Title
Fragile Associations Coexist With Robust Memories for Precise Details in Long-Term Memory

Journal
Journal of Experimental Psychology Learning Memory and Cognition, 42(3)

ISSN
0278-7393

Authors
Lew, Timothy F
Pashler, Harold E
Vul, Edward

Publication Date
2016-03-01

DOI
10.1037/xlm0000178

Peer reviewed

Fragile associations coexist with robust memories for precise details in long-term memory

Timothy F. Lew, Harold E. Pashler, & Edward Vul
Department of Psychology
University of California, San Diego

Abstract

What happens to memories as we forget? They might gradually lose fidelity, lose their associations (and thus be retrieved in response to the incorrect cues), or be completely lost. Typical long-term memory studies assess memory as a binary outcome (correct/incorrect), and cannot distinguish these different kinds of forgetting. Here we assess long-term memory for scalar information, thus allowing us to quantify how different sources of error diminish as we learn, and accumulate as we forget. We trained subjects on visual and verbal continuous quantities (the locations of objects and the distances between major cities, respectively), tested subjects after extended delays, and estimated whether recall errors arose due to imprecise estimates, misassociations, or complete forgetting. Although subjects quickly formed precise memories and retained them for a long time, they were slow to learn correct associations, and quick to forget them. These results suggest that long-term recall is especially limited in its ability to form and retain associations.

Keywords: Visual memory, Long-term memory, Associative memory

What happens to memories as we forget? If, for instance, you return from a trip and try to remember where you left your car keys, there are different ways your memory of their location could have deteriorated. You may misremember the location of the keys by several feet (imprecise recall of the correct location). Perhaps you will look for your car keys in the place where you left your umbrella (associate objects with the incorrect locations). Or maybe you will completely forget where you left your keys, and randomly guess where they might be. How much do imprecise recall, misassociations, and altogether losing locations contribute to memory errors?

Most investigations of long-term memory examine recollection in an all-or-none manner: either a memory is recalled/recognized or it is not. Consequently, these studies rely on indirect measures and qualitative manipulations to estimate association fidelity and memory precision. For instance, by comparing recall for individual items with cued recall for paired associates, researchers have tried to isolate failure to recall an item from failure to correctly associate that item (Tulving & Wiseman, 1975). Similarly, others have qualitatively estimated memory precision by comparing people’s ability to distinguish categorically (e.g., two different mailboxes) and perceptually (e.g., a mailbox when it is open vs. closed) similar images (Brady, et al., 2008). Superficially, it would seem that the application of signal detection theory to recognition memory provides a framework for estimating the strength of memories via binary accuracy rates at different confidence judgments (Green & Swets, 1966; Wickelgreen & Norman, 1966). However, this “memory strength” could be interpreted either as memory precision or as association fidelity. Although these studies have provided important insights into the content and structure of memory, they can only indirectly assess how memories degrade over time by using confidence judgments as proxies for precision or by comparing accuracy rates in qualitatively different conditions.

In contrast, recent visual working memory studies have used continuous report tasks in which subjects recall the exact features of objects (e.g., color, orientation, size) to test how different types of errors affect memory. Analyses of such continuous report data via mixture models can then estimate the extent to which errors arose due to imprecise responses about the correct feature value, misassociations and random guesses (Bays & Husain, 2008; Zhang & Luck, 2008; Anderson, Vogel, & Awh, 2011; Bays, Wu, & Husain 2011; see Ma, Husain & Bays, 2014, for a review).

Despite the recent explosion of interest in continuous report tasks in visual *working memory*, relatively few studies have investigated how different types of errors contribute to forgetting in visual *long-term memory*. Brady, et al. (2013) used a continuous report task to examine the extent to which the fidelity of memories and complete forgetting affected memory, finding that the rate of random guesses increases with delays but long-term memory precision matches that of working memory when it is least precise. However, Brady, et al.'s retention intervals did not exceed about an hour, so they could not assess forgetting over longer intervals. Moreover, they did not examine misassociations and consequently may have mischaracterized misassociations as random guesses, and underestimated how much information long-term memory retained.

