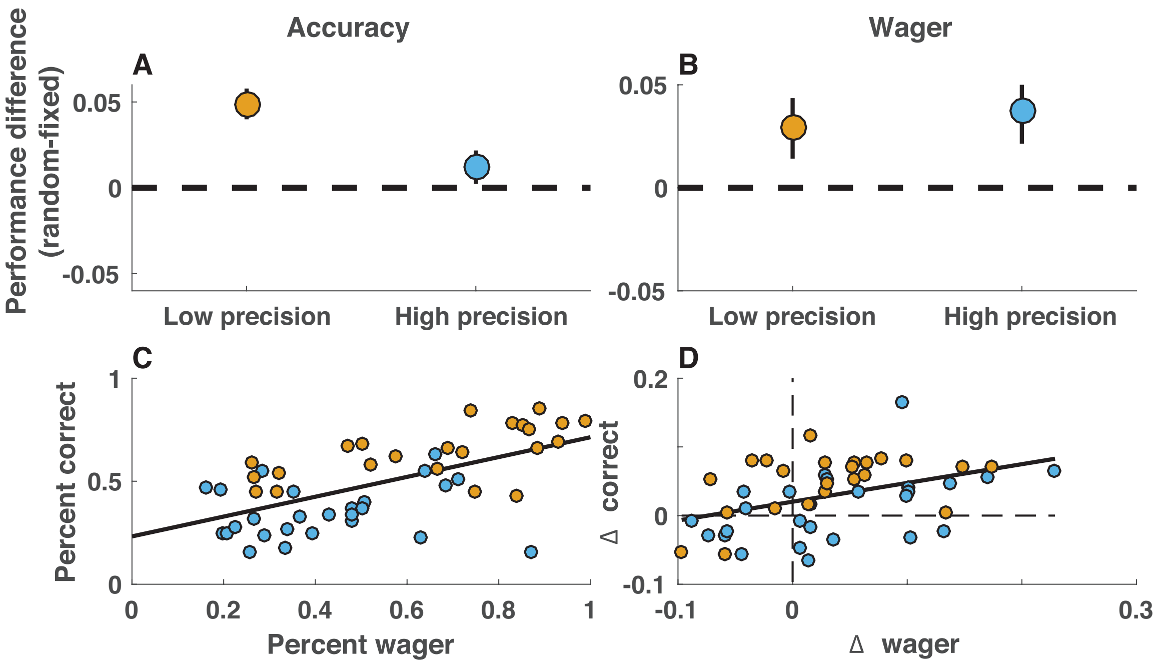

Figure S1: Uniformly spaced stimulus configurations degrade task performance and confidence in human subjects. Subject performance was assessed in terms of accuracy (percent of trials eliciting positive feedback) and confidence (percent of trials eliciting high post-decision wagers) separately according to precision condition (23 subjects were required to achieve an error of less than π/3 to elicit positive feedback [low precision], whereas 24 subjects were required to achieve an error of less than π/8 to elicit positive feedback [high precision]). **A)** Subjects in the low precision condition were more accurate for random spacing, as opposed to fixed spacing, stimulus configurations (orange; t=5.6, $p < 10e-4$), whereas subjects in the high precision condition attained similar overall performance in both configurations (blue; t=1.5, p = 0.15). Points/lines indicate group mean/SEM. **B)** Subjects in both conditions indicated higher confidence for random-spacing, as opposed to fixed-spacing, stimulus configurations (t = [2.3, 2.0] and p = [0.03, 0.06] for high and low precision conditions, respectively). **C)** Subjects that were most accurate, as assessed online according to a fixed error threshold, also tended to make higher post-decision wagers. Orange and blue points indicate subjects in low and high precision conditions, respectively. **D)** Furthermore, the improvement in accuracy from fixed- to random-spaced arrays was greater for subjects that showed the largest increase in confidence across the same conditions.

