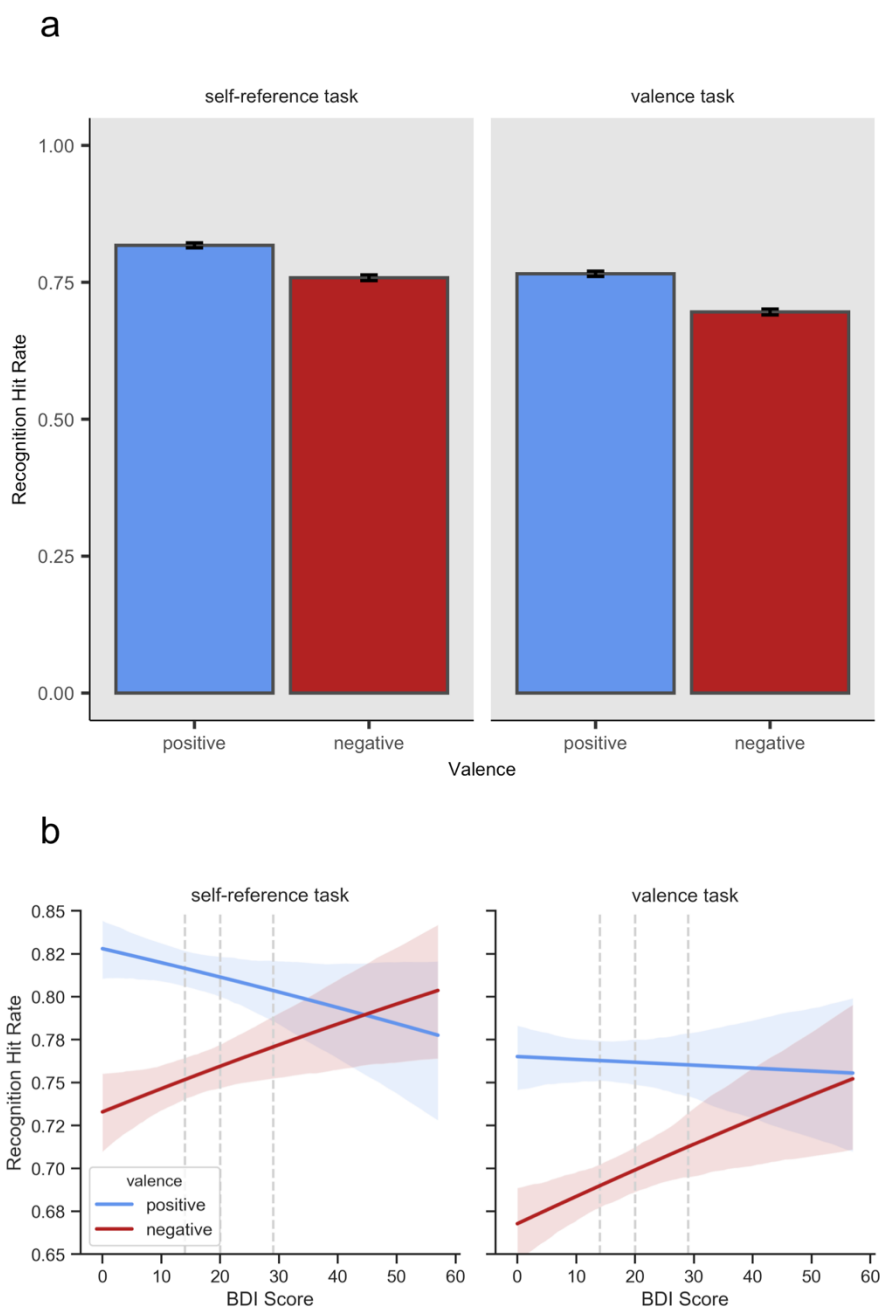

Figure S2. Hit rate at recognition. In each panel, columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). Panel (a) depicts a higher average hit rate for positive vs. negative words, and for words from the self-reference vs. valence task. Panel (b) demonstrates that as BDI scores increase, hit rates increase for negative words regardless of task but weakly decrease for positive words from the self-reference task.

Recognition Performance for Non-Recalled Words

It is possible that by testing free recall before recognition memory, item effects were introduced for recalled words. We therefore present here a re-analysis of the recognition data reported in the main text, with recalled words excluded.

Hit Rates

As shown in Figure S3a, participants correctly recognized more old positive vs. old negative words, and more words from the self-reference task versus valence task. These effects do not appear to interact: The recognition advantage for positive versus negative words appears similar for words from both tasks. As shown in Figure S3b, higher BDI scores were associated with better recognition of negative words from both tasks, and with slightly poorer recognition of positive words from the self-reference task.

These observations were confirmed by a Bayesian mixed effects logistic regression model: $p("old") \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$. Participants are more likely to correctly recognize positive versus negative words from the self-reference task (95% HDI: 0.356:0.744). Participants are also estimated to correctly recognize fewer negative words from the valence task than the self-reference task (95% HDI: -0.451:-0.274), with no difference for positive words (95% HDI: -0.165:0.101). As BDI score increases, the number of correctly recognized positive words from the self-reference task decreases (95% HDI: -0.016:-0.005), but the number of correctly recognized negative words increases (95% HDI: 0.002:0.012). There were no additional interactions of BDI with encoding task for either positive (95% HDI: -0.005:0.010) or negative (95% HDI: -0.004:0.005) words.

HDDM

Figure S4a presents the estimated mean drift rates by condition. 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 words (95% HDI: 0.977:1.031) relative to old negative words (95% HDI: 0.707:0.759). In contrast, the right panel shows that evidence accumulated slightly less efficiently towards a "new" response (i.e., was less negative) for new positive words (95% HDI: -0.879:-0.826) versus new negative words (95% HDI: -0.962:-0.909).

Figure S4b 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. These impressions were largely supported by a Bayesian mixed effects regression model: $v \sim age + gender + valence \cdot study \cdot BDI + (1 \mid subject)$. BDI score is not clearly associated with drift rates for new negative words (95% HDI: -0.001:0.002), but there is a relative increase for old negative words (95% HDI: 0.001:0.005). We note that this is very similar to the findings in the main text, in which the 95% HDI bordered zero for each coefficient. In contrast, the model estimates that, relative to negative words, drift rates decrease with BDI score for new positive words (95% HDI: -0.005:-0.001), with no estimated difference for old positive words (95% HDI: -0.004:0.002).

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 no effect on thresholds (95% HDI: -0.002:0.001), a weakly increased bias towards "old" responses (95% HDI: 0.000:0.001), and weakly increased non-decision times (95% HDI: 0.000:0.002).

SDT

Figure S5a 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 S5b reveals that memory strength for positive and negative words was unaffected by BDI score, as was

response bias for positive words; however, *c* decreased with BDI score for negative words, indicating greater willingness to endorse old and new negative words as "old" as the severity of depression increases. This selective effect of BDI on response bias for negative words is consistent with the positive association between BDI scores and hit rates for negative words, but not positive words (see Figure S3).

Each parameter was analyzed with a Bayesian mixed effects regression model: $parameter \sim age + gender + valence \cdot BDI + (1 \mid subject)$. Consistent with the observations above, positive words are associated with higher *d'* scores (95% HDI: 0.57:1.66) a stronger bias to respond "old" (95% HDI: -0.172:-0.107) relative to negative words. The model does not estimate a relationship between BDI score and *d'* for either positive (95% HDI: -0.002:0.003) or negative (95% HDI: -0.004:0.002) words. There was a modest negative association between BDI scores and *c* for negative words (95% HDI: -0.004:-0.001), confirming the impression from the right panel of Figure S5b. As in the main text, 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 S5b.

