Supplemental Results

All Experiments

Error by typicality: While we had strong predictions that category typicality would influence images' *direction* of error, it was less clear if category typicality would similarly affect their *magnitude* of error. For example, the use of category knowledge may have created small, systematic distortions for typical category members, but atypical category members may have exhibited the same extent of error in all directions. Here, we plot average error by typicality in the four experiments (Figure S1). We found that in Experiments 1 – 3, error was greater for typical category members relative to atypical category members in the experimental group, suggesting that the influence of category knowledge resulted in greater error in addition to systematic bias. Statistics are reported in the main text for Experiments 1, 2 and 4, and in the Supplemental Results specific to Experiment 3.

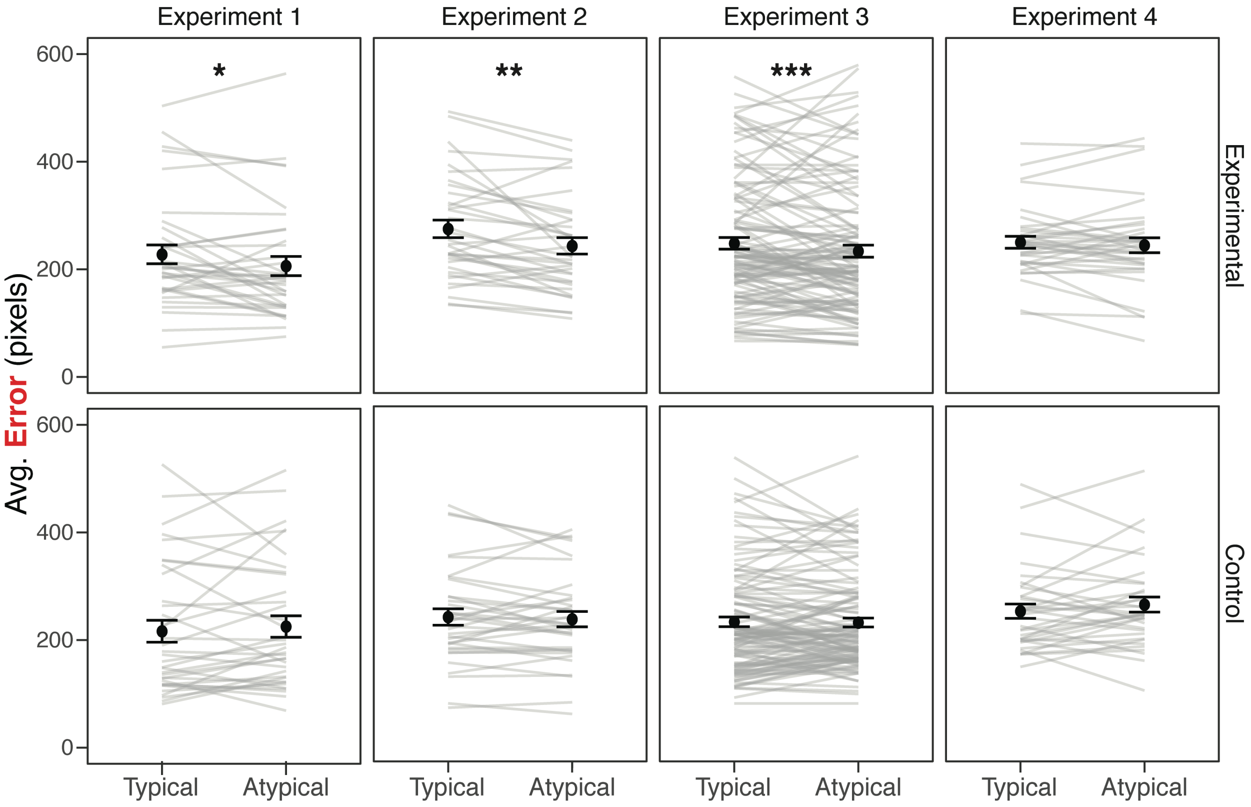


Figure S1. Average memory accuracy by category typicality in each experiment. Gray lines signify participants. Error bars indicate SEM. * $p < .05$; ** $p < .01$; *** $p < .001$

Error by spatial consistency and confidence: In the main text, we reported how confidence and spatial consistency with category membership interacted to influence error in Experiment 2, finding that the difference in error by spatial consistency was greatest when reported confidence was weakest. We aimed to replicate this effect in Experiments 3 and 4. See Figure S4 for a comparison of the effect sizes across the three experiments.

In Experiment 3, focusing on the experimental group, we entered spatial consistency, confidence, and their interaction into a mixed-effect model with error as the dependent measure (Figure S2B). This revealed main effects of consistency, $F_{(1,\ 149.87)}$ = 81.47, p < .001, and confidence, $F_{(2,\ 129.49)}$ = 99.17, p < .001, qualified by an interaction, $F_{(2,\ 1625.85)}$ = 3.32, p = .04. This interaction was driven by the same pattern of effects as Experiment 1, with less of a difference in error between spatially consistent and inconsistent images for increasingly confident responses (very confident: $t_{(708.66)}$ = -2.63, p = .009; somewhat confident: $t_{(325.63)}$ = -7.14, p < .001; guessed: $t_{(303.20)}$ = -7.77, p < .001; Bonferroni-corrected α = .017 for 3 tests). This replicates the effects observed in Experiment 2 and suggests that when memory for specific image locations was weaker, participants relied more on their knowledge of a category's likely location on the grid.

In Experiment 4, using the same model (Figure S2C), we found main effects of spatial consistency, $F_{(1,\ 50.52)}$ = 49.08, p < .001, and confidence, $F_{(3,\ 39.35)}$ = 8.63, p < .001, but no consistency by confidence interaction, $F_{(3,\ 1129.60)}$ = 1.77, p = .15. For consistency with Experiments 2 and 3, we computed pairwise tests of error by consistency for each level of reported confidence. This revealed less error for spatially consistent images relative to inconsistent images at all levels of confidence (all t < -2.84, all p < .005; Bonferroni-corrected α = .0125 for 4 tests). The absence of an interaction in Experiment 4 may be driven by the change in stimuli for this experiment, which may have changed the way participants used the confidence responses. For example, visual inspection suggests that the main effect of confidence is much stronger in Experiment 2 and 3 relative to Experiment 4, as the spread of error across levels of confidence is much more narrow in Experiment 4, possibly signifying that confidence is a less informative marker of memory strength in this experiment.