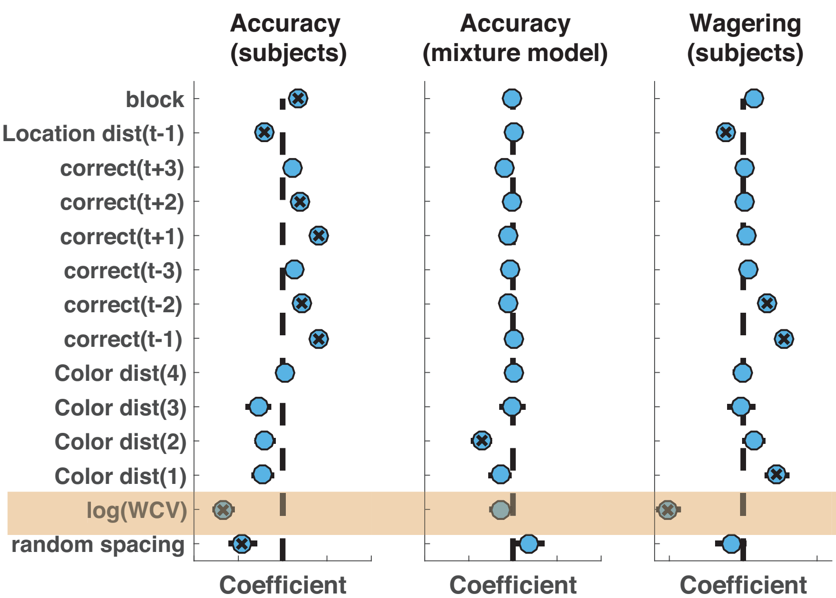

Figure S2: Stimulus clustering and recent feedback impact accuracy and confidence after controlling for other confounded factors in a GLM. Effects of color clustering on performance (left) and confidence (right) persist after accounting for potential confounding factors and feedback-dependent performance adjustments. Coefficients from a mixed-effects logistic regression model of binary accuracy and wager are plotted on the abscissa. Circles/lines reflect mean/SEM, and X marks indicate coefficients significantly different from zero ($p < 0.05$). Coefficients for log(WCV), a proxy for stimulus clustering, are highlighted in orange. Coefficients for log(WCV) are significantly lower than zero, indicating better performance / higher wagering on trials where stimulus colors were more clustered. This effect is not present in accuracy data simulated from a mixture model that includes binding errors (center).