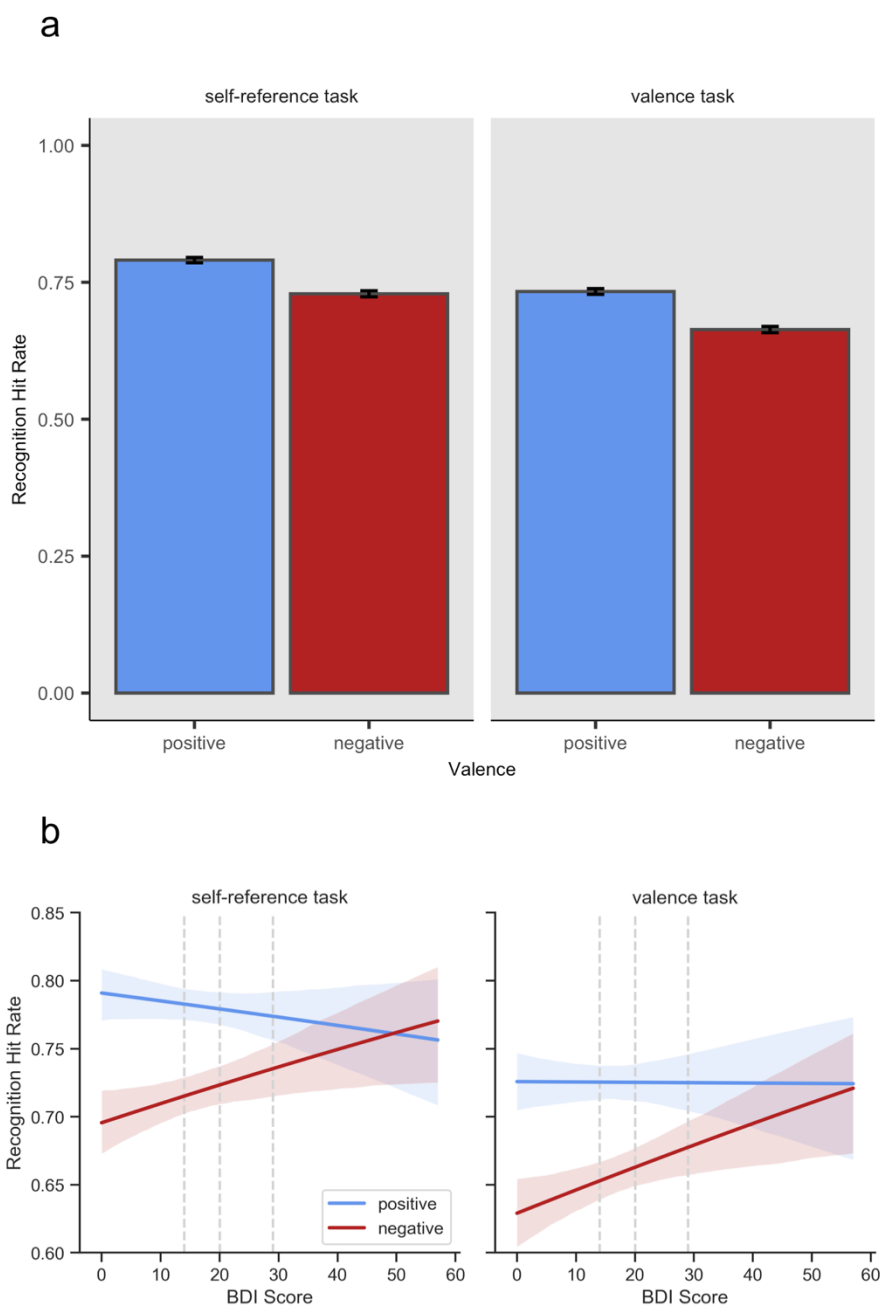

Figure S3. Hit rate at recognition, excluding recalled words. In each panel, columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). Panel (a) depicts a higher average hit rate for positive vs. negative words, and for words from the self-reference vs. valence task. Panel (b) demonstrates that as BDI scores increase, hit rates increase for negative words from both tasks but weakly decrease for positive words from the self-reference task.

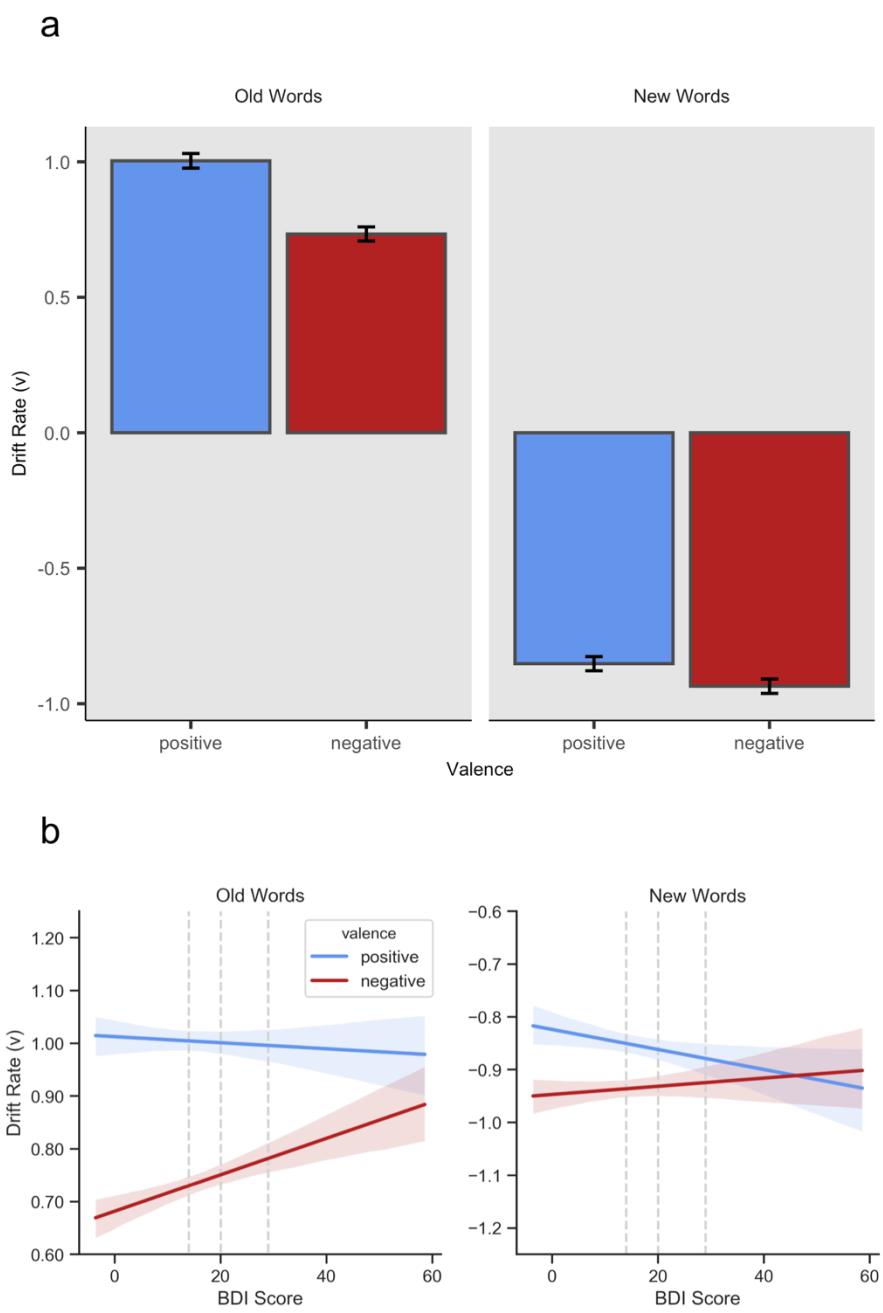

Figure S4. HDDM drift rate estimates for the recognition data. Positive values indicate evidence accumulation towards an "old" response, while negative values indicate evidence accumulation towards a "new" response. In each panel, columns denote study status (left: old/studied words, right: new/unstudied 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). 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. The impact of BDI on drift rate is especially apparent for old negative words.