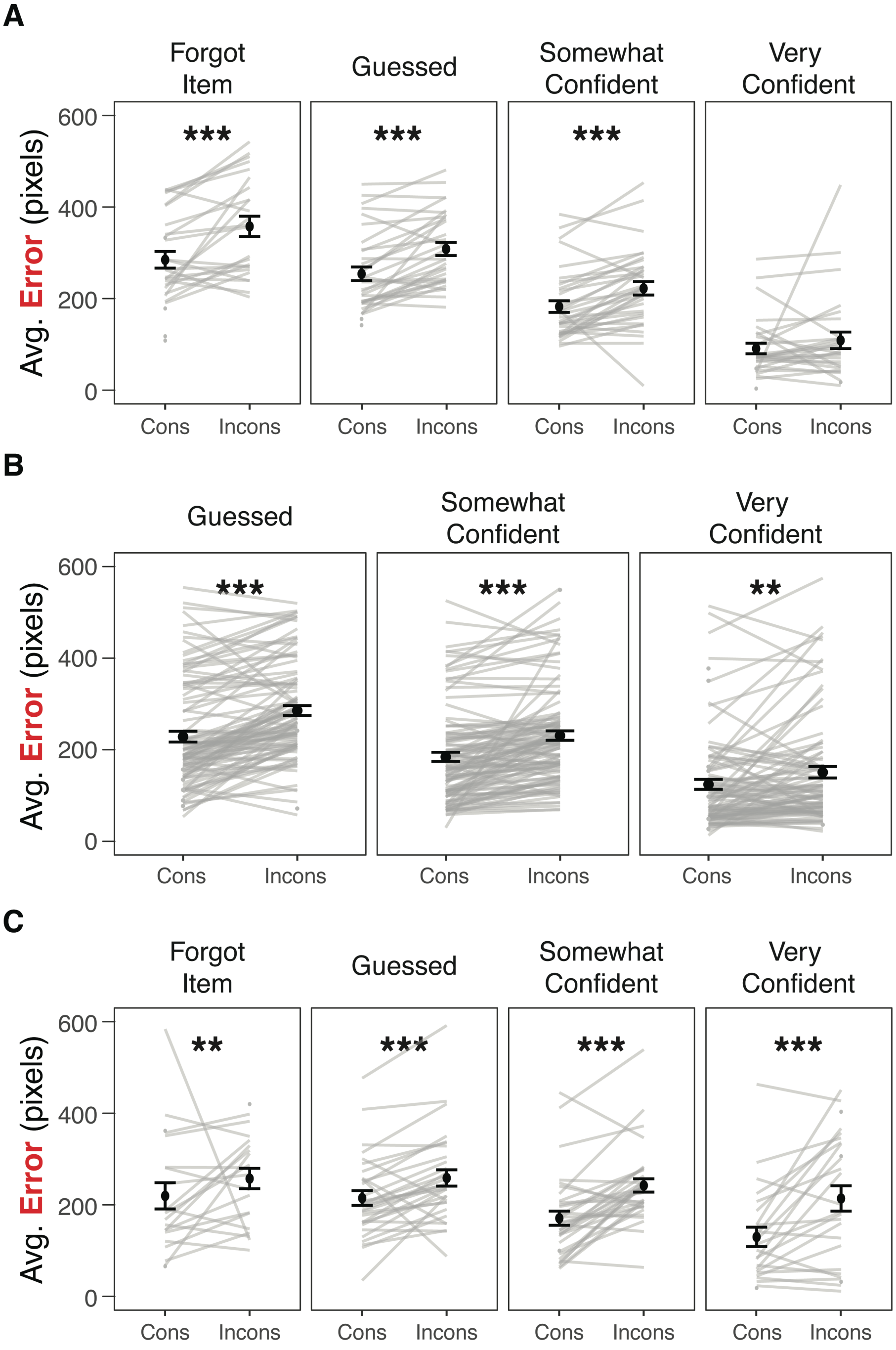


Figure S2. Error by consistency and confidence. Average error by confidence and consistency in Experiment 2 **(A)**, Experiment 3 **(B)**, and Experiment 4 **(C)**. The data from Experiment 2 (replotted from Figure 3D) is included above for comparison with the other experiments. Gray lines signify participants. Gray dots signify participants with no responses in the other bin (e.g. a dot in Forgot Item for consistent images indicates that the participant did not respond 'Forgot Item' for any inconsistent images). Error bars indicate SEM. Statistics reflect simple effects that survive Bonferroni correction for the number of tests conducted within each experiment (Experiments 2 & 4: α = .0125, Experiment 3: α = .0167). ** *p* < .01; *** *p* < .001

Bias by typicality and confidence: We also sought to replicate the influence of confidence on bias observed in Experiment 2, which revealed a main effect of confidence such that there was less bias for the most confident responses. See Figure S4 for a comparison of the effect sizes across the three experiments.

In Experiments 3 and 4, we computed separate mixed effects models with confidence and typicality with bias as the dependent variable, again focusing only on the experimental group. In Experiment 3 (Figure S3B), this model revealed a main effect of confidence, $F_{(2,\ 98.64)} = 3.80$, $p = .03$, and a main effect of category typicality, $F_{(1,\ 1921.76)} = 21.85$, $p < .001$. There was no reliable interaction, $F_{(2,\ 2908.71)} = 2.21$, $p = .11$. As in Experiment 2, the main effect of confidence was driven by less bias for 'very confident' responses relative to 'somewhat confident' responses, $t_{(78.79)} = -2.47$, $p = .016$, and relative to 'guessed' responses, $t_{(101.64)} = -2.52$, $p = .013$, both of which survive correction for multiple corrections (3 tests, $\alpha = .0167$). There was no difference in bias between 'somewhat confident' and 'guessed' responses, $t_{(91.66)} = -0.67$, $p = .50$. The main effect of typicality was driven by greater bias for typical category members relative to atypical category members for 'very confident' responses, $t_{(622.79)} = 3.27$, $p = .001$, and for 'somewhat confident' responses, $t_{(329.80)} = 3.56$, $p < .001$, but not for 'guessed' responses, $t_{(279.10)} = 1.18$, $p = .24$ (3 tests, $\alpha = 0.167$). In Experiment 4 (Figure S3C), the same group x typicality ANOVA revealed no main effects or interaction, all $F < 1.85$, all $p > .15$.

Retrieval in Experiments 2 and 3 was modulated by reported confidence such that stronger memories, as indexed by high confidence, were less biased towards their category centers. This suggests that retrieval of weaker memories relied more on knowledge of each category's general location and thus memory for their locations was more distorted. Furthermore, the main effect of category typicality in Experiment 3 suggests that confidence and category typicality may separately and independently bias retrieval, although this was not observed in Experiment 2. Interestingly, in Experiment 4, bias was not affected by confidence at all, in contrast to observations that higher confidence was related to less error. This may be explained by the possibility that confidence was not as strong as an index of memory as it was in Experiments 1 – 3.

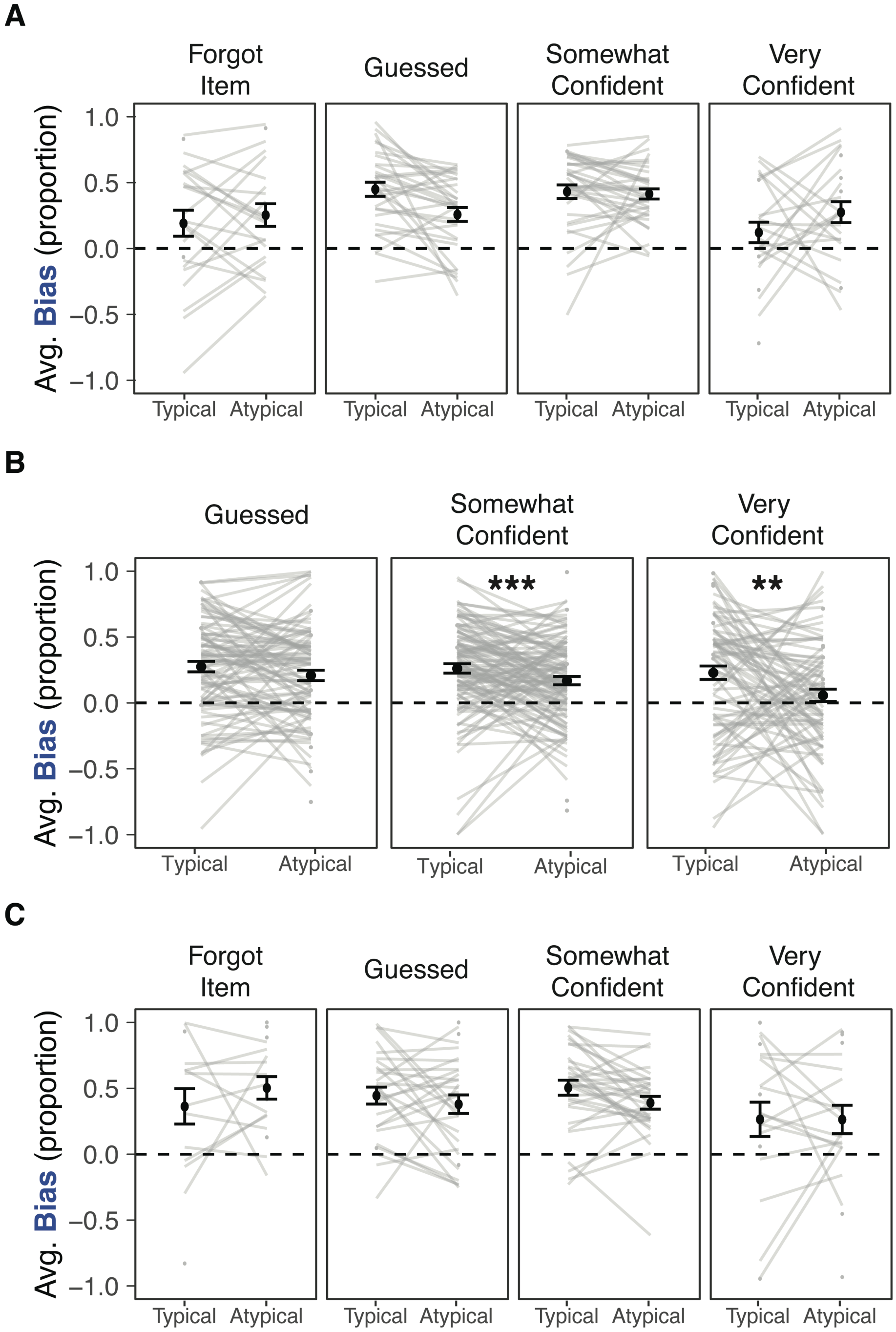


Figure S3. Bias by typicality and confidence. Average bias (as a proportion of error) by confidence and typicality in Experiment 2 **(A)**, Experiment 3 **(B)**, and Experiment 4 **(C)**. Results from Experiment 2 (re-plotted from Figure 3B) are included here for comparison with other experiments. Gray lines signify participants. Gray dots signify participants with no responses in the other bin (e.g. a dot in Forgot Item for typical category members indicates that the participant did not respond 'Forgot Item' for any atypical category members). Error bars indicate SEM. Statistics reflect simple effects that survive Bonferroni correction for the number of tests conducted within each experiment (Experiments 2 & 4: α = .0125, Experiment 3: α = .0167). ** *p* < .01; *** *p* < .001