Here we examine the time course over which memories are acquired, gain precision, and form associations during training, and how these memories then deteriorate over time. We asked subjects to learn and later recall the locations of objects (Experiments 1, 2, 3) or the distances between cities (Experiment 4). We then used a mixture model to estimate the precision of their memories, as well as the proportion of their responses that reflected imprecise reports of the correct item, imprecise reports of one of the *other* items (a misassociation), or a random guess.

Experiments 1 & 2

To assess how memories formed over the course of learning and were lost over time, we used a cued recall task to train subjects on the locations of objects until they reached a performance criterion (Experiment 1) and test them after delays up to one week (Experiment 2). On both training and testing trials, subjects recalled the location of cued objects, but they received the correct location as feedback only on training trials.

Methods

Subjects. 40 subjects from the Amazon Mechanical Turk marketplace participated in Experiment 1. Because Experiment 2 required subjects to participate in 3 sessions spanning a week, we recruited 35 members of the UCSD Psychology Department's online subject pool. In both experiments subjects received a flat payment as well as a bonus based on their performance.

Design. Experiment 1 focused on the acquisition of memories. Each subject learned the locations of ten objects using testing with feedback over multiple blocks. Each subject proceeded through as many of these training blocks as required to recall the locations of all the objects in a block sufficiently precisely (see Procedure). The order of the ten objects was randomized within each block.

Experiment 2 focused on the forgetting of memories. The training session was similar to Experiment 1 with the exception that objects dropped out when they were recalled correctly three blocks in a row. Once subjects learned the locations of all the objects to sufficient precision, they performed a distractor task (12 addition and subtraction problems, each containing two operands that were whole numbers between 0 and 40), and were then tested on the object locations. Subjects then returned for two testing sessions after delays of one day (test day-1) and seven days (test day-7).