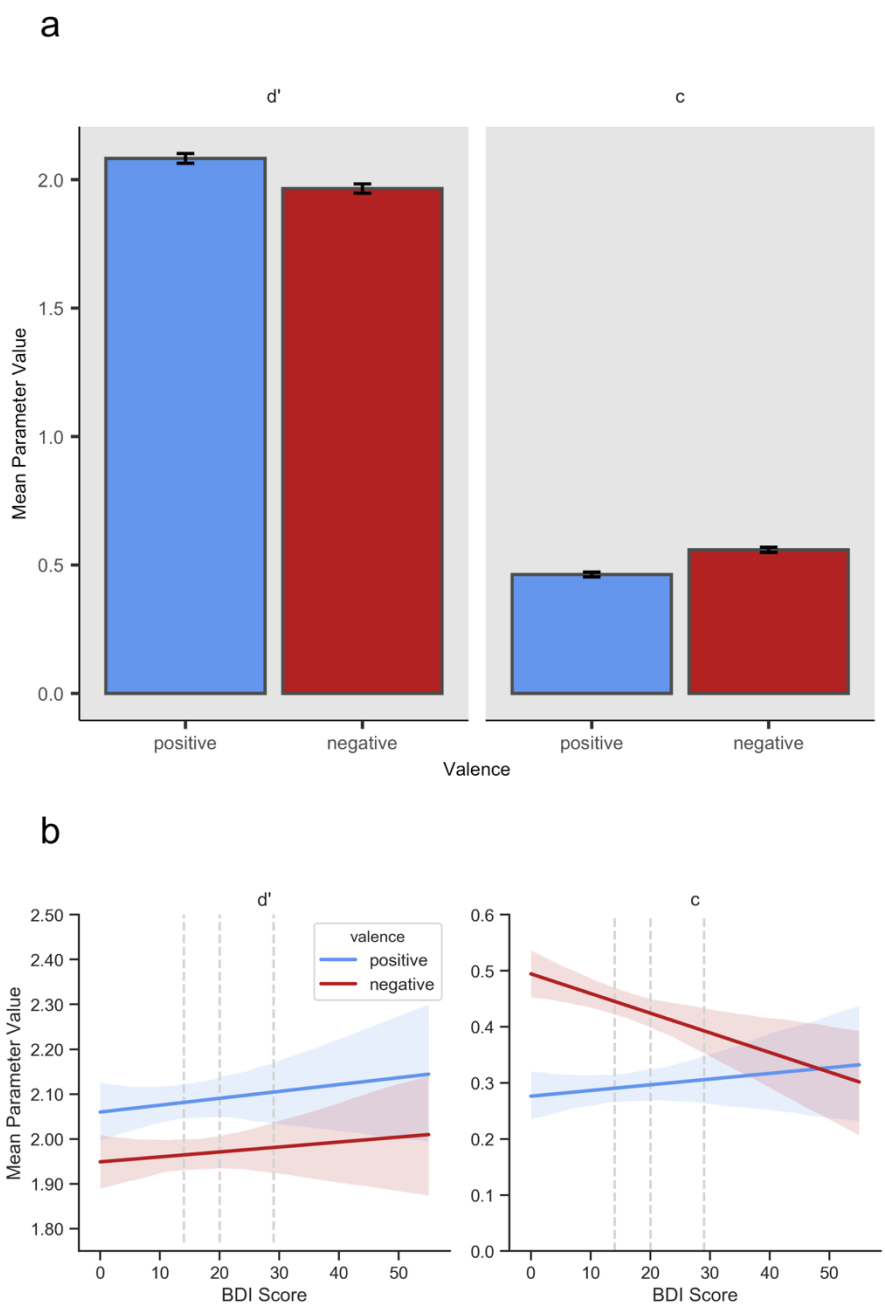

Figure S5. SDT parameter values for the recognition data. In each panel, columns denote parameter (left: *d'*, right: *c*), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). Panel (a) depicts higher 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.

Source Accuracy for Recognition Hits

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 present here an analysis of the response proportions from each task separately to supplement the models. As predicted, Figure S6a reveals greater source 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.

The above observations were confirmed by a Bayesian mixed effects logistic regression model: $p(correct) \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$. Participants are more likely to correctly identify the source of positive vs. negative words from the self-reference task (95% HDI: 1.143:1.390). Participants are more likely to correctly identify the source of negative words from the valence task vs. the self-reference task (95% HDI: 0.730:0.919), with a corresponding decrease for positive words (95% HDI: -2.192:-1.926). As BDI score increases, accuracy increases in the self-reference task for negative words (95% HDI: 0.011:0.021), but decreases for positive words (95% HDI: -0.039:-0.029). These effects flip in the valence task, such that relative accuracy decreases for negative words (95% HDI: -0.040:-0.031) and increases for positive words (95% HDI: 0.052:0.066).

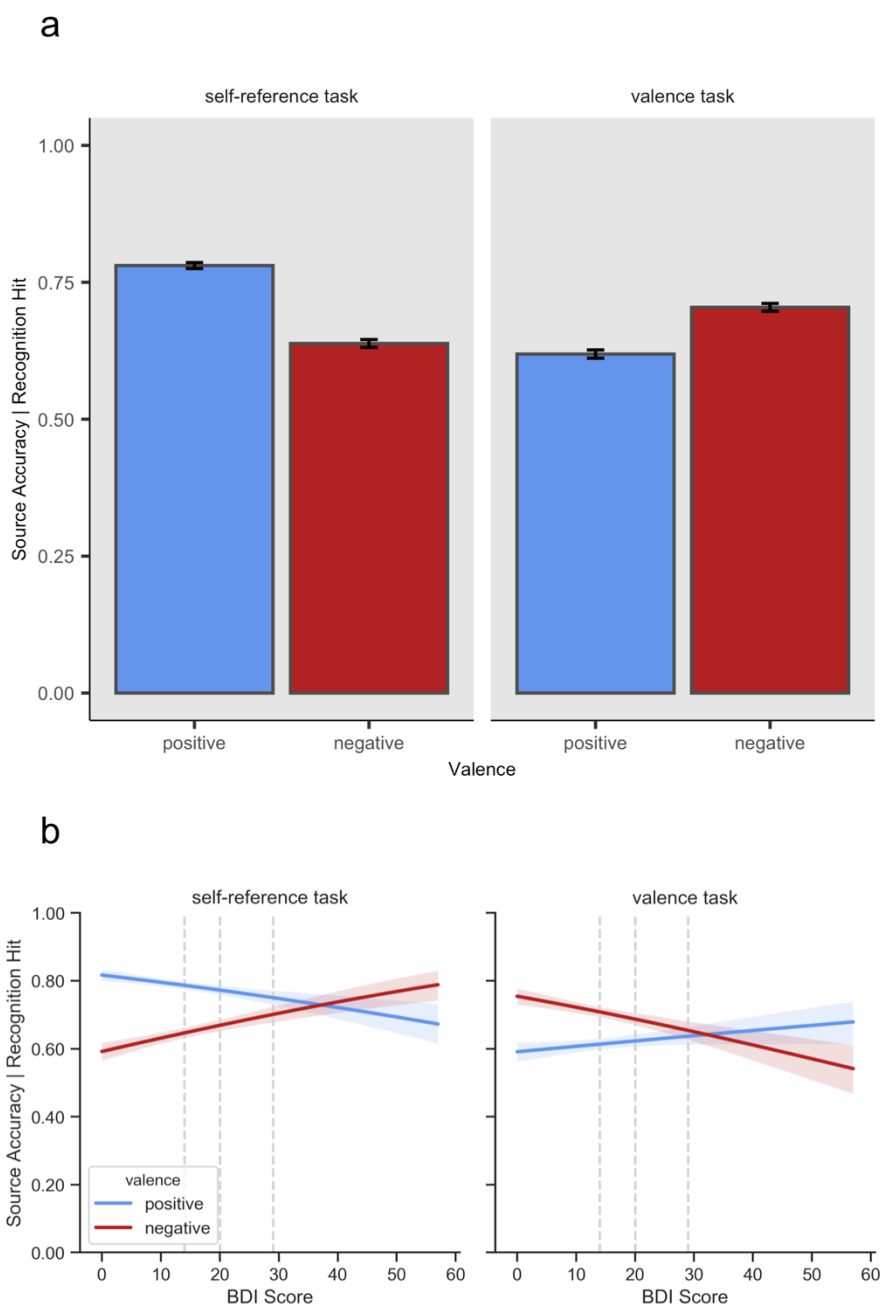

Figure S6. Source accuracy for recognition hits. Panel (a) presents average accuracy per condition. Panel (b) presents accuracy as a function of BDI score. In each panel, columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). Panel (a) depicts a higher average source accuracy for positive vs. negative words in the self-reference task, with the opposite effect in the valence task. Panel (b) demonstrates that as BDI scores increase, hit rates increase for negative words but weakly decrease for positive words in the self-reference task, again with an opposite effect in the valence task.

Source Attributions for Recognition False Alarms

The analyses of recognition and source memory reported in the main text consistently indicate a role for bias. Recall that at recognition, participants were asked to report the encoding task for each word that they reported as "old", regardless of whether the word had actually been studied. Though source attributions for studied words (recognition hits) can be driven by either memory strength or response bias, such attributions for unstudied words (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. We therefore present an analysis of source attributions for recognition false alarms to further explore the role of bias.

Because the data includes only unstudied words, there are no encoding task conditions to assign items to. Thus, a signal detection analysis (comparing the proportion of self-reference attributions from each task) was not feasible. HDDM fits were conducted, however drift rates varied only by valence.

Proportion of Self-Reference Attributions

Figure S7a reveals that the proportion of self-reference source attributions was greater for positive versus negative false alarms. Figure S7b shows that as BDI scores increased, the proportion of self-reference source attributions increases for negative false alarms, but slightly decreases for positive false alarms. These observations were confirmed by a Bayesian mixed effects logistic regression model: $p(correct) \sim age + gender + valence \cdot BDI + (1 \mid subject) + (1 \mid word)$. Participants are more likely to attribute positive versus negative false alarms to the self-reference task (95% HDI: 0.920:1.427). As BDI score increases, the probability of a self-reference source attribution increases for negative false alarms (95% HDI: 0.021:0.038), but decreases for positive false alarms (95% HDI: -0.050:-0.032).

HDDM

Figure S8a presents the estimated mean drift rates. Recall that positive and negative values indicated evidence accumulation towards "self-reference" and "valence" responses, respectively.

Participants had numerically positive, but near-zero, drift rates for positive false alarms, suggesting very slow accumulation that trends towards a "self-reference" response on average (95% HDI: 0.013:0.098). In contrast, participants had negative drift rates for negative false alarms, suggesting accumulation towards a "valence" response on average (95% HDI: -0.377:-0.225).

As shown in Figure S8b, as BDI score increased participants tended to accumulate evidence in favor of a "self-reference" response for negative false alarms and weakly in favor of a "valence" response for positive false alarms. These conclusions were supported by a Bayesian mixed effects regression model: $v \sim age + gender + valence \cdot BDI + (1 \mid subject)$. Higher BDI scores lead to more positive drift rates (more "self-reference" responses) for negative words (95% HDI: 0.005:0.009) but relatively more negative drift rates (more "valence" responses) for positive words (95% HDI: -0.012:-0.006).

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 no discernable changes in thresholds (95% HDI: -0.002:0.001), no discernable changes in starting point bias (95% HDI: 0.000:0.000), and weakly increased non-decision times (95% HDI: 0.000:0.001).