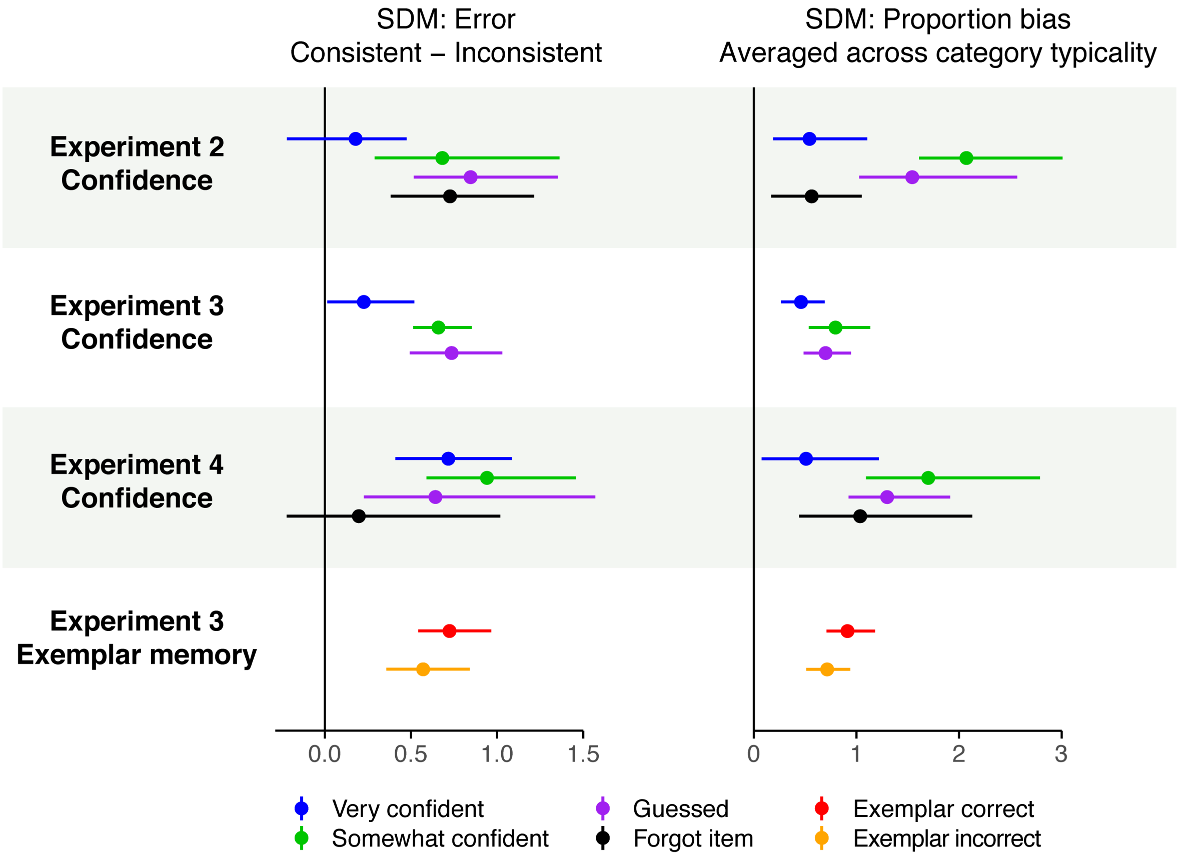


Figure S4. Effect sizes of error and bias across experiments, by confidence and exemplar memory. **(A)** Standardized difference in mean error for spatially consistent versus inconsistent images. 0 on the x-axis reflects no difference in error by consistency. **(B)** Standardized difference in the mean bias relative to 0, averaged across typical and atypical images. 0 on the x-axis reflects no reliable bias towards category clusters. Points represent mean effect size. Lines represent bootstrapped 95% confidence intervals.

Expt	Condition	Forgot Item	Guessed Location	Somewhat Confident of Location	Very Confident of Location
2	Consistent	10.5 (11.3)	34.9 (21.3)	41.8 (20.0)	12.8 (15.0)
	Inconsistent	10.1 (10.1)	35.3 (22.6)	41.8 (21.8)	12.9 (14.4)
3	Consistent		35.1 (30.1)	45.5 (24.1)	19.4 (19.9)
	Inconsistent		36.6 (31.0)	43.9 (24.4)	19.4 (20.5)
4	Consistent	9.3 (14.6)	30.4 (24.5)	49.1 (24.5)	11.2 (14.9)
	Inconsistent	10.1 (17.0)	30.7 (24.3)	48.5 (25.4)	10.7 (12.2)

Table S1. Mean (SD) percentage of all encoded images, by spatial consistency and confidence.

Expt	Condition	Forgot Item	Guessed Location	Somewhat Confident of Location	Very Confident of Location
2	Typical	10.4 (11.2)	36.4 (24.9)	42.1 (24.1)	11.1 (13.1)
	Atypical	9.8 (10.4)	34.2 (22.4)	41.4 (21.9)	14.6 (16.6)
3	Typical		38.0 (32.5)	44.1 (26.1)	17.9 (19.3)
	Atypical		35.3 (31.3)	43.9 (25.9)	20.8 (23.1)
4	Typical	8.9 (16.0)	32.7 (28.8)	48.3 (28.4)	10.1 (12.1)
	Atypical	11.3 (18.9)	28.5 (22.7)	48.7 (24.8)	11.4 (13.6)

Table S2. Mean (SD) percentage of all spatially inconsistent images, by category typicality and confidence.

Swap errors: One open question is whether the observed distortions in memory are due to small but systematic changes in retrieval, or large errors where participants mistook an image for a different one – a swap error. Although there is no clear way to identify swaps in these experiments, one way to adjudicate between these possibilities is by approximating the profile of a swap error by identifying retrieved locations that were closer to their category cluster. Since most members of a category were clustered in the same area, a swap error would look like a very large bias towards its category cluster. While this profile cannot directly pinpoint swap errors, decomposing retrieval in this way may provide a course hint at possible differences in biases.

To quantify this, we drew a line between each image's encoded location and cluster center, and we projected its retrieved location onto the line. We then sorted the projected distances of all images into four bins, separated by three landmarks along the line: the image's encoded location, its category's cluster center, and the halfway point between the two (Figure S5). The resulting four bins were: (1) 'Biased away', or images whose projected retrieval was father from their cluster center relative to their encoded locations, (2) 'Closer to Encoded', or images whose retrieval was biased towards their cluster center, but was retrieved closer to its encoded location than its cluster center, (3) 'Closer to Cluster', or images whose retrieval was closer to their cluster center than their encoded location, and (4) 'Beyond Cluster', for images retrieved in locations extending beyond their cluster center. The proportions of trials in these bins were entered into ANOVAs separately for each experiment. Planned paired t-tests of the difference by typicality were performed separately for each bin. Significance for these tests was Bonferroni-corrected for four comparisons ($\alpha = .0125$).

In Experiment 1, a 4 (bin) x 2 (typicality: typical, atypical) ANOVA revealed a main effect of bin, $F_{(3,102)} = 15.93$, $p < .001$, $\eta_p^2 = 0.50$, no main effect of typicality, $F_{(1, 34)} = 0.47$, $p = .50$, $\eta_p^2 = 0.00$, and a reliable interaction, $F_{(3, 102)} = 2.93$, $p = .04$, $\eta_p^2 = 0.08$. This interaction was driven by more atypical category members biased away from their cluster center relative to typical ones, $t_{(34)} = -3.04$, $p = .004$, $d = -0.51$. No other comparisons reached significance, $p > 0.08$.

In Experiment 2, the same ANOVA revealed similar effects: a main effect of bin, $F_{(3,102)} = 37.12$, $p < .001$, $\eta_p^2 = 0.77$, no reliable main effect of typicality, $F_{(1, 34)} = 0.19$, p

= .67, $\eta_p^2 = 0.00$, and a reliable interaction, $F_{(3,\ 102)} = 10.54$, $p < .001$, $\eta_p^2 = 0.24$. Here, there were more atypical trials than typical trials that were biased away from their cluster center, $t_{(34)} = -3.65$, $p < .001$, $d = -0.62$, and biased towards the cluster center but still retrieved closer to their encoded location, $t_{(34)} = -3.05$, $p = .004$, $d = -0.52$. However, more typical trials were retrieved closer to their cluster center, $t_{(34)} = 3.27$, $p = .002$, $d = 0.55$, and beyond their cluster center, $t_{(34)} = 3.40$, $p = .002$, $d = 0.58$.

In Experiment 3, we again found a main effect of bin, $F_{(3,342)} = 35.62$, $p < .001$, $\eta_p^2 = 0.43$, no reliable main effect of typicality, $F_{(1,\ 114)} = 0.01$, $p = .92$, $\eta_p^2 = 0.00$, and a reliable interaction, $F_{(3,\ 342)} = 5.21$, $p = .002$, $\eta_p^2 = 0.04$. Here, more atypical trials were biased away from their cluster center, $t_{(114)} = -3.82$, $p < .001$, $d = -0.36$, and more typical trials were retrieved beyond their cluster center $t_{(114)} = 3.11$, $p = .002$, $d = 0.29$. No other comparisons reached significance, all $p > 0.28$.