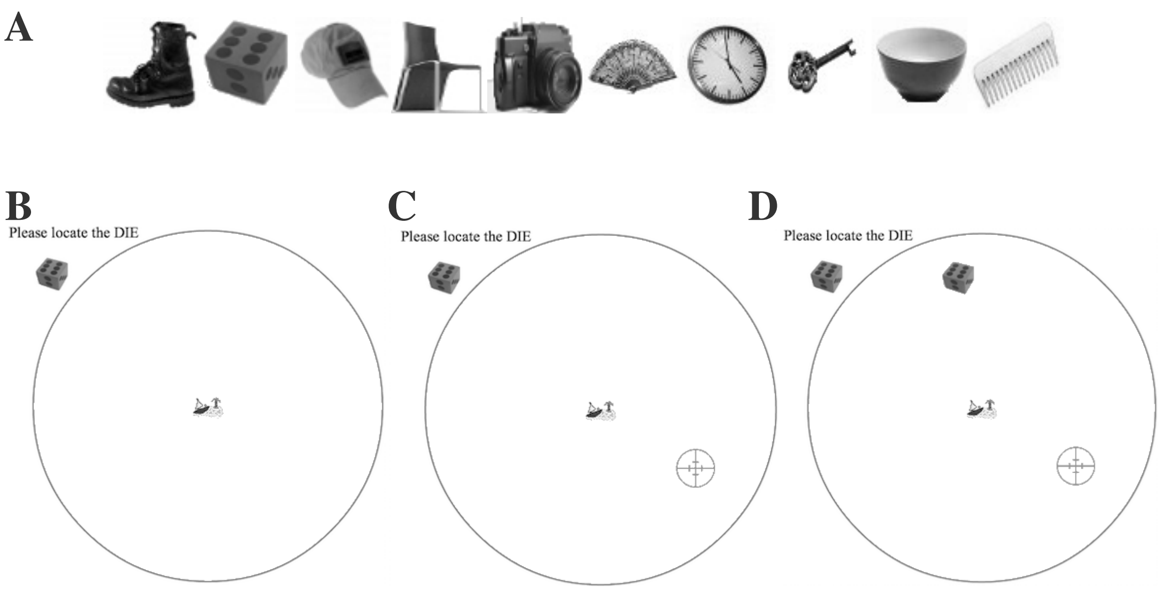


Figure 1. The objects used in Experiments 1, 2 and 3 and an example trial for Experiments 1 and 2. (A) The ten objects used in the visuospatial memory task: boot, die, hat, chair, camera, fan, clock, key, bowl and comb (in the experiment objects were presented in full color). In each trial, (B) subjects were cued to recall the location of the item indicated in the top-left (here: a die). (C) Subjects then clicked a location to respond and a red crosshair marked their selection. (D) During training trials, subjects were then

shown the item in its correct location for one second (this feedback was omitted during test trials).

Stimuli. In both experiments subjects trained on the locations of ten everyday objects (Figure 1A). The cover story for the task was that the subject had lost several of their personal belongings in the ocean and had to remember where those objects were underwater. Objects were presented in a light blue circle with an island in the center that acted as a central location landmark and enhanced engagement with the cover story (see Figure 1B). Apart from their role in the cover story, the color of the background and the island in the center were unrelated to the task.

Because our focus was on learning over many repeated presentations under free-viewing, we did not ask subjects to maintain fixation. Additionally, because each participant performed the study in their own web browser, screen size and viewing distance were not explicitly controlled but subjects were instructed to adjust their browser window size such that the entire experiment display would fit on the screen.

Each object was represented by a 60 × 60 px image of an everyday object (drawn from a stock image website: www.freeimages.com). We selected ten perceptually and semantically distinct objects to minimize their confusability, and every subject saw those same ten objects. The circle containing the objects had a radius of 450 px and the island was 50 × 50 px.

For each subject, we generated the locations of objects from a uniform distribution across the circle (with the constraint that they did not overlap with the island).

Procedure. Subjects were trained and then tested on the locations of objects using a cued recall task (Figure 1B). During the training phase of both experiments, on each trial subjects saw an image of an object and reported that object's location by clicking within the display circle. After the response, a 50 × 50 px red crosshair appeared at the selected location, and an image of the object appeared at the correct location. If the response was within 50 px of the correct location (such that the crosshair overlapped with the object image), the response was considered correct.

In the training experiment (Experiment 1), a subject completed the training phase (and thus the experiment) once she recalled all the objects correctly in one block.

In the training phase of the retention experiment (Experiment 2), an object was "dropped" out of the training loop after it was correctly recalled in three consecutive blocks, and the training phase was complete once all objects had been dropped.

Trials in the testing phase of Experiment 2 were the same as training trials, but lacked corrective feedback (instead the subject's response was indicated by a red crosshair onscreen for an extra second).