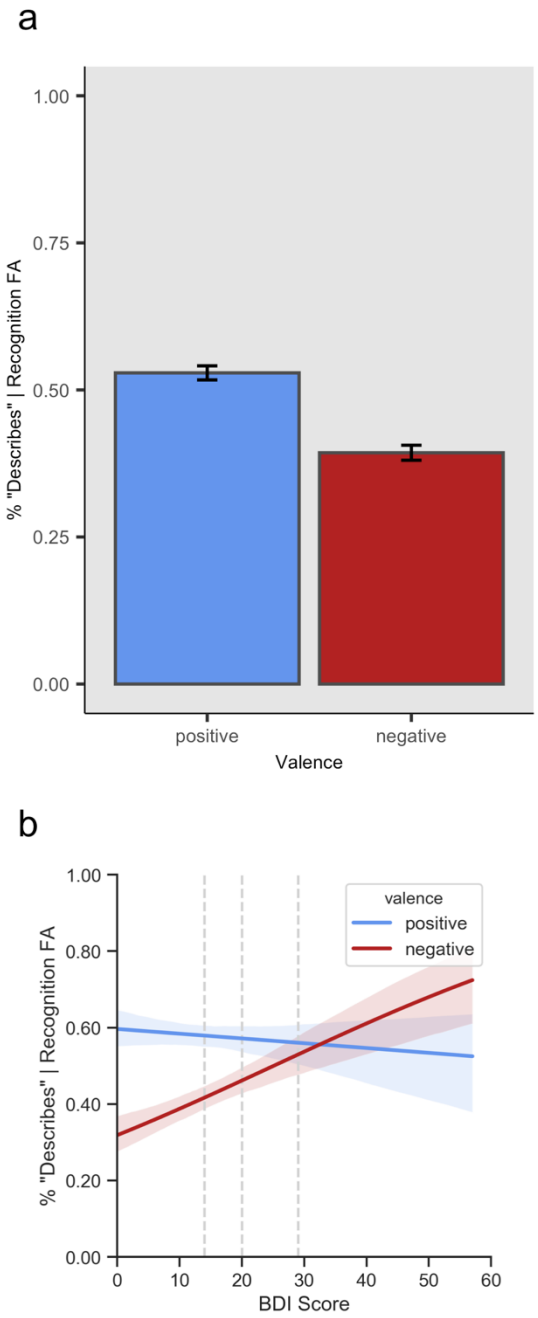

Figure S7. Source attributions for recognition false alarms. Panel (a) presents the average proportion of self-reference task attributions per condition. Panel (b) presents the proportion of self-reference task attributions as a function of BDI score. Colors denote normative valence (blue: positive, red: negative). Error bars and bands represent 95% bootstrap confidence intervals (Waskom et al., 2017). Panel (a) depicts a higher average proportion of self-reference source attributions for positive vs. negative false alarms. Panel (b) demonstrates that as BDI score increase, the proportion of self-reference source attributions increases for negative false alarms but weakly decreases for positive false alarms.

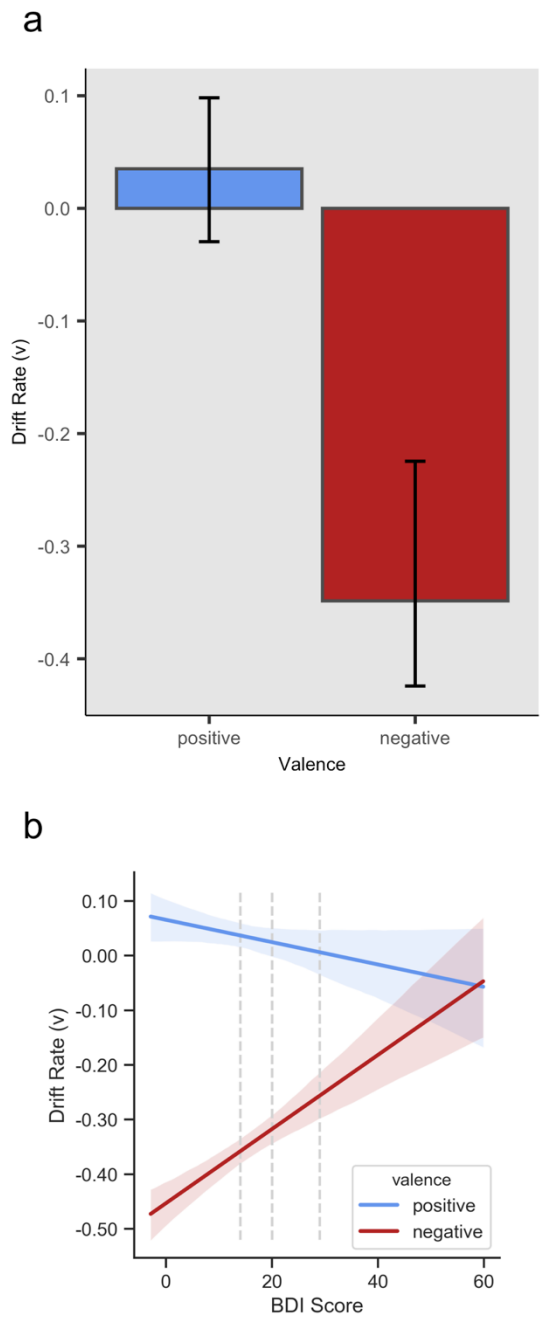

Figure S8. HDDM drift rate estimates for source attributions for recognition false alarms. Colors denote normative valence (blue: positive, red: negative). Error bars in panel (a) represent 95% HDIs. Error bands in panel (b) represent 95% bootstrap confidence intervals (Waskom et al., 2017). Panel (a) depicts sluggish evidence accumulation for positive false alarms and accumulation towards a "valence" response for negative false alarms. Panel (b) demonstrates that as BDI score increases, drift rate increases (towards a "self-reference" response) for negative false alarms but weakly decreases (towards a "valence" response) for positive false alarms.

Additional Samples without BDI Scores

The data included in this portion of the supplement constitute two pilot studies that preceded the study reported in the main text. These studies are highly similar to the main study, but (a) used much smaller samples and (b) do not include BDI measurements. Further details are included in the Methods section below. The Results section presents all analyses from the main text not pertaining to BDI scores. The pattern of results is consistent with those from the main study, with some exceptions related to the substantially reduced power and greater number of trials in these earlier studies.

Methods

In the following subsections, we report only that information which deviates from the main text.

Participants. As in the sample reported in the main text, participants were recruited through the online psychology experiment platform, TestMyBrain.org (Germine et al., 2012), in two samples (N_1=95, N_2=111). Prior to analyses, in each Sample (1, 2) we excluded (21, 25) participants for being younger than 18 years old, (5, 0) for not completing all study phases, and (2, 1) for having a negative recognition *d'* score, leaving a total of (67, 85) participants in the final analyses. The final samples were diverse: (52, 61)% were female and (67, 51)% were of European descent, with mean (SD) ages of 35.3 (14.4) and 38.1 (16.5) years in Samples 1 and 2, respectively. Limited analyses of these samples were previously described in an open access report by Passell et al (2019).

Self-Report Measures. Participants in all three samples provided basic demographic information, including age, gender, and ethnicity, prior to completing the task.

Memory Task. Samples 1 and 2 completed 100 encoding trials, instead of the 50 trials completed by participants in the main text. Due to a coding error, the word lists were not randomly assigned to each of the two encoding tasks for Sample 1; that is, old words were assigned to the same encoding task for almost all participants. This error was corrected in Sample 2 and for the participants in the main text.

Participants completed 200 recognition trials (100 words from encoding plus 100 lures), reflecting the increased number of studied words.

Results

Encoding. As shown in Figure S9 (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 or not a word was positive.

To quantify these impressions, the encoding responses were analyzed with a Bayesian mixed effects logistic regression model: $p("yes") \sim age + gender + valence \cdot task + (1 \mid subject) + (1 \mid word)$. The patterns observed above were all statistically supported. That is, the model indicated that participants are more likely to respond "yes" to a positive versus negative word in the self-reference task (95% HDIs: Sample 1, 2.565:3.309; Sample 2, 2.597:3.174). Participants are less likely to respond "yes" to negative words in the valence task relative to the self-reference task (95% HDIs: Sample 1, -1.806:-1.019; Sample 2, -1.288:-0.884), but are more likely to respond "yes" to positive words (95% HDIs: Sample 1, 2.438:3.518; Sample 2, 2.216:2.766).

Recall. As shown in Figure S10, participants recalled more positive than negative words, as well as more words from the self-reference task versus the valence task. The recall advantage for positive versus negative words appears to be reduced for words recalled from the valence task compared to the self-reference task in Sample 1, however this interaction does not appear in Sample 2. Relative to the sample reported in the main text, participants from Samples 1 and 2 recalled a greater number of words on average, reflecting the use of twice as many encoding trials as in the main study.