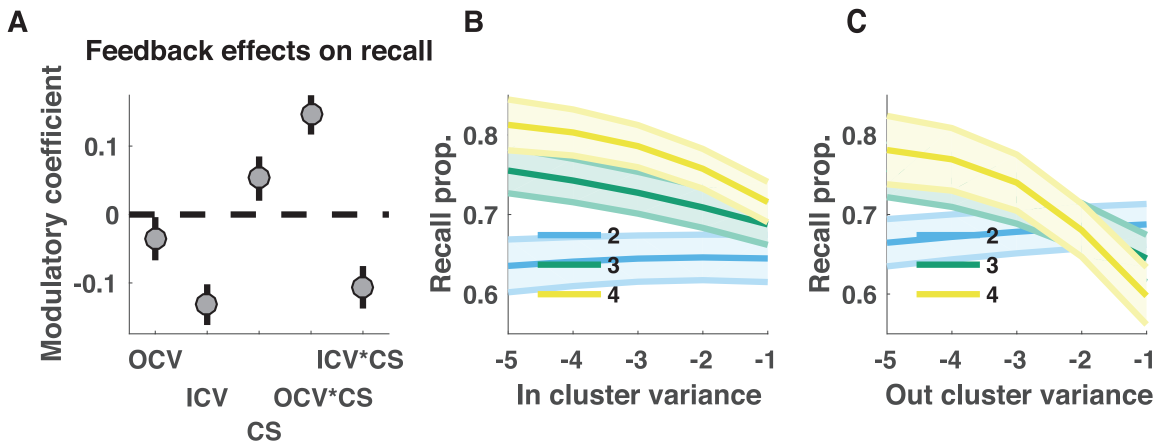

Figure S3: Recall is affected by clustering of both probed and non-probed stimuli. To further examine the source of the recall benefits, subject data were fit with a mixture model that considered reports to come from a mixture of processes including 1) a uniform "guess" distribution, 2) a "memory+binding" distribution centered on the color of the probed target, and 3) a "binding error" distribution including peaks at each non-probed target. Additional terms were included in the model to allow the recall probability to vary as a logistic function of various descriptive aspects of stimulus clustering that all factor into the within-cluster variance measurements reported in figure 5. To do so, the color array from each trial was divided into two (minimal variance) clusters to compute 1) the variance of the cluster that did not contain the probed item [OCV], 2) the variance of the cluster that did contain the probed item [ICV], and the number of colors in the cluster that contained the probed target [CS]. **A)** Mean/SEM coefficients across subjects indicated that these three factors, along with their interactions, were systematically related to trial-to-trial fluctuations in subject recall rates. **B&C)** The predicted recall rates from model fits are plotted as a function of ICV (B) and OCV (C) color coded according to the number of items in the relevant cluster (the cluster containing the probed item for B, and the cluster that did not contain the probed item for C). Recall bonuses are evident for low values of both ICV and OCV, although these benefits scale with the number of colors contained in the relevant cluster.

