**Data Processing**

First, the recall portion of the study was coded as zero for incorrect responses, one for correct responses, and NA for participants who did not complete the recall section (all responses were blank). All word responses to judgment items were deleted and set to missing data. The final dataset was created by splitting the initial data file into six sections (one for each of the three experimental blocks and their corresponding recall scores). Each section was individually melted using the reshape package in *R* and was written as a csv file. The six output files were then combined to form the final dataset. Code is available at xx. With 112 participants, the dataset in long format included 7,056 rows of potential data (i.e., 112 \* 63 judgments). One incorrect judgment data point (> 100) was corrected to NA. Missing data for judgments or recall was then excluded from the analysis, which includes word responses to judgment items (i.e. responding with *cat* instead of a number). These items usually excluded a participant from receiving Amazon Mechanical Turk payment, but were included in the datasets found online. 787 data points were excluded (188 judgment only, 279 recall only, 320 both), leading to final *N* of 105 participants and 6,269 observations. Recall and judgment scores were screened for outliers using Mahalanobis distance at *p* < .001, and no outliers were found (T&F). To screen for multicollinearity, we examined correlations between judgment items, COS, LSA, and FSG. All correlations were *r*s < .50.

(descriptives added online)

**Results**

**Hypothesis 1**

Our first hypothesis sought to replicate bias and sensitivity findings from previous research while expanding the JAM function to include three types of memory. FSG, COS, and LSA were used to predict each type of judgment. Judgment values were divided by 100, so as to place them on the same scale as the database norms. Slopes and intercepts were then calculated for each participant’s ratings for each of the three judgment conditions, as long as they contained at least nine data points out of the 21 that were possible. Single sample *t*-tests were then conducted to test if slope and intercept values significantly differed from zero. See table xx for means and standard deviations. Slopes were then compared to the JAM function, which is characterized by high intercepts (between 40 and 60 on a 100 point scale) and shallow slopes (between 20 and 40). Because of the scaling of our data, to replicate this function, we should expect to find intercepts ranging from .40 to .60 and slopes in the range of 0.20. to 0.40. Intercepts for associative, semantic, and thematic judgments were each significant, and all fell within or near the expected range. Thematic judgments had the highest intercept at .67, while associative judgments had the lowest intercept at .51.

The JAM slope was successfully replicated for FSG in the associative judgment condition, with FSG significantly predicting association, although the slope was slightly higher than expected at 0.49. COS and LSA did not significantly predict association. For semantic judgments, each of the three database norms were significant predictors. However, JAM slopes were not replicated for this judgment type, as FSG had the highest slope at 0.12, followed by LSA, and then COS. These findings were mirrored for thematic judgments, as each database norm was a significant predictor, yet slopes for each predictor fell below range of the expected JAM slopes. Again, FSG had the highest slope, this time just out of range at 0.19, followed closely by LSA at 0.19. Interestingly, COS slopes were found to be negative for this judgment condition. Overall, although JAM slopes were not successfully replicated in each judgment type, the high intercepts and shallow slopes present in all three judgment conditions are still indicative of overconfidence and insensitivity in participant judgments.

Additionally, we examined the frequency that each predictor was the maximum strength for each judgment condition. For the associative condition, FSG was the strongest predictor for 64% of the participants, with COS and LSA being the strongest for only 16% and 20% of participants respectively. These differences were less distinct when examining the semantic and thematic judgment conditions. In the semantic condition, FSG was highest at 45% of participants, LSA was second at 32%, and COS was least likely at 24%. Finally, in the thematic condition, LSA was most likely to be the strongest predictor with 45% of participants, with FSG being the second most likely at 37%, and COS again being least likely at 19%. Interestingly, in all three conditions, COS was least likely to be the strongest predictor, even in the semantic judgment condition.

**Hypothesis 2**

As a result of the overlap between variables in Hypothesis 1, the goal of Hypothesis 2 was to test for an interaction between the three database norms when predicting participant judgment ratings. First, the database norms were mean centered. The nlme package and lme function were used to calculate these analyses (CITE). A maximum likelihood multilevel model was used to test the interaction between FSG, COS, and LSA when predicting judgment ratings while covarying for type of judgment, with participant number being used as the random intercept factor. Multilevel models were used to retain all data points (rather than averaging over items and conditions), while controlling for correlated error due to participants, as these models are advantageous for multiway repeated measures designs (Gelman). This analysis resulted in a significant three-way interaction between FSG, COS, and LSA (*β* = 3.324, *p* < .001), which is examined below in a simple slopes analysis. Table xx includes values for main effects, two-way, and three-way interactions.