These observations were partially supported by a Bayesian mixed effects Poisson regression model: $\#\,recalled \sim age + gender + valence \cdot task + (1 \mid subject)$. Participants recall more positive than negative words from the self-reference task (95% HDIs: Sample 1, 0.165:0.537; Sample 2, 0.052:0.317). There was no estimated effect on the number of negative words recalled from the valence task vs. the self-reference task in Sample 1 (95% HDI: -0.417:0.011), however participants in Sample 2 were estimated to recall fewer negative words in the valence task (95% HDI: -0.586:-0.163), with no estimated difference for positive words in either sample (95% HDIs: Sample 1, -2.000:0.093; Sample 2, -0.190:0.379).

Recognition. *Hit rates*. As shown in Figure S11, participants correctly recognized more old positive versus old negative words, and more words from the self-reference task versus valence task. However—likely due to reduced power—these effects are reduced relative to the sample reported in the main text (see Figure S2). These effects do not appear to interact: The recognition advantage for positive versus negative words appears similar for words from the self-reference and valence tasks.

These observations were largely confirmed by Bayesian mixed effects logistic regression models: $p(\text{"}old\text{"}) \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$. Participants are more likely to correctly recognize positive versus negative words from the self-reference task for Sample 2 (95% HDI: 0.189:0.564) but not Sample 1 (95% HDI: -0.052:0.511). Participants are also estimated to correctly recognize fewer negative words from the valence task than the self-reference task (95% HDIs: Sample 1, -0.629:-0.091; Sample 2, -0.461:-0.192), with no difference for positive words (95% HDIs: Sample 1, -0.369:0.399; Sample 2, -0.091:0.304).

HDDM. Threshold a, starting point bias z, and non-decision time t_0 were all fixed within subjects. Drift rate v was allowed to vary by valence and study status. A positive value indicated evidence accumulation towards an "old" response and a negative value indicated evidence accumulation towards a "new" response. Figure S12 presents the estimated mean drift rates by condition; note here that error bars represent 95% HDIs. As shown in the left panel, though numerically in the same direction as the effect

reported in the main text, evidence accumulated about equally efficiently towards an "old" response for old positive words (95% HDIs: Sample 1, 0.760:1.053; Sample 2, 0.718:0.919) compared to old negative words (95% HDIs: Sample 1, 0.569:0.862; Sample 2: 0.477:0.678). Similarly, the right panel shows that evidence accumulated equally efficiently towards a "new" response (i.e., was more positive) for new positive words (95% HDIs: Sample 1, -0.848:-0.558; Sample 2: -0.744:-0.541) compared to new negative words (95% HDIs: Sample 1, -0.807:-0.515; Sample 2, -0.833:-0.629). We attribute these differences from the main text to reduced power, as evidenced by the wider intervals.

Signal detection. Figure S13 presents the SDT parameters for memory strength (*d'*, left column) and bias (*c*, right column). As can be seen, participants exhibited greater memory strength for positive than negative words, with a conservative response bias across valences. Contrary to the sample reported in the main text, bias appears to be about the same for positive and negative words. Further, *d'* values were reduced in Samples 1 and 2 relative to the main text, potentially due to the larger number of trials.

Each parameter was analyzed with a Bayesian mixed effects regression model: $parameter \sim age + gender + valence + (1 \mid subject)$. Consistent with the observations above, positive words are associated with higher *d'* scores (95% HDIs: Sample 1, 0.037:0.270; Sample 2, 0.104:0.313) relative to negative words. There was no estimated difference in bias between positive and negative words for Sample 1 (95% HDI: -0.078:0.069), however participants in Sample 2 were estimated to have a stronger bias to respond "new" for negative versus positive words (95% HDI: -0.162:-0.022).

Source Accuracy. *Source accuracy for recognition hits.* As predicted, Figure S14 reveals greater source accuracy for positive than negative words from the self-reference task. By contrast, and consistent with the data reported in the main text (see Figure S6), source accuracy was higher for negative than positive words from the valence task. These observations were confirmed by Bayesian mixed effects logistic regression models: $p(correct) \sim age + gender + valence \cdot task \cdot BDI + (1 \mid subject) + (1 \mid word)$. Participants are more likely to correctly identify the source of positive versus negative words from the self-reference task (95% HDIs: Sample 1, 0.579:1.175; Sample 2, 0.731:1.076). Participants are

more likely to correctly identify the source of negative words from the valence task versus the self-reference task (95% HDIs: Sample 1, 0.250:0.819; Sample 2, 0.193:0.526), with a corresponding decrease for positive words (95% HDIs: Sample 1, -1.589:-0.793; Sample 2, -1.570:-1.100).

HDDM. As for the recognition model, threshold a, starting point bias z, and non-decision time t_0 were all fixed within subjects. Drift rate v was allowed to vary by valence and the source (encoding) task. Positive and negative values indicated evidence accumulation towards "self-reference" and "valence" responses, respectively. Figure S15 presents the estimated mean drift rates by condition; note here that error bars represent 95% HDIs. 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 (95% HDIs: Sample 1, 0.530:0.911; Sample 2: 0.563:0.866) relative to negative words (95% HDIs: Sample 1, -0.086:0.283; Sample 2, -0.049:0.244). For words from the valence task, evidence accumulated less efficiently (i.e., was less negative) for positive words (95% HDIs: Sample 1, -0.781:-0.403; Sample 2: -0.502:-0.210) versus negative words (95% HDIs: Sample 1, -0.825:-0.452; Sample 2: -0.834:-0.532), though there is some overlap between these values for Sample 1.

Signal detection. Figure S16 presents source discriminability between the self-reference and valence tasks (d', left panel) as well as bias towards a "self-reference" 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.

A Bayesian mixed effects regression model was used to analyze each parameter: $parameter \sim age + gender + valence + (1 \mid subject)$. As seen in Figure S16, positive words are associated with higher values of d' (95% HDIs: Sample 1, 0.222:0.605; Sample 2, 0.124:0.429) and lower values of c (95% HDIs: Sample 1, -0.498:-0.146; Sample 2, -0.569:-0.249).

References

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

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/

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, A., Cole, J. B., Warmenhoven, J., Ruiter, J. de, Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., … Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (September 2017). In *Zenodo* (p. https://zenodo.org/record/883859#.XwOHPyhKg2w). https://doi.org/10.5281/zenodo.883859

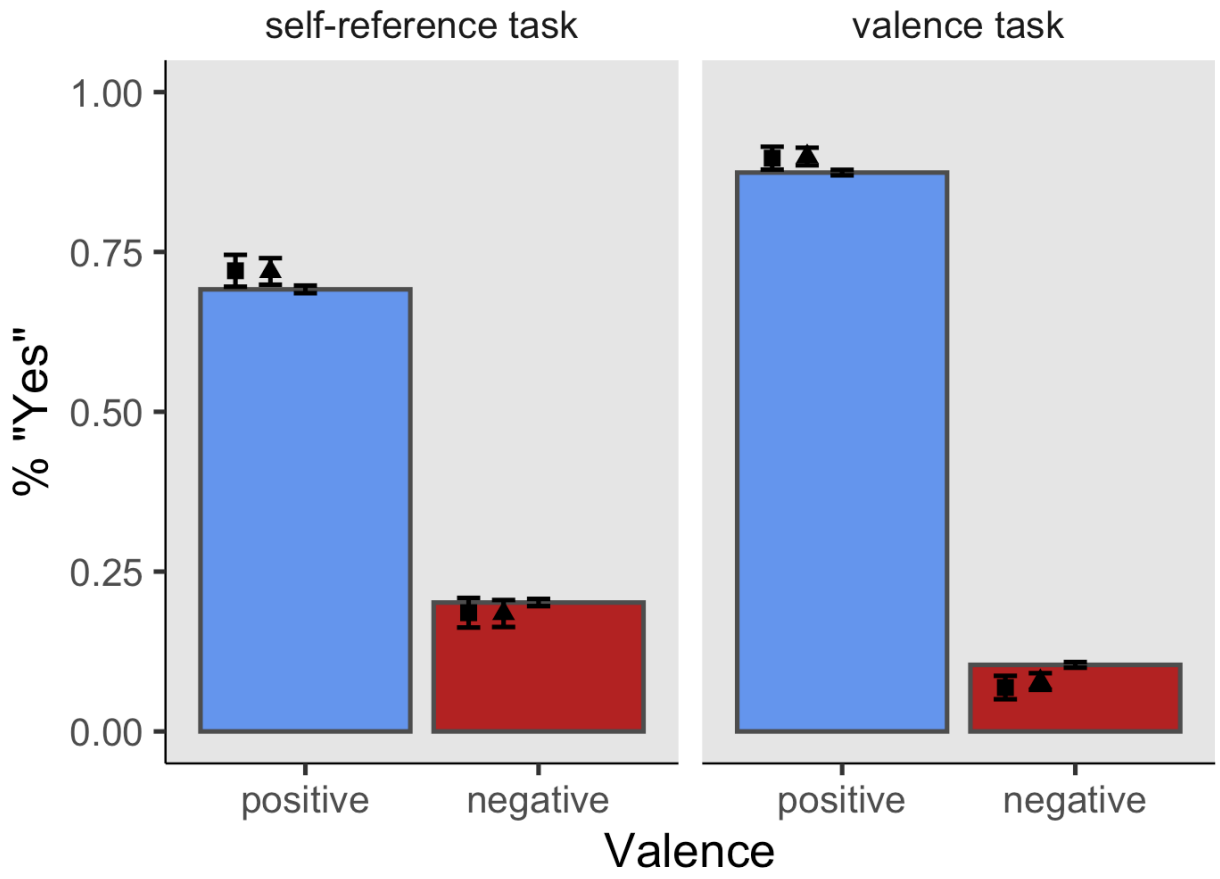

Figure S9. Proportion of "yes" responses at encoding for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% bootstrap confidence intervals (Waskom et al., 2017). The average proportion of "yes" responses is higher for positive vs. negative words in both encoding tasks.