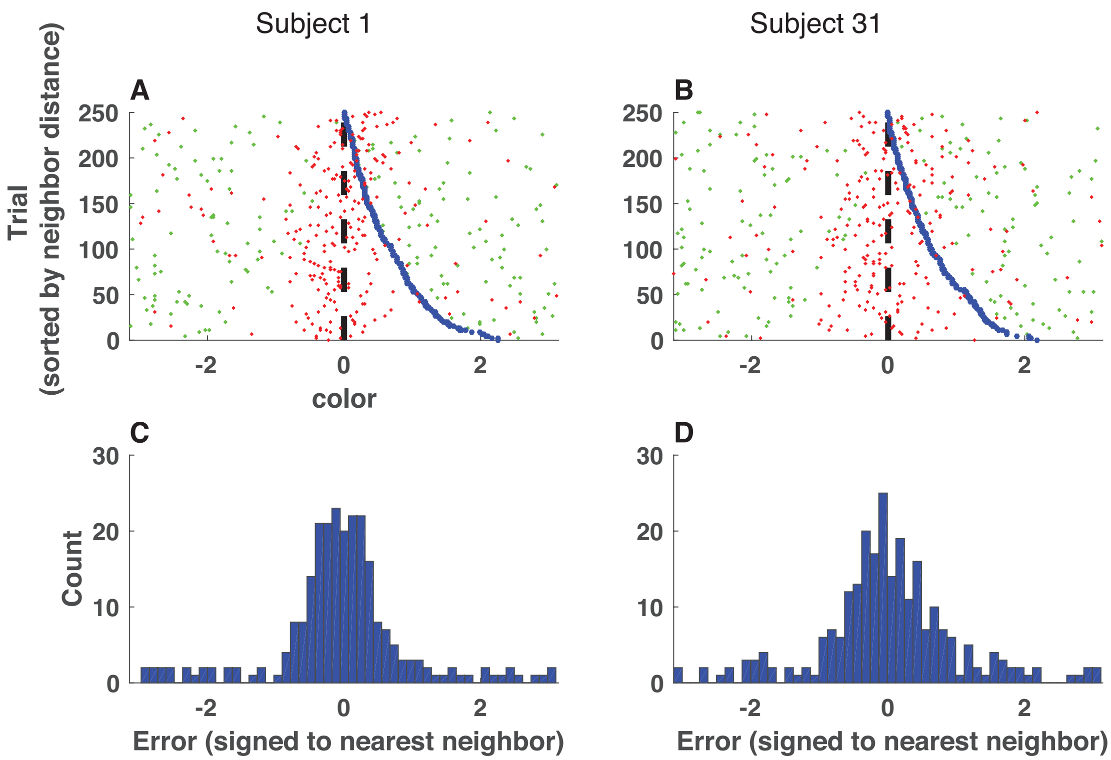

Figure S4: Sorting trials according to the nearest neighbor non-probed target color reveals structure in memory reports. A&B: Signed error of memory reports (red points) for all trials completed by two sample subjects (left = subject 1, right = subject 31). Trial errors are sorted by the distance from the probed target to the most similar color in the target array (nearest neighbor distance, NND) and transformed according to the direction of the nearest neighbor target (blue points). Green points reflect the positions of other colors in the target array, relative to the probed color and transformed as described above. Note the asymmetry in error distributions appears to change as a function of the nearest neighbor distance. **C&D:** Error histograms for the same two example subjects, transformed as described above. Note that in some cases apparent structure in the sorted errors (A) is no longer visible after collapsing across nearest neighbor distances (C).

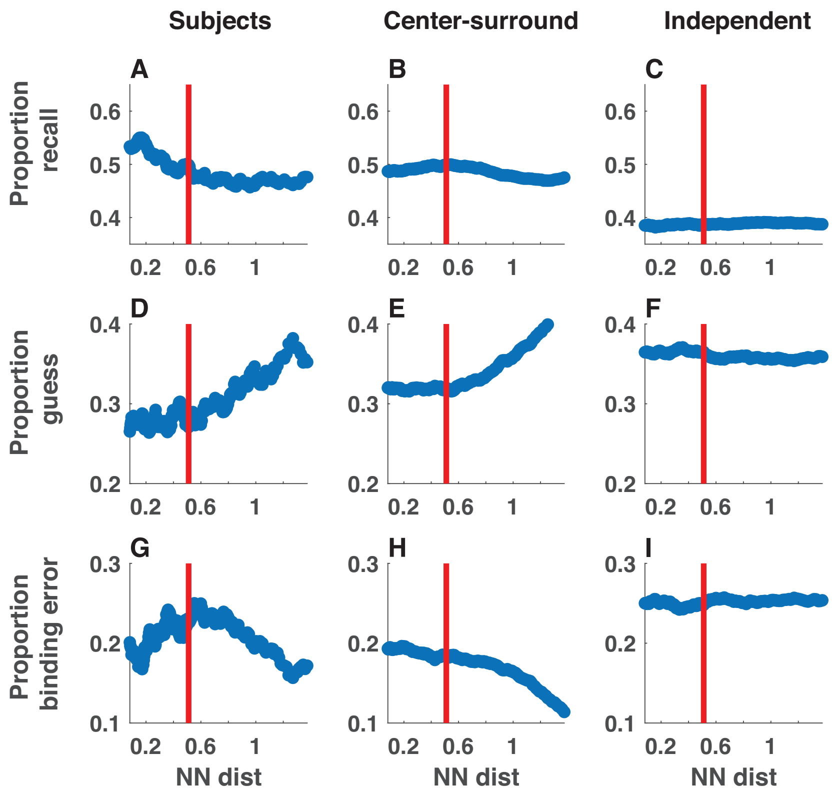

Figure S5: Neighboring stimulus features affect fits of mixture model. Subject (left) and simulated (center = center-surround, right = independent encoding) data were collapsed across all sessions and binned in sliding windows according to the absolute distance between the probed target color and the most similar non-probed target color (NN dist; abscissa). Data in each bin were fit with a mixture model that included free parameters to estimate the proportion of reports generated from 1) the von Mises "memory distribution" (**A-C**), 2) the uniform "guess distribution" (**D-F**), or 3)the mixture of von Mises "binding error distribution" (**G-I**). Parameter estimates for precision and bias terms are reported in the main text (figure 8).