To investigate this interaction, simple slopes were calculated for low, average, and high levels of COS. This variable was chosen for two reasons: first, it was found to be the weakest of the three predictors in hypothesis one, and second, manipulating COS would allow us to track changes across FSG and LSA. Significant two-way interactions were found between FSG and LSA at both low COS (*β* = -1.492, *p* < .001), average COS (*β* = -.569, *p* < .001), and high COS (*β* = 0.335, *p* = .013). A second level was then added to the analysis in which simple slopes were created for each level of LSA, allowing us to assess the effects of LSA at different levels of COS on FSG. When both COS and LSA were low, FSG significantly predicted judgment ratings (*β* = 0.663, *p* < .001). At low COS and average LSA, FSG decreased but still significantly predicted judgment ratings (*β* = 0.375, *p* < .001). However, when COS was low and LSA was high, FSG was not a significant predictor (*β* = .087, *p* = .079). This finding suggests that at low COS, LSA and FSG create a seesaw effect in which increasing levels of thematics is counterbalanced by decreasing importance of association when predicting recall. FSG was not a significant predictor when COS was high and LSA was low (*β* = .099, *p* = .088). At high COS and average LSA, FSG significantly predicted judgment ratings (*β* = .167, *p* < .001), and finally when both COS and LSA were high, FSG increased and was a significant predictor of judgment ratings (*β* = 0.236, *p* < .001). Thus, at high levels of COS, FSG and LSA are complimentary when predicting recall, increasing together as COS increases.

**Hypothesis 3**

Given the results of Hypothesis 2, we then sought to extend the analysis to participant recall scores. A multilevel logistic regression was used with the lme4 package and glmer() function (CITE), testing the interaction between FSG, COS, and LSA when predicting participant recall. As with the previous hypothesis, we covaried type of judgement and, additionally, covaried judgment ratings and used participants as the random intercept factor. MAIN EFFECTS OF COVARY (judgment value) STUFF. A significant three-way interaction was detected between FSG, COS, and LSA (*β* = 24.572, *p* < .001). See table xx for main effects, two-way, and three-way interaction values.

The moderation process from hypothesis 2 was then repeated, with simple slopes first calculated at low, average, and high levels of COS. This set of analyses resulted in significant two-way interactions between LSA and FSG at low COS (*β* = -7.845, *p* < .001) and high COS (*β* = 5.812, *p* = .010). No significant two-way interaction was found at average COS. Following the design of hypothesis two, simple slopes were then calculated for low, average, and high levels of LSA at the low and high levels of COS, allowing us to assess how FSG effects recall at varying levels of both COS and LSA. When both COS and LSA were low, FSG was a significant predictor of recall (*β* = 4.116, *p* < .001). At high COS and average LSA, FSG decreased as a predictor, though it was still significant (*β* = 2.601, *p* < .001). FSG was weakest as a predictor when COS was low but LSA was kept high (*β* = 1.086, *p* = .031). As with hypothesis 2, LSA and FSG counterbalanced one another, wherein the increasing levels of thematics led to a decrease in the importance of association in predicting recall. At high COS and low LSA, FSG was a significant predictor (*β* = 2.447, *p* = .002). When COS was high and LSA was average, FSG increased as a predictor and remained significant (*β* = 3.569, *p* < .001). This finding repeated when both COS and LSA were high, with FSG increasing as a predictor of recall (*β* = 4.692, *p* < .001). Therefore, at high levels of COS, LSA and FSG are complimentary predictors of recall, increasing together and extending the findings of hypothesis 2 to participant recall.

**Hypothesis 4**

In our fourth and final hypothesis, we investigated whether the judgment slopes and intercepts obtained in Hypothesis 1 would be predictive of recall ability. Whereas Hypothesis 3 indicated that word relatedness was directly related to recall performance, this hypothesis instead looked at whether or not participants’ sensitivity and bias to word relatedness could be used a predictor of recall (Maki). This analysis was conducted with a multilevel logistic regression, as described in hypothesis 3 where each database slope and intercept was used as predictors of recall, separated by judgment type. These values were obtained from Hypothesis 1 where each participant’s individual slopes and intercepts were calculated for associative, semantic, and thematic judgment conditions. Table xx shows average slopes and intercepts for recall for each of the three types of memory. In the associative condition, FSG slope significantly predicted recall (*β* = .898, *p* = .008), while COS slope and LSA slope were non-significant. In the semantic condition, COS slope and LSA slope were both found to be significant predictors of recall (*β* = 2.04, *p* < .001, *β* = .1.06, *p* = .019). FSG slope was non-significant in this condition. Finally, no predictors were significant in the thematic condition, though LSA slope was found to be the strongest (*β* = .896, *p* = .089).