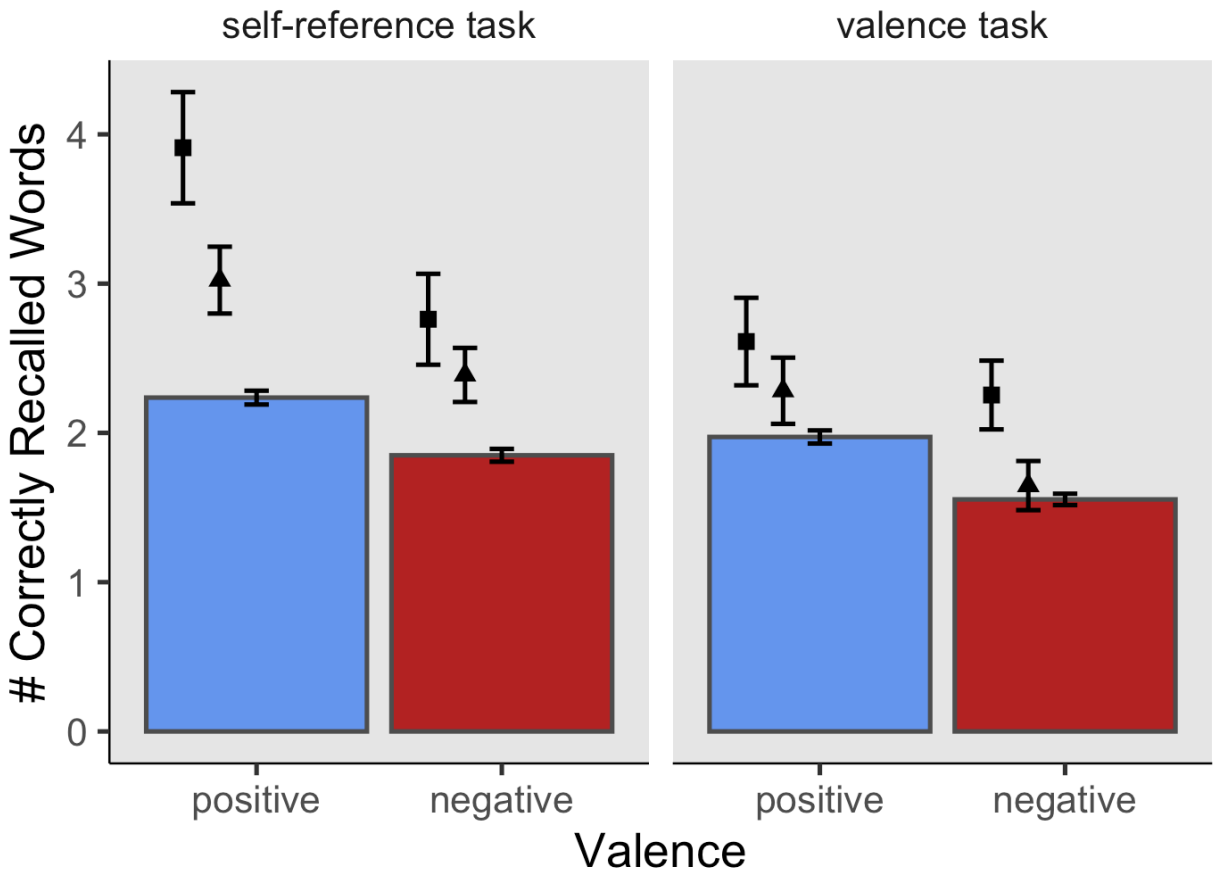

Figure S10. Number of correctly recalled words for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% bootstrap confidence intervals (Waskom et al., 2017). There is a higher average number of positive vs. negative words recalled from each encoding task, and a higher average number of words recalled from the self-reference task overall. The number of words recalled is lower for Sample 3 overall, likely due to the shorter study list length.

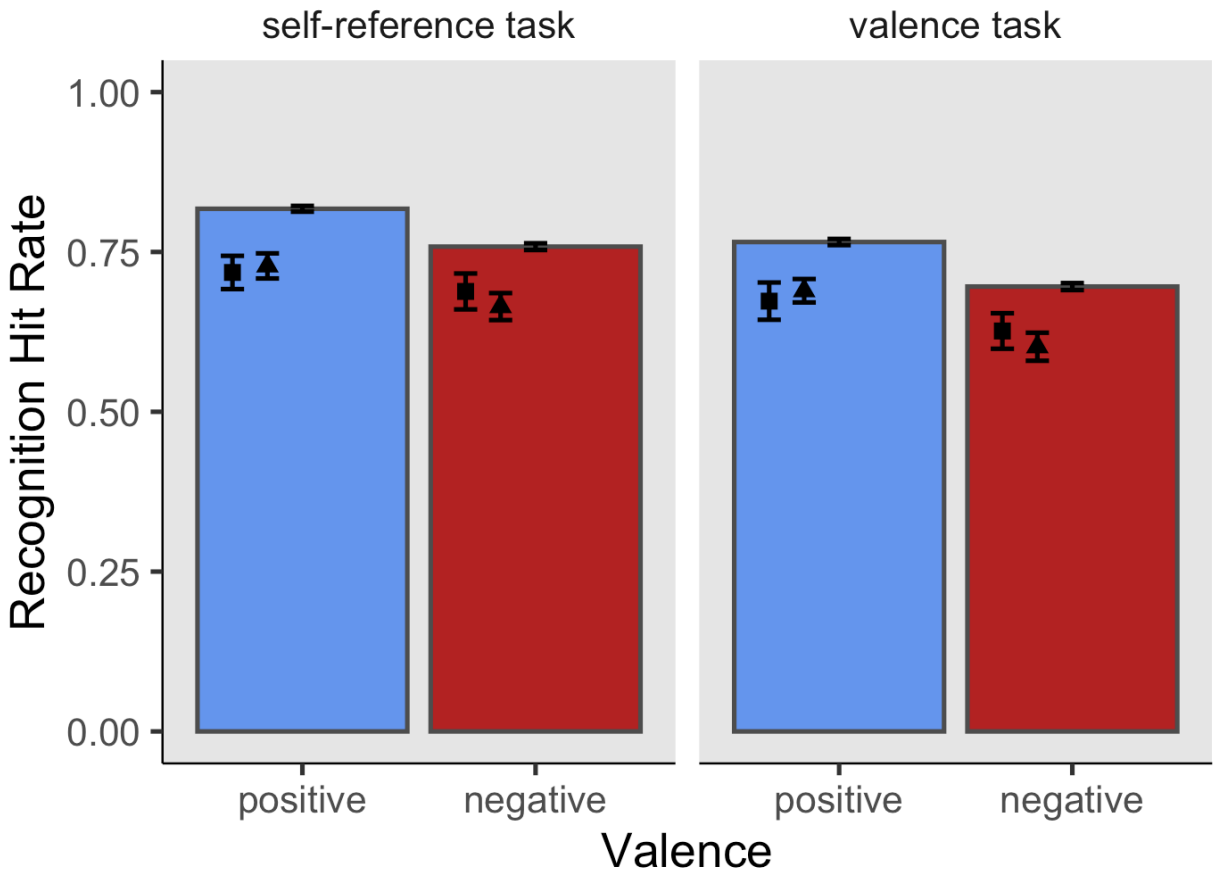

Figure S11. Hit rate at recognition for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% bootstrap confidence intervals (Waskom et al., 2017). The average hit rate is higher for positive vs. negative words, and for words from the self-reference vs. valence task.