We found a different pattern in Experiment 4. The same ANOVA revealed a main effect of bin, $F_{(3,102)} = 15.59$, $p < .001$, $\eta_p^2 = 0.54$, no main effect of typicality, $F_{(1,\ 34)} = 0.51$, $p = .49$, $\eta_p^2 = 0.00$, and no interaction, $F_{(3,\ 102)} = 0.22$, $p = .88$, $\eta_p^2 = 0.006$. Unlike in Experiments 1 – 3, more trials were retrieved closer to their cluster center than to their encoded locations.

Taken together, these results suggest that a mixture of different errors drives the observed biases in retrieval. It seems likely that many of participants' memories for both typical and atypical category members can be characterized as smaller distortions rather than larger swap errors, as the plurality of trials were biased towards their cluster center but still remained closer to their encoded locations. However, the large proportion of trials retrieved closer to or beyond their category cluster also highlights the possibility that swap errors were also common. When considering typicality, we find that typical category members are more likely to be retrieved closer to or beyond their cluster center relative to atypical ones in Experiments 2 and 3 (and in the same direction in Experiment 1), suggesting that swap errors may contribute more to retrieval of typical category members relative to atypical category members. Finally, memory in Experiment 4 may have been more prone to swap errors, likely due to visually similar exemplars of the six objects that were encoded, rather than visually distinct members of six categories.

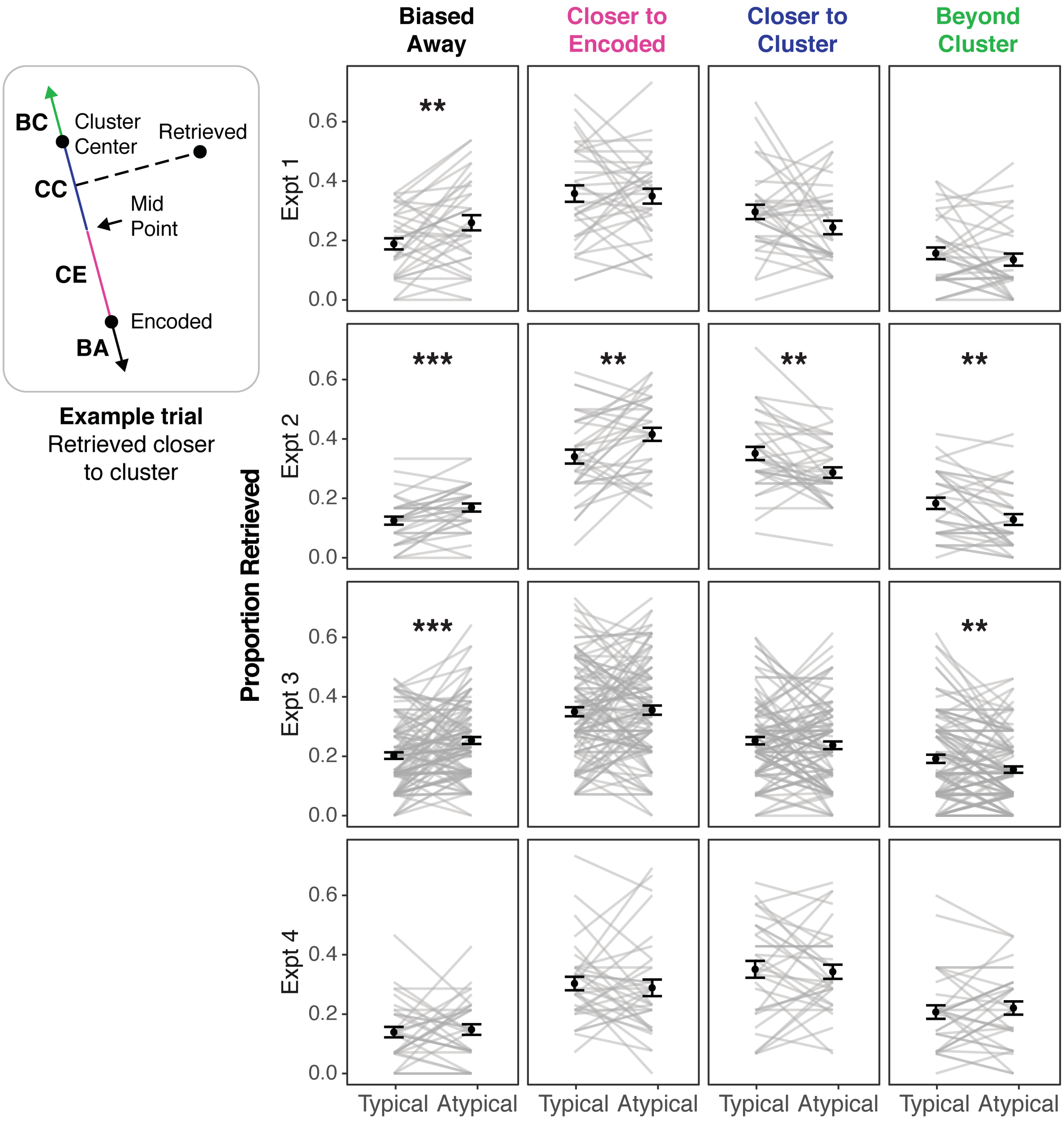


Figure S5. **Swap error analyses.** Proportion of images retrieved, relative to encoded locations and cluster centers. 'Biased Away' indicates the proportion of trials retrieved farther from the cluster center than what was encoded. 'Closer to Encoded' indicates the proportion of retrieved closer to their encoded location than to their cluster center. 'Closer to Cluster' indicates the proportion retrieved closer to their cluster center than their encoded locations. 'Beyond Cluster' indicates the proportion retrieved beyond their category cluster relative to their encoded location. Gray lines signify participants. Error bars indicate SEM. Statistics reflect two-tailed paired-sample t-tests (α = .0125. Bonferroni-corrected for 4 comparisons per experiment). ** *p* < .01 *** *p* < .001

Interactions with superordinate category: Experiments 1 – 3 comprised animal and object images. The purpose of this was to increase power by doubling the number of image locations tested as a function of spatial consistency and category typicality. However, there are systematic differences between natural kinds and manmade objects, one being that visual similarity is a stronger organizing dimension of natural kinds. If the biases in location memory reported in the main text are driven by visual similarity, such that typical category members are retrieved closer to their category's location because they are more often confused with visually similar category members, we reasoned that this effect would be stronger for animals over objects.

We tested this alternative prediction separately in the Experimental groups from Experiments 1 – 3, using 2 (typicality: typical, atypical) x 2 (superordinate category: animals, objects) repeated-measures ANOVAs with bias as the dependent measure. In all three experiments, there were significant main effects of typicality, all $F > 4.59$, all $p < .04$, as observed in the main text. There were no main effects of superordinate category, all $F < 1.81$, all $p > .19$. Critically, in Experiments 1 and 3 there were no reliable interactions between superordinate category and typicality, both $F < 2.25$, both $p > 0.14$, although there was a trend for an interaction in Experiment 2, $F_{(1,\ 34)} = 3.48$, $p = .07$, $\eta_p^2 = 0.09$. Interestingly, this was driven by greater bias for typical over atypical category members for object images, $t_{(34)} = 2.57$, $p = .01$, d = 43, as compared to animal images, $t_{(34)} = 0.35$, $p = .73$, d = 0.06. This is the opposite of what would be predicted if bias in memory was driven by higher visual similarity between category members.

For completeness, we also queried whether superordinate category modulated the influence of spatial consistency on error, by conducting separate 2 (spatial consistency: consistent, inconsistent) x 2 (superordinate category: animals, objects) repeated-measures ANOVAs over the Experimental groups of each of the three experiments. All revealed a main effect of spatial consistency as reported in the main text, all $F > 53.54$, all $p < .001$, and no main effect of superordinate category, all $F < 1.96$, all $p > .17$. In Experiments 1 and 2, there was no interaction, both $F < 1.38$, both $p > .24$, and in Experiment 3, there was a trend, $F_{(1,\ 114)} = 3.52$, $p = .06$, $\eta_p^2 = 0.03$. Here, the difference in error by spatial consistency was stronger for animals, $t_{(114)} = -7.65$, $p < .001$, d = -0.71, over objects, $t_{(114)} = -6.10$, $p < .001$, d = -0.57, although robust in both.

Experiment 1

Control analyses: We sought to assess whether the observed differences in error and bias could be explained by images' proximity to landmarks. To do this, we divided the grid into areas that were near and far from landmarks on the grid. Locations near landmarks fell within a 100-pixel border around the edge of the grid, or within 100 pixels of the vertical border dividing its right and left sides. As the entire grid was 600 x 1200 pixels, the areas far from landmarks constituted the two 400 x 400 pixel squares in the center of each side of the screen.