Results

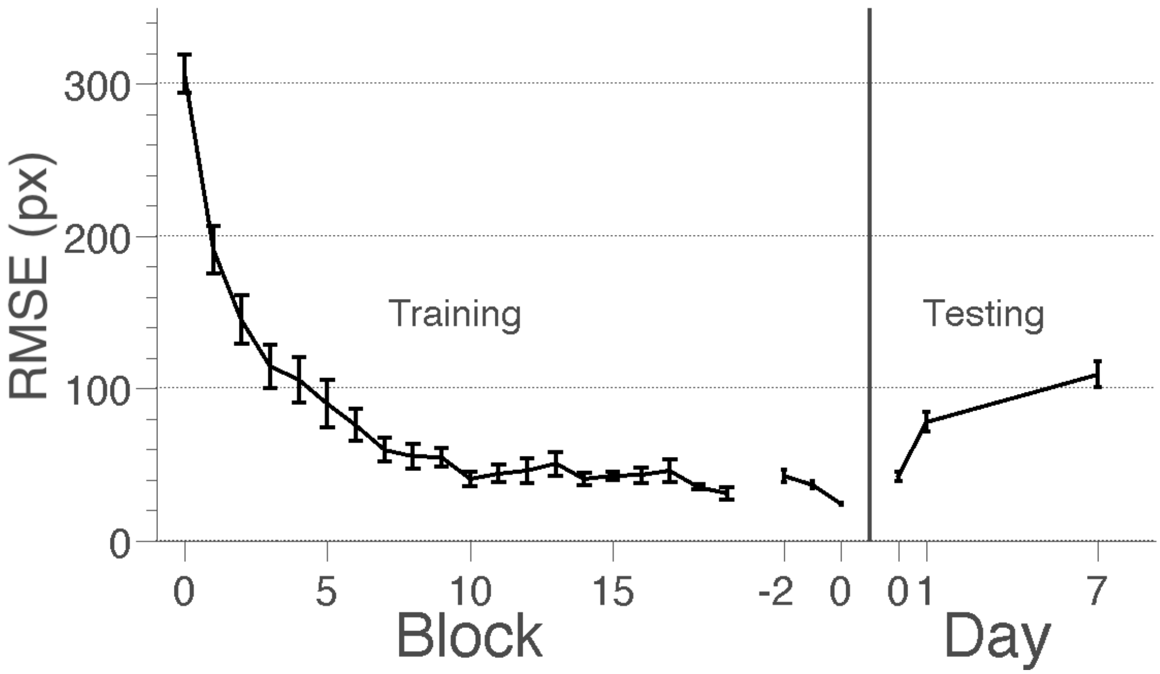


Figure 2. Learning curves from Experiment 1 and the forgetting curve from Experiment 2. Y-axis is the across-subject mean (±1 SEM across-subjects) of the root mean squared error (Euclidean pixel distance between the recalled and the correct object location). Training performance from Experiment 1 is shown on the left (in Blocks) and testing performance from Experiment 2 on the right (in Days). Training performance is shown relative to the beginning of training (Blocks 0 to 20) and relative to the end of training (Blocks -2 to 0) to illustrate criterion performance. Root mean square error (RMSE) decreased during training and increased during testing, indicating the subjects learned and forgot the locations of objects.

Did subjects learn and forget the locations of objects? To coarsely assess learning and forgetting, we can consider the average distance between the reported and correct locations (calculated as the root mean squared error across objects; RMSE). This coarse measure of learning shows that subjects learned the locations of objects over approximately 12.25 blocks (*SEM=1.08*) of training in Experiment 1 (Figure 2, Training) and forgot some, but not all, of what they learned during the 1-week retention interval in Experiment 2 (Figure 2, Testing). Because the number of blocks it took subjects to finish training varied, we examined how well subjects recalled the locations once they completed training by calculating the RMSE of each subject's last three blocks of training (Figure 2, Training, blocks -2—0). Performance was worse during the first testing block (Experiment 2) compared to the end of training (Experiment 1) ($t(75)=6.45$, $p<.001$), though we cannot say how much this should be attributed to rapid forgetting or subtle differences in the training protocol between the two experiments. While this coarse error measure shows that subjects are indeed learning and forgetting something about the locations of objects, it cannot discern whether errors are attributable to imprecision, misassociations, or complete forgetting.