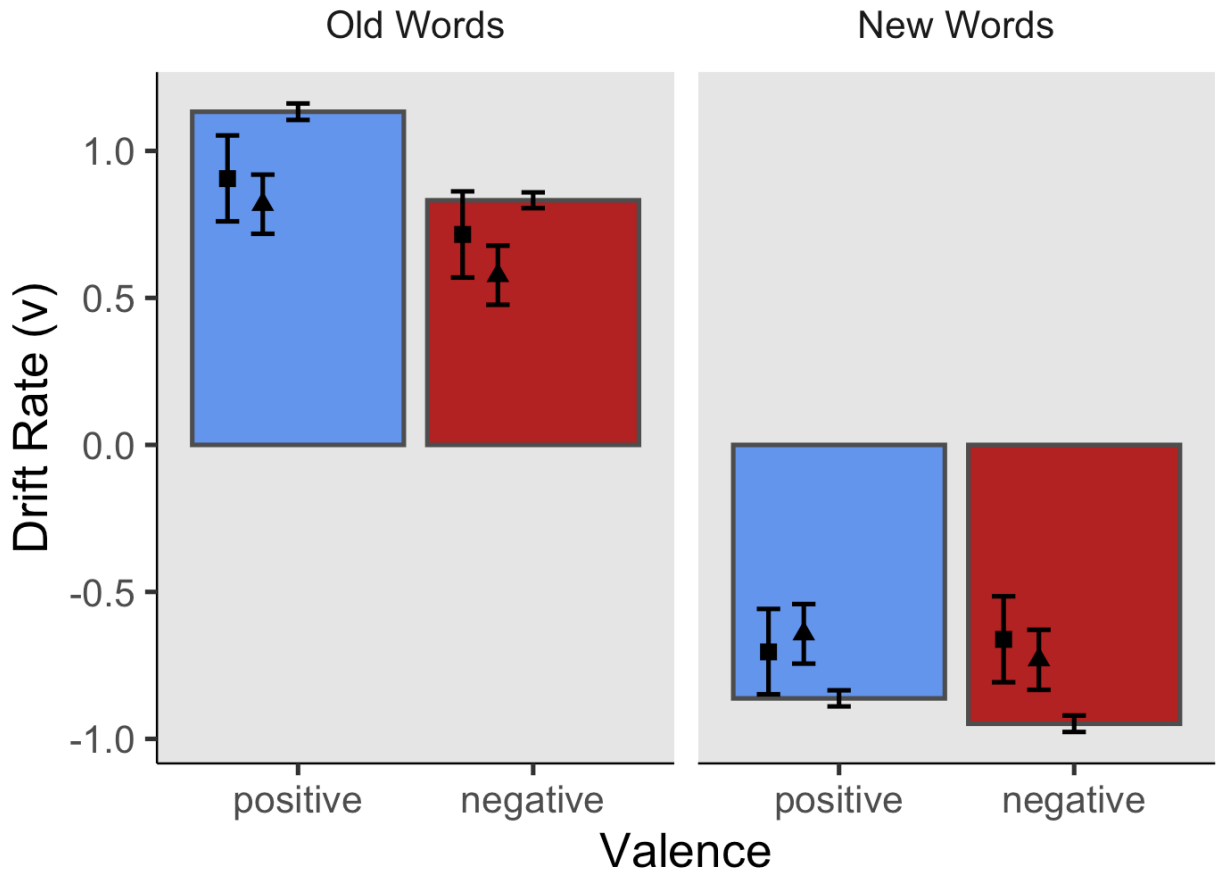

Figure S12. HDDM drift rate estimates for the recognition data for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote study status (left: old/studied words, right: new/unstudied words), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% HDIs. Evidence accumulates about as efficiently towards an "old" response for old positive vs. old negative words, but less efficiently towards a "new" response for new positive vs. new negative words.

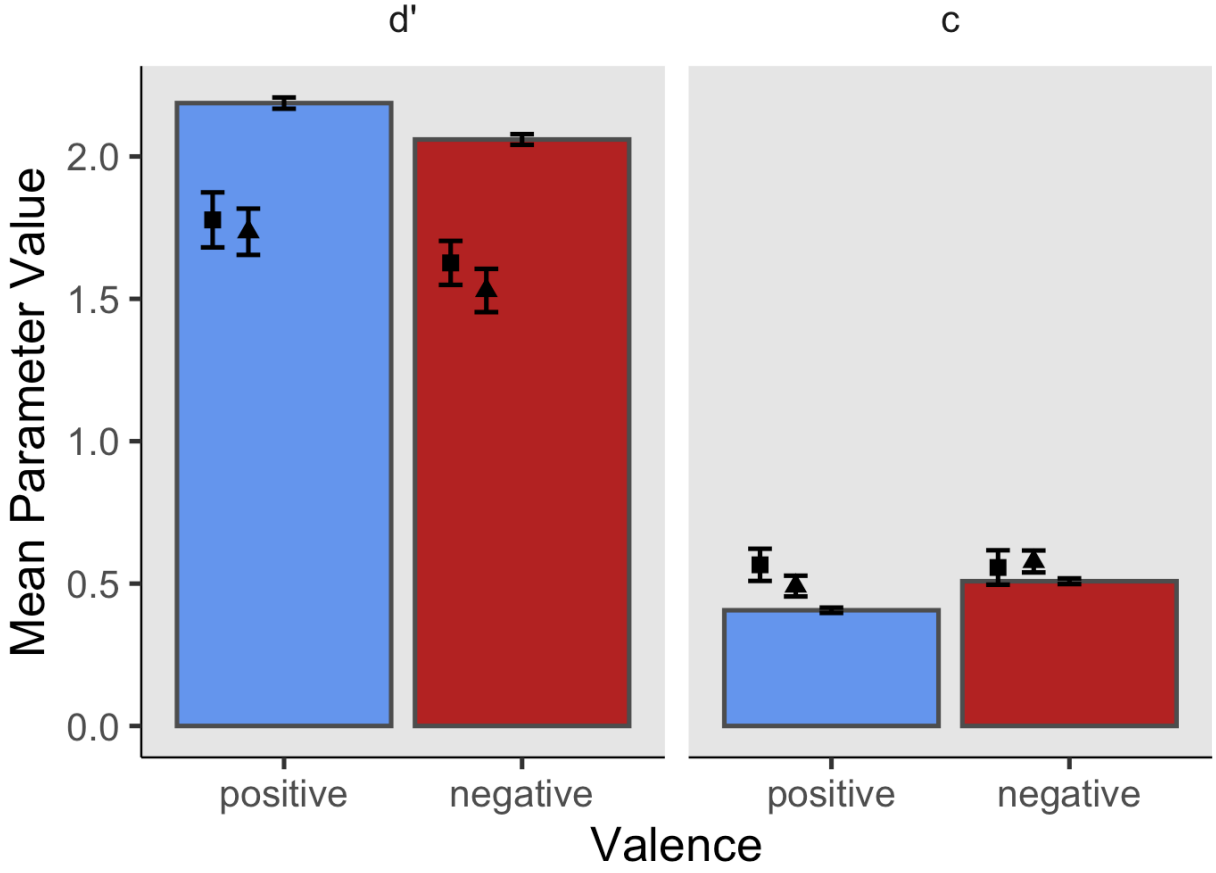

Figure S13. SDT parameter values for the recognition data for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote parameter (left: *d'*, right: *c*), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% bootstrap confidence intervals (Waskom et al., 2017). On average, values of *d'* are higher and values of *c* are lower (more liberal) for positive vs. negative words.

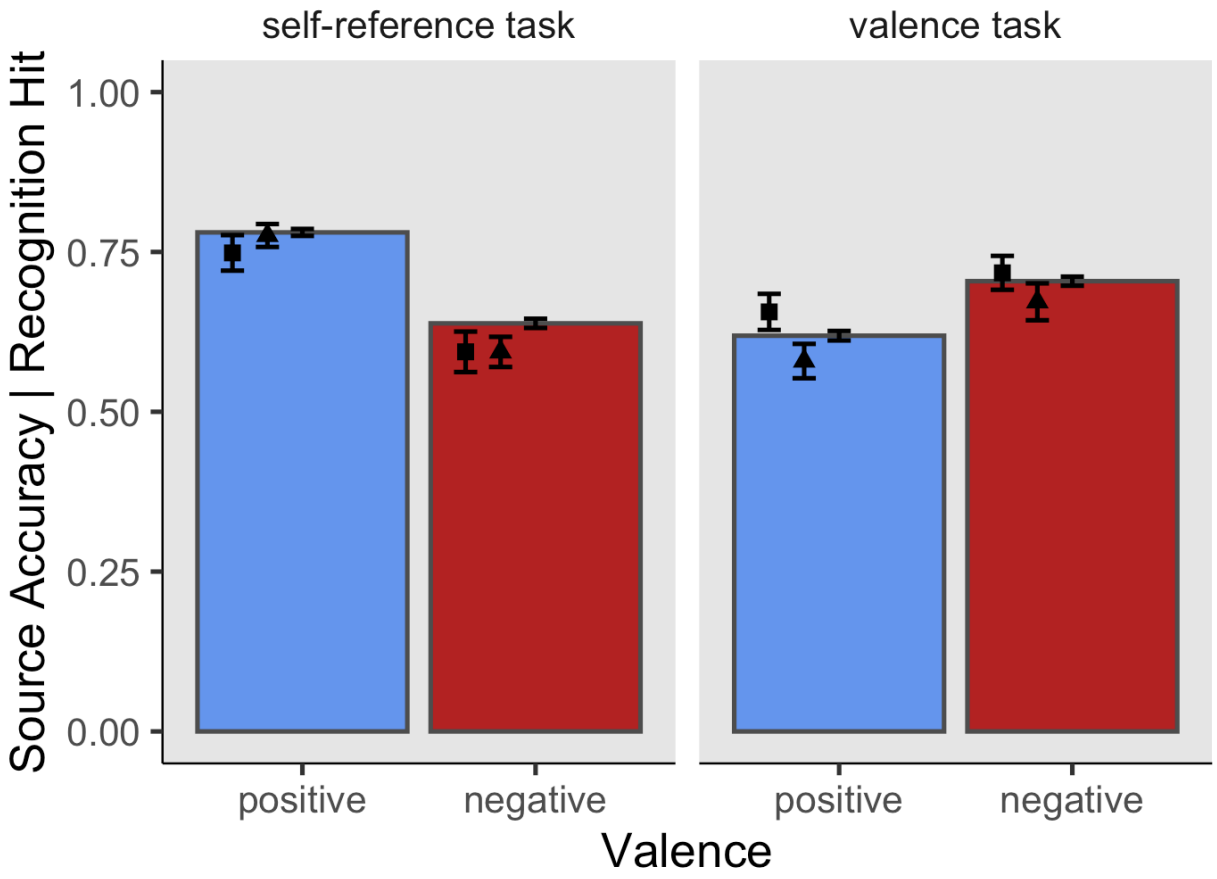

Figure S14. Source accuracy for recognition hits for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote encoding task (left: self-reference task, right: valence task), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% bootstrap confidence intervals (Waskom et al., 2017). The average source accuracy is higher for positive vs. negative words in the self-reference task, with the opposite effect in the valence task.