We chose this configuration because it gave rise to roughly equal numbers of typical and atypical category members near a landmark (typical: mean = 6.86, SD = 1.33; atypical: mean = 6.17, SD = 0.95) and far from a landmark (typical: mean = 7.37, SD = 1.60; atypical: mean = 7.60, SD = 0.91). Because there were uneven numbers of trials in each condition, we analyzed bias with a mixed effects linear model with typicality (typical, atypical), landmark (near, far), and their interaction as fixed effects. In addition to the effect of typicality observed in the main text, $F_{(1,\ 920.64)} = 9.20$, $p = .002$, there was a effect of landmark, $F_{(1,\ 925.35)} = 12.59$, $p < .001$. The effect of landmark reflected more bias for images located near landmarks relative to images far from landmarks. As many of the locations near landmarks were close to the edge of the grid, this effect may be driven by the fact that there was less area for these images to be placed that would be biased away from their category's center. Critically, there was no interaction, $F_{(1,\ 924.76)} = 0.00$, $p = .99$, suggesting that proximity to landmarks did not influence the effect of category typicality on bias observed in the main text.

Note that because the category centers were not located near a landmark, there is a large imbalance of spatially consistent and spatially inconsistent locations in each bin. Specifically, there were fewer spatially consistent locations near landmarks (consistent: mean = 8.60, SD = 0.20; inconsistent: mean = 13.03, SD = 0.29) and more spatially consistent locations far from landmarks (consistent: mean = 33.4, SD = 0.20; inconsistent: mean = 14.97, SD = 0.29). Nevertheless, we analyzed error in a mixed-effects model with spatial consistency (consistent, inconsistent), landmark (near, far), and their interaction as fixed effects. As in the main text, there was an effect of spatial consistency, $F_{(1,\ 296.33)} = 39.14$, $p < .001$. In addition, there was an effect of landmark,

$F_{(1,\ 35.05)}$ = 4.28, p = .046, with more error for images located near landmarks relative to images located farther away. There was no reliable interaction, $F_{(1,\ 2336.99)}$ = 3.35, p = .07, suggesting that proximity to a landmark did not reliably affect the extent that spatial consistency modulated error. However, since there was a trend, we conducted post-hoc tests. First, there was less error for spatially consistent over inconsistent images for those located both near to and far from landmarks, both t < -3.86, p < .001. However, we found that across spatially inconsistent images, there was more error for ones near a landmark relative to those far from one, $t_{(88.88)}$ = 2.76, p = .007, while proximity to landmarks did not influence error for spatially consistent images, $t_{(90.0)}$ = 0.48, p = .63.

Experiment 3

Exemplar memory by typicality and confidence: We used participants' reports of confidence about their chosen images in the exemplar memory test to gain confidence in this measure as an indication of memory strength. To do so, we computed a mixed-effects binary logistic regression with exemplar memory as the dependent variable and confidence, typicality, and their interaction as fixed-effects variables. Type II Wald χ^2 tests were used to determine significance. These revealed an effect of confidence, $\chi^2_{(2)}$ = 119.93, p < .001, with a greater likelihood of correct exemplar memory for images with higher confidence ratings. There was also an effect of typicality, $\chi^2_{(1)}$ = 4.56, p = .03, with a greater likelihood of correct exemplar memory for atypical images relative to typical images, as reported in the main text. There was no reliable interaction, $\chi^2_{(2)}$ = 1.68, p = .43. Overall, participants' confidence ratings reflected the likelihood that they chose the correct exemplar, providing justification for our decision to use exemplar memory as an index of memory for the perceptual details of the images.

Error by spatial consistency with semantic knowledge: Before considering location memory as a function of exemplar memory, we first sought to replicate the observations from Experiments 1 and 2 that memory was more accurate for images that were spatially consistent with their category membership. We computed a 2 (group: experimental, control) x 2 (spatial consistency: consistent, consistent) ANOVA. This revealed a main effect of spatial consistency, $F_{(1,228)}$ = 96.14, p < .001, η_p^2 = 0.30, and no main effect of group, $F_{(1,228)}$ = 0.007, p = .93, η_p^2 = 0.001, qualified by an interaction, $F_{(1,228)}$ = 28.48, p < .001, η_p^2 = 0.11. Both groups exhibited greater error for spatially inconsistent images relative to inconsistent images (experimental: $t_{(114)}$ = -10.47, p < .001, d = -0.98; control: $t_{(114)}$ = -3.23, p = .002, d = -0.30), but the interaction indicated that the effect was reliably stronger in the experimental group. As in Experiment 4, the dense clustering of images in spatially consistent locations may have aided retrieval in both groups. However, the reliably stronger effect in the experimental group demonstrates that retrieval was more accurate for images that were spatially consistent with their category membership, replicating Experiments 1 and 2.

Bias by category typicality: Before considering how exemplar memory related to bias in location memory, we first aimed to replicate the finding that typical category members were retrieved closer to their category's cluster relative to atypical members, as was observed in Experiments 1 and 2. We computed a 2 (group: experimental, control) x 2 (typicality: typical, atypical) ANOVA amongst spatially inconsistent images, with the average proportion of bias as the dependent variable. This revealed no main effect of group, $F_{(1,228)}$ = 0.002, p = .96, η_p^2 = 0.00, but a main effect of typicality, $F_{(1,228)}$ = 13.76, p < .001, η_p^2 = 0.06, qualified by a reliable interaction, $F_{(1,228)}$ = 10.53, p = .001, η_p^2 = 0.04. As in Experiments 1 and 2, retrieval was more biased towards the category center for typical images relative to atypical images in the experimental group, $t_{(114)}$ = 4.96, p < .001, d = .46, but not the control group, $t_{(114)}$ = 0.33, p = .74, d = .03.

Error by category typicality: We also aimed to replicate the observations of Experiments 1 and 2 showing more error for typical category members relative to atypical category members (Figure S1). An 2 (group: experimental, control) x 2 (typicality: typical, atypical) ANOVA revealed a main effect of typicality, $F_{(1,228)}$ = 4.09, p = .04, η_p^2 = 0.02, no main effect of group, $F_{(1,228)}$ = 0.85, p = .36, η_p^2 = 0.04, and no reliable interaction, $F_{(1,228)}$ = 2.84, p = .09, η_p^2 = 0.01. To compare with Experiments 1 and 2, we ran post-hoc t-tests to compare error by typicality and group. As in Experiments 1 and 2, we found greater error for typical images relative to atypical ones in the experimental group, $t_{(114)}$ = 2.31, p = .02, d = 0.22, but not the control group, $t_{(114)}$ = 0.28, p = .78, d = 0.03.

Supplemental Methods

All Experiments

Amazon Mechanical Turk: All data was collected on Amazon Mechanical Turk (AMT). AMT participants more closely match the demographics of adults in the United States relative to traditional lab experiments (Buhrmester, Kwang, & Gosling, 2011) and the quality of data is equivalent to that collected from lab participants across a number of learning and memory experiments (Brady & Alvarez, 2011; Crump, McDonnell, & Gureckis, 2013). All experiments were restricted to participants who were located in the United States, had completed > 100 HITs with at least a 95% acceptance rate, and had not participated in any prior pilots or other lab experiments that used the same stimuli. These restrictions were chosen to increase the likelihood that participants were native or fluent English speakers who could understand all provided instructions and were naïve to the aim of the experiments. Depending on the length of the task, participants were given a >5-hour time window to complete each experiment so they could choose a suitable time to complete the tasks without interruption.

Experiment 1: Stimulus Development

To create the image-location associations used in Experiment 1, we developed a data-driven approach to estimate how images would be spatially organized if their locations adhered to their category membership. To do this, we used relatedness ratings from a separate set of participants who completed an odd-man-out task. These were used both to define images' locations and identify their category membership and typicality.

Participants

24 participants (23 – 49 years old, 9 female) completed the experiment. The University of Pennsylvania Institutional Review Board (IRB) approved all consent procedures.

Materials

Images: Stimuli consisted of 70 150 x 150 pixel color images on white backgrounds (35 animals, 35 common objects) from the Bank of Standardized Stimuli (Brodeur, Dionne-

Dostie, Montreuil, & Lepage, 2010; https://sites.google.com/site/bosstimuli/home). These images were re-sized to 100 x 100 pixels for the memory experiments.

Procedure

Picture-naming: This task was developed to (1) ensure that participants could correctly identify the animals and objects, (2) familiarize the participants to the range of stimuli over which they would be making similarity judgments, and (3) create a pool of qualified participants that would be invited to complete the odd-man-out task. In this task, participants were instructed to name each image as specifically as possible. For example, if they viewed a picture of a bird, they were instructed to name the species of bird if possible. On each trial, participants viewed an image and chose from three options presented below the image: "I know this", "I know this but can't think of its name", or "I don't know what this is". If they chose "I know this", they also typed the name of the animal or object next to their choice. Either pressing the 'enter' key after filling in the name, or clicking on the button corresponding to the other two choices, automatically advanced the participant to the next trial with a 200ms inter-trial-interval (ITI). Otherwise, the trial advanced automatically after 15 seconds of inactivity. Trials were divided into separate animal and object blocks that were randomized across participants. The order of images in each block was randomized for each participant. On average, the task was completed in 2 - 4 minutes.