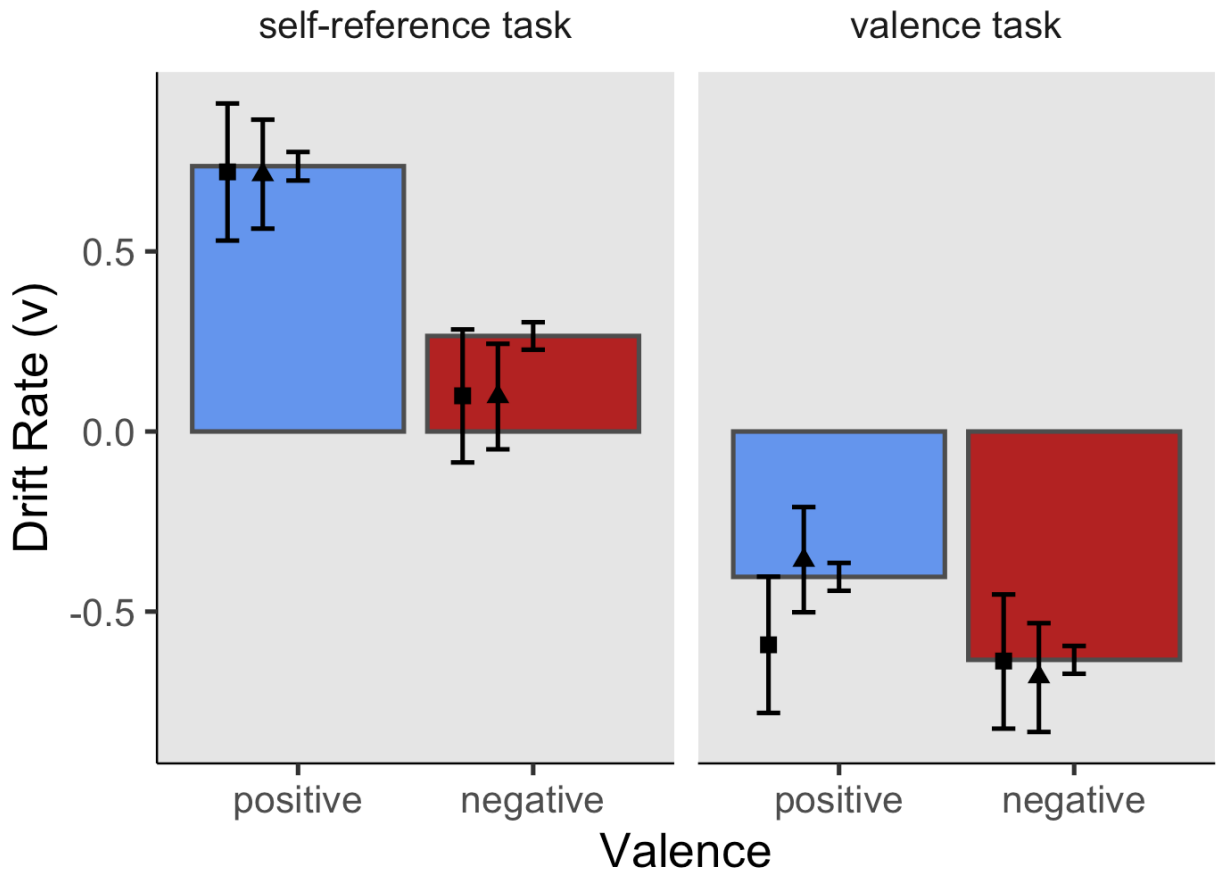

Figure S15. HDDM drift rate estimates for the source accuracy data for Sample 1 (square points), Sample 2 (triangle points), and the main text (bars). Columns denote study status (left: old/studied words, right: new/unstudied words), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% HDIs. Evidence accumulates more efficiently towards a "self-reference" response for positive vs. negative words from the self-reference task, but less efficiently towards a "valence" response for positive vs. negative words from the valence task.

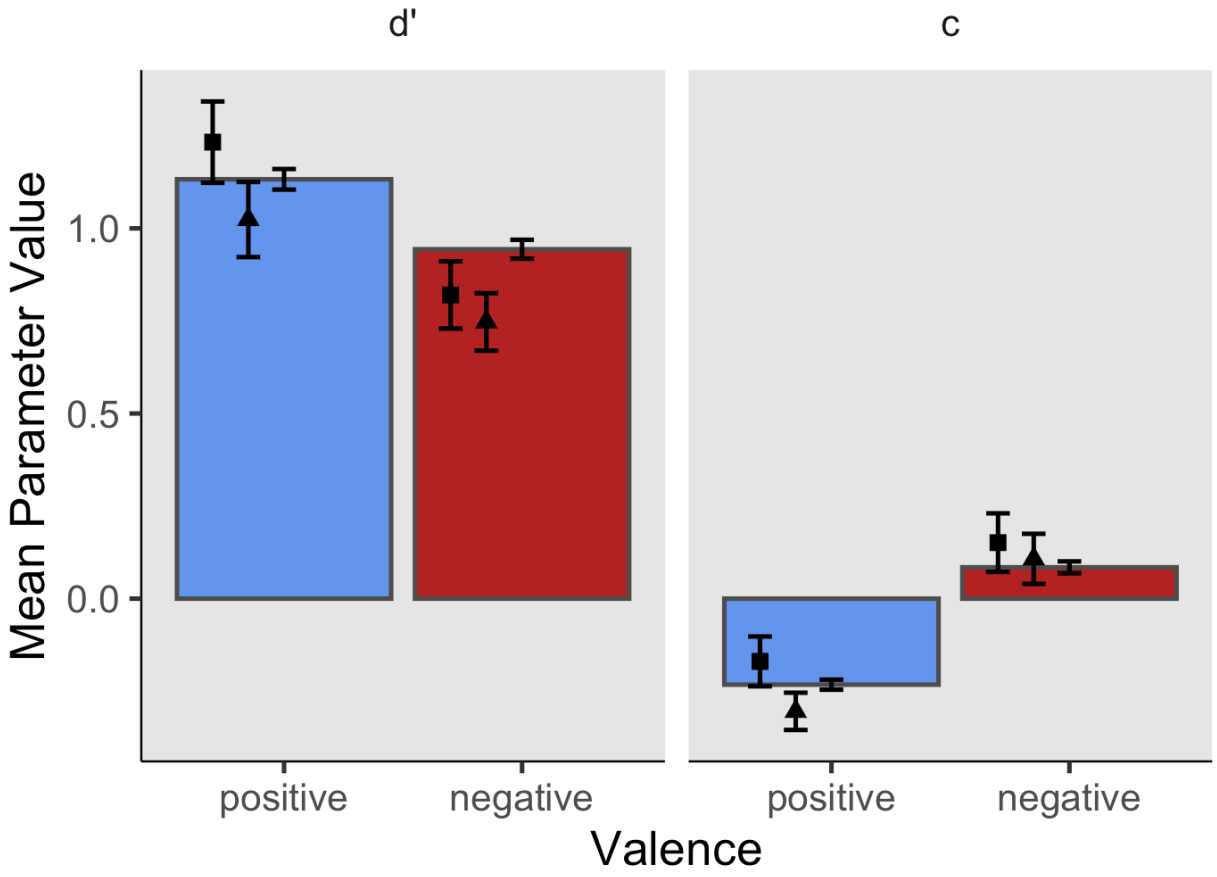

Figure S16. SDT parameter values for the source accuracy data for Sample 1 (squares), Sample 2 (triangles), and the main text (bars). Columns denote parameter (left: *d'*, right: *c*), and colors denote normative valence (blue: positive, red: negative). Error bars represent 95% bootstrap confidence intervals (Waskom et al., 2017). On average, values of *d'* are higher and values of *c* are lower (more liberal) for positive vs. negative words.