40 participants were recruited to participate in the picture-naming task. They were compensated with $1.50. One participant was excluded for entering nonsense words for the majority of animal trials, and an extra participant was recruited as a replacement. As the aim of this task was to ensure that participants had sufficient knowledge of the animals and objects represented in each image, which participants may have even if they do not know the specific name of an image, we developed a lenient criterion for analyzing this task. Close neighbors of an image, such as crocodile for alligator, lizard for gecko, utility knife or razor for box cutter, were accepted as correct. Misspelled names were included if they matched the correct image name with a maximum generalized Levenstein edit distance of 0.3. Using these criteria, 94.93% of images (SD: 8.94%) were named correctly. When including "I know this but can't think

of its name" as accurate responses, accuracy rose to 97.71% (SD: 5.47%), signifying near-perfect knowledge of the images. 35 participants who accurately named at least 65 out of 70 images were invited to participate in the next phase of the experiment; the excluded 5 participants named 39 - 62 images correctly.

Odd-man-out: Eligible participants were invited to participate in the odd-man-out (OMO) task. In this task, participants were presented with three animals or three objects and were instructed to click on the picture that 'did not belong', or in other words, the animal or object that was least similar to the other two. Once an image was chosen, the three images were replaced by three new images after a 200ms ITI. Trials were untimed, but participants were instructed to respond in 2 – 4 seconds and make their decision as quickly as possible while still being accurate. They were told that there were no right answers, but that some trials would seem easier than others. They were instructed to make their decisions based on many factors, instead of focusing on one. For animal trials, this included factors like whether they are in the same family, share similar habitats, and are predators or prey. For object trials, this included factors like whether they serve a similar purpose, where they are commonly located, and their size.

For each category, a complete testing of all triplets of images would require 6,545 trials (choose 3 given 35). Piloting of this task with a prior cohort (N = 6) revealed that for each participant, similarity matrices derived from a randomly sub-sampled set of two-fifths of these combinations were highly correlated with similarity matrices derived from the participant's full set of trials (for all 6 pilot participants, animals: all r > 0.90; objects: all r > 0.81). Thus, we presented participants with a random sample of 2,620 combinations per category (two-fifths of all possible triplets). This task was divided into 20 separate batches (10 per category) expected to take 12 - 15 minutes each. Each batch comprised 262 trials of a given category and 5 attention checks. In an animal batch, an attention check comprised two animals and one object. Participants were instructed to choose the object. In an object batch, participants were instructed to choose the animal presented with two objects.

Although it is impossible to control the order of HITs that AMT workers choose to work on, we attempted to randomize the order of batches that workers had access to by creating six different sequences of batches, and assigning a worker to one of the sequences based on the last letter or number in their account ID. AMT IDs are randomly generated string of letters and digits, so assigning groups by the last letter or number of a worker's ID is akin to sampling from a uniform distribution of 1 to 36 (26 letters, 10 digits). Within each batch, the trial order was randomized for each participant.

Participants were given 7 days to complete all 20 batches. Upon the invitation to complete the batches, participants were told that they would be compensated $2.50 per batch, and if they completed all 20 within 7 days, they would receive a bonus of $5, for a total of $55. Two reminders were sent over the course of the week, and a third reminder was sent the day before the deadline, only to participants who had already completed >15 batches. If a participant missed 2 or more attention checks in a batch, it was rejected and a message was sent to the participant offering them an opportunity to re-do the batch. Out of the participants whose data were used in the memory task, only one participant needed to re-do a rejected batch. At the end of the week, 24 out of 35 participants had completed all 20 batches. The remaining 11 participants had completed 1 to 13 batches (mean = 4, SD = 4.44).

The OMO judgments were then used to create pairwise similarity matrices for each subject and superordinate category (Figures S6A, S6D). Starting with a 35 x 35 matrix of zeros, for each trial in which an image was chosen, the similarity value for the other two presented images increased by one. These similarity values were summed across all trials and then divided by the number of times the two images appeared in the same trial. Similarity matrices thus ranged from 0 to 1, with higher values corresponding to greater similarity between images. To quantify the test-retest reliability of the ratings for each participant, split-half correlations were derived by re-computing pairwise similarity using the first and last half of the batches that were completed for each subject, and then correlating the lower triangle of the two matrices together. Correlations between the first and last half of batches were high (animals: mean r = 0.66, SD = 0.05; objects: mean r = 0.54, SD = 0.06) and greater than correlations between each participant's ratings and similarity matrices derived from Latent Semantic Analysis (LSA) and Word2Vec (Figures S6C, S6F). Three subjects, whose split-half correlations for both categories were > 2 SD lower than the group mean, were excluded (all *r*'s < 0.08), leaving 21 subjects with data to be used to derive the locations for the memory test.

Category membership and typicality: The odd-man-out results were used to derive a 2-dimensional representation of participants' semantic knowledge (Figures S6B, S6E). To do this, the similarity matrices from the 21 participants who completed the odd-man-out task were averaged into two group-level similarity matrices, one for each superordinate category. Classical multi-dimensional scaling was used to project the matrices into 2D spaces. To evaluate how well this projection reflected participants' decisions in the OMO task, we computed the goodness of fit of the projected data to the full similarity space using subsets of the dimensions produced by the scaling procedure. This measure was operationalized as the R^2 between the full similarity matrix and a reconstructed matrix using 1 to 10 coordinate dimensions weighted by their eigenvalues (Figure S6G). While there were no obvious bends in this elbow plot, moving from one to two dimensions resulted in the steepest change in fit, and two dimensions captured 85.5% of the variance in ratings for animals and 70.0% of the variance for objects.

K-means clustering was applied to the 2D coordinates to derive data-driven clusters of images. The optimal number of clusters was determined by computing the within-cluster sum of squared error for solutions with 1 to 10 clusters and using an elbow plot to identify the smallest number of clusters that explained the largest portion of error (Figure S6H). For both categories, the data was best explained with 3 clusters. This data-driven procedure resulted in animal clusters corresponding to birds, land mammals, and sea creatures, and object clusters corresponding to kitchen utensils and appliances, tools and personal care items, and office supplies.

The cluster solutions were also used to identify images that were typical and atypical category members. First, a category was operationalized as images belonging to the same cluster (e.g. birds). The center of each cluster was derived by computing the mean x and y coordinate over all cluster images. Then, all images were sorted by their distance to their cluster center. For each cluster, the closest 20% of images were labeled as 'typical' category members (e.g. cardinal), and the furthest 20% of images were labeled as 'atypical' members (e.g. toucan and ostrich). Note that this data-driven procedure resulted in differing numbers of category members in each group, and thus differing numbers of typical and atypical category members. Table S3 lists the number of images per category and condition.

Counterbalancing: For some categories, the number of category members did not evenly divide into 40%, such that there were an odd number of images assigned as spatially inconsistent that could not be evenly assigned to typical and atypical category members. To account for this and for other variations in the stimulus display, we created 6 counterbalancing groups. In each group, we (1) randomized the assignment of the odd image to a typical or atypical category member, (2) generated a new, random set of spatially inconsistent locations, and (3) counterbalanced whether animals appeared on the left side of the screen and objects on the right, or vice versa. AMT workers were assigned to one of the six groups based on the last letter or number in their Worker ID. Worker IDs are randomly generated string of letters and digits, so using the last letter or number of a worker's ID is akin to sampling from a uniform distribution of 1 to 36 (26 letters, 10 digits).

Picture-naming responses sorted by category typicality: The picture-naming task was developed to ensure that participants could correctly identify the animals and objects, and to familiarize the participants to the stimuli before making similarity judgments. We also used responses from this task to investigate whether responses to the images differed based on their category typicality, which was subsequently calculated based on the responses to the odd-man-out judgments. Paired t-tests revealed no difference in average accuracy between typical and atypical category members, $t_{(39)} = 0.52$, $p = .61$, and faster median response times for typical category members relative to atypical ones, $t_{(39)} = 5.18$, $p < 0.001$. The faster response times for typical category members over atypical ones is consistent with a highly reliable body of work showing a word frequency effect, such that pictures naming is faster for high-frequency objects (e.g. Oldfield & Wingfield, 1965). The lack of a difference in accuracy may be due to the fact that accuracy was at ceiling overall, and that even though the images range in their frequency, on average they were all relatively easy to identify.

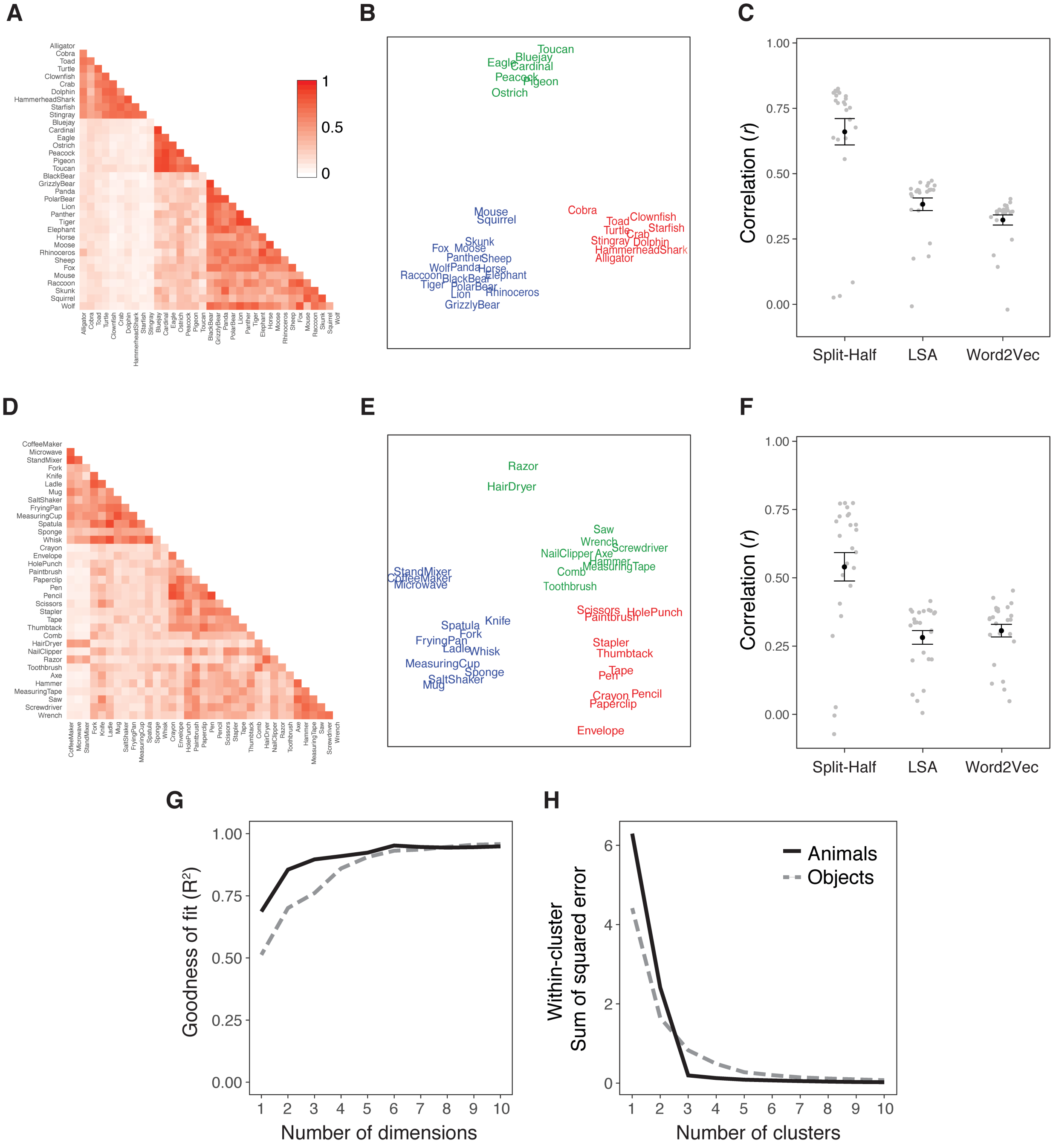


Figure S6. Odd-man-out procedure. (A, D) Similarity matrix of all animals and all objects derived from odd-man-out judgments. **(B, E)** 2D projection of animal and object similarity matrices using classical multi-dimensional scaling. Colors indicates 3 clusters derived with k-means clustering. **(C, F)** Split-half correlations of the similarity matrices within each participant plotted alongside each participant's correlation with similarity matrices derived from LSA and Word2Vec. Gray dots signify participants. Error bars indicate SEM. **(G)** Goodness of fit of projected similarity ratings derived from the group-averaged correlation matrix, as a function of the number of reconstructed dimensions. **(H)** Elbow plot of within-cluster sum of squared error for the k-means clustering algorithm as a function of number of clusters chosen.

Experiments 1 & 3				
Category	**Total**	**Spatially Consistent**	**Typical**	**Atypical**
Birds	7	4	1 or 2	1 or 2
Mammals	18	11	3 or 4	3 or 4
Sea Creatures	10	6	2	2
Tools/personal care	11	7	2	2
Kitchen	13	7	3	3
Office	11	7	2	2
Experiment 2				
Category	**Total**	**Spatially Consistent**	**Typical**	**Atypical**
Birds	20	14	3	3
Mammals	20	14	3	3
Sea Creatures	20	14	3	3
Insects	20	14	3	3
Clothes	20	14	3	3
Furniture	20	14	3	3
Kitchen	20	14	3	3
Office	20	14	3	3

Table S3. Number of images in each category in each experiment. **Top.** Experiments 1, 3 & 4. 40% of items in a category (the top and bottom 20%) were assigned as spatially inconsistent. When this proportion gave rise to an odd number of items (e.g. in the bird and mammals category), the odd one was randomly assigned as typical or atypical for each counterbalancing group. In other words, a given counterbalancing group encoded one typical bird and two atypical birds, and another group encoded two typical birds and one atypical bird. Experiment 3 used the same categories, locations, and counterbalancing groups developed for Experiment 1. In Experiment 4, the number of items in each color category was randomly matched to a semantic category such that the number of items in each category varied in the same was as in Experiments 1 and 3. **Bottom.** Experiment 2. As categories were predetermined in Experiment 2, all categories comprised 20 items. 30% of items in a category (the top and bottom 15%) were assigned as spatially inconsistent.

Experiment 2: Stimulus Development

We aimed to replicate Experiment 1 with validated approaches for defining category membership and typicality. To this end, a separate cohort of participants completed a list ranking task to identify typical and atypical members in a fixed set of categories.

Methods

Participants

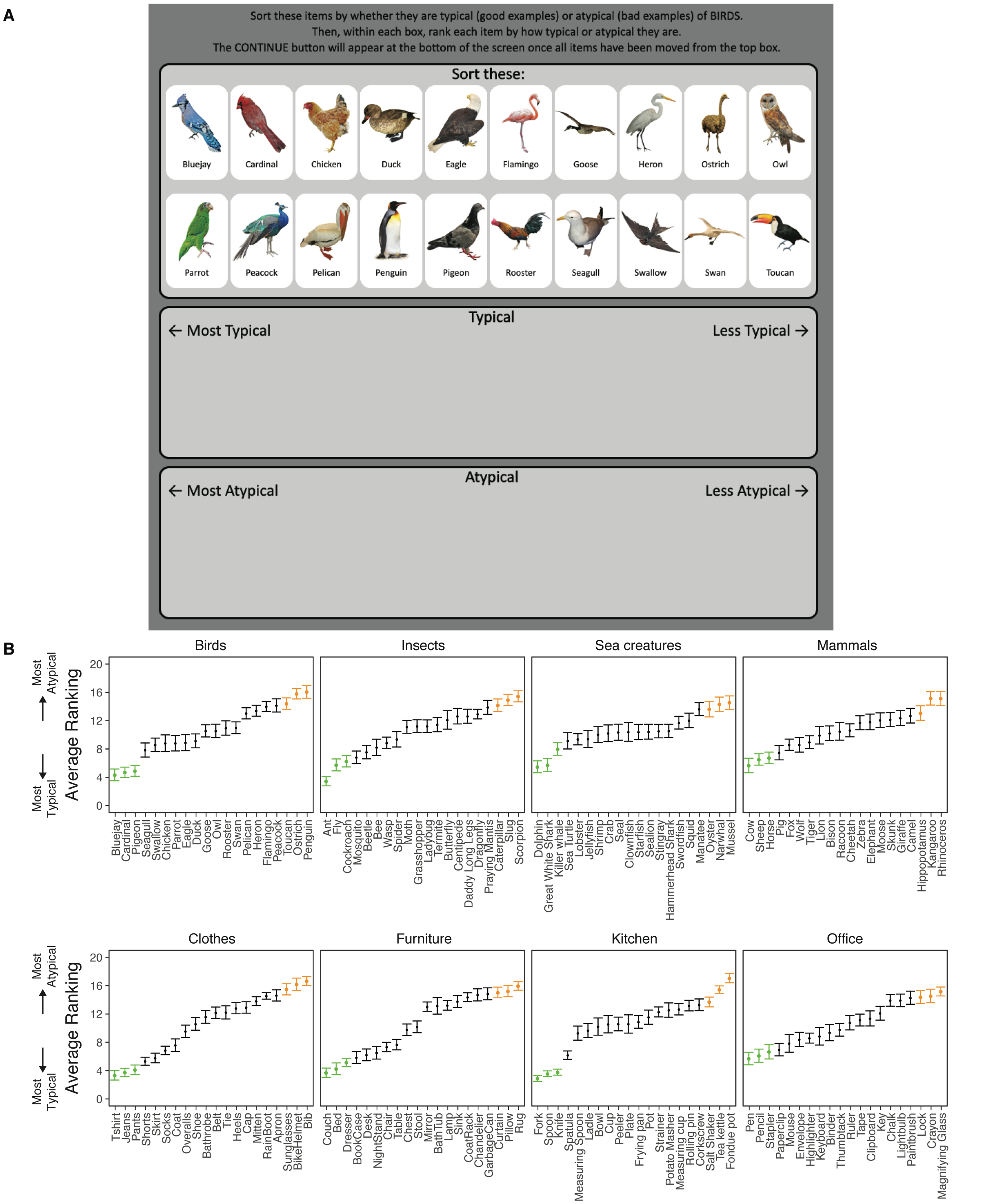
216 participants (27 per category) completed this task. The University of Pennsylvania IRB approved all consent procedures. Demographics were not collected due to experimenter error.

Materials

Stimuli comprised 160 100x100-pixel color images on white backgrounds. These comprised 8 categories of 20 images: birds, insects, sea creatures, mammals, clothes, furniture, kitchen utensils, and office supplies. The categories were selected from prior studies of categorization norms (Uyeda & Mandler, 1980; Deyne et al., 2008).

Procedure

We employed a validated list ranking task (Figure S7A; Djalal, Ameel, & Storms, 2016). Extensive instructions with examples were given to ensure participants understood the concept of category typicality. For each category, participants viewed 20 images in a box labeled ‘Sort these’. Underneath, there were two empty boxes labeled ‘Typical’ and ‘Atypical’. Participants were instructed to drag 10 images into each box. They were allowed to drag images freely across the boxes in any order. Then, within each box, participants sorted the 10 images as most (a)typical to less (a)typical. Arrows and labels indicated the direction that images were to be sorted. The average completion time was 2.6 minutes (SD = 1.9). The resulting positions of the images were concatenated into a ranked list of category typicality and averaged across participants. The top three and bottom three images in each list were assigned as ‘typical’ and ‘atypical’ category members that would be ‘spatially inconsistent’ in the memory experiment, and the images in the middle of the list were assigned as ‘spatially consistent’ (Figure S7B).



Sort these items by whether they are typical (good examples) or atypical (bad examples) of BIRDS.
Then, within each box, rank each item by how typical or atypical they are.
The CONTINUE button will appear at the bottom of the screen once all items have been moved from the top box.

Figure S7. Ranking procedure (A) Ranking task and instructions for the bird category. **(B)** Average typicality rankings for each category, which could range from 1 (all participants ranked image as most typical) to 20 (all participants ranked image as most atypical). Error bars indicate SEM. Color indicates condition in the memory experiment: green = spatially inconsistent/typical; black = spatially consistent; orange = spatially inconsistent/atypical.

Experiment 3

Power analysis: We chose a new sample size for this experiment to account for the fact that the planned analyses were less powered relative to the analyses in the first two experiments. Specifically, because we planned to analyze error and bias as a function of exemplar memory accuracy, requiring the division of each condition by whether the encoded exemplars were correctly or incorrectly chosen at test, there would be fewer trials in each condition. To address this, we first collected data from subset of the shuffled control group (N = 35) to procure an estimate of exemplar memory accuracy in the absence of informative category knowledge during encoding. We found that accuracy was on average 67.6% (SD 16.6%). We then re-analyzed the influence of category typicality on bias from the experimental group in Experiment 1 (N = 35) with a subset of their trials that matched the number of exemplar incorrect trials from the pilot group. We chose to conduct this re-analysis using incorrect trial counts because exemplar accuracy in the initial subset was above 50%, meaning there were fewer incorrect trials than correct trials and any analyses including incorrect trials would be the least powered.

The subsampling procedure was as follows. First, we matched each participant in Experiment 1 to a participant in the Experiment 3 pilot cohort. We randomly subsampled the number of typical and atypical images in the Experiment 1 participant to match the number of incorrect exemplars that the Experiment 3 pilot participant chose, and then we derived average bias separately for these 'incorrect' typical and atypical category members. We repeated this subsampling procedure 5,000 times per participant. We then computed the effect size of the difference in bias for typical versus atypical category members for each permutation and averaged them. This resulted in an estimated effect size of the difference in bias by typicality that would have been observed in Experiment 1 when limiting the analysis to trials with incorrect exemplar memory, assuming both groups of participants exhibited equivalent memory for the specific images. As expected, this procedure resulted in a smaller effect size (Cohen's *d* = 0.263) than what was observed in Experiment 1 when including all trials (Cohen's *d* = 0.548). Using a power analysis ($\alpha = 0.05$, power = 0.80), we established that a sample size of 115 participants per group was needed to recover this smaller effect.

Experiment 4

Counterbalancing: In Experiment 4, we aimed match as many features of the experimental design to Experiment 1 as possible in order to understand whether biases in memory are driven by category knowledge or by newly-learned visual similarities of encoded images. Because a data-driven clustering algorithm determined category membership of the stimuli in Experiment 1, the number of category members varied across categories (Table S3). Thus, to match this in Experiment 4, we created 6 counterbalancing groups, where in each one, the categories in Experiment 1 were randomly matched to Experiment 4. For example, because in Experiment 1 there were 7 birds, 18 mammals, and 10 sea creatures, in one counter-balancing group there were 7 lamps, 18 clocks, and 10 chairs, and in another counter-balancing group there were 7 trains, 18 cars, and 10 planes.

As described in the main text, both the assignment of the six color ranges to the six categories and the assignment of colors to images within each category were also randomized separately for the 6 groups. Thus, each group encoded a unique set of stimuli with (1) different numbers of category members per category, (2) categories with different ranges of colors, and (3) different images assigned typical and atypical colors.

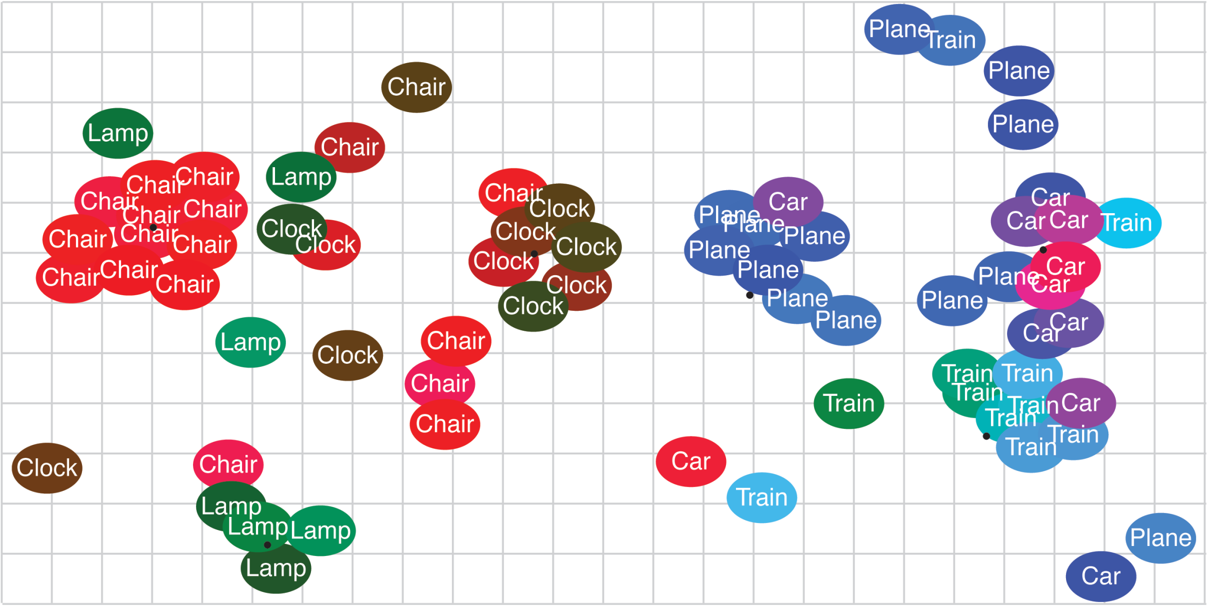


Figure S8. Experiment 4 stimulus display. Spatial locations for the six categories (white text) and the dominant color of its associated image (ovals).

References

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLOS ONE*, *5*(5), e10773.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), e57410.

Deyne, S. D., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Djalal, F. M., Ameel, E., & Storms, G. (2016). The typicality ranking task: A new method to derive typicality judgments from children. *PLOS ONE*, *11*(6), e0157936.

Oldfield, R. C., & Wingfield, A. (1965). Response Latencies in Naming Objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Uyeda, K. M., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, *12*(6), 587–595.