Measuring imprecision, misassociations, and random guessing

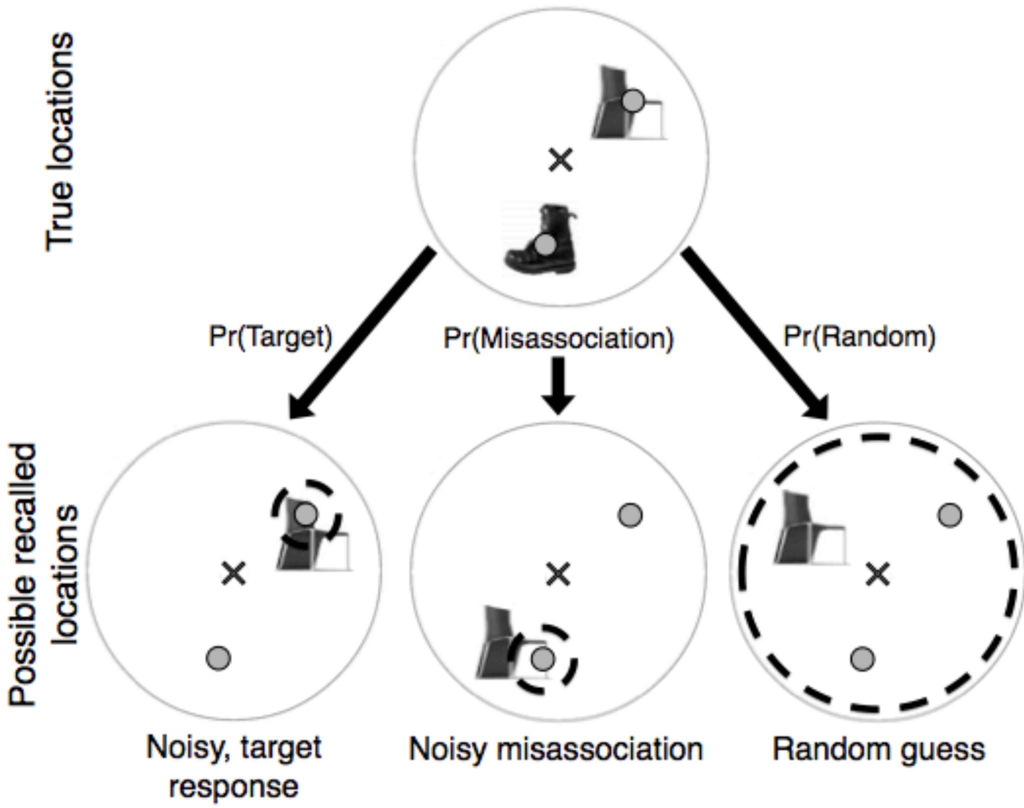


Figure 3. Schematic of the types of errors we aim to characterize with the mixture model: imprecise report of the target, misassociation, or random guess (illustrated with just two objects that are displayed disproportionately large for visual clarity). The top-row shows the true locations of the objects and the bottom-row shows possible types of responses. Grey dots represent the locations of the target object (the chair) and a possible non-target object (the boot). The grey X indicates the center of the environment. We use the mixture model to estimate the probability of each type of error, denoted here by Pr(Target), Pr(Misassociation) and Pr(Random). If there are multiple non-target objects, Pr(Misassociation) is divided evenly among them. Under our error model, target reports and misassociations are recalled with isotropic, two-dimensional Gaussian noise around the selected location (small, grey dashed circles). The model treats random guess responses as samples from a broad, truncated, two-dimensional Gaussian distribution around the display center (large, dashed circle).

To characterize the contributions of imprecision, misassociation, and complete forgetting of memories during learning and forgetting, we analyzed subjects' responses with a mixture model, similar to that used in Bays, et al. (2011) (Figure 3; see Appendix A, Model overview, for technical details). Under this model, each response is either an imprecise report of the target item, an imprecise report of one of the *other* items (a misassociation), or a random guess. A report of the target object location or a misassociated location is assumed to be distributed as an isotropic two-dimensional Gaussian centered on an object's location. Random guesses are assumed to be samples from a truncated two-dimensional Gaussian distribution centered in the environment and bound by the environment's edge[*]. The model estimates a single parameter for the precision of location memories; thus it assumes that correctly associated responses and responses when objects are associated with the wrong location have the same precision around their latent location[†]. The model also estimates the mixture weights of each type of response, corresponding to the probabilities that subjects report the location of the target item, make a misassociation, and randomly guess. Thus, by analyzing responses via this mixture model, we can estimate the precision of location memory, the probability of misassociations, and the probability of complete forgetting (random guessing). We fit the model in the native coordinate space rather than to the distribution of response errors (as in Zhang & Luck, 2008), though our results do not depend on this distinction.

In several of our analyses, we report the posterior distributions of the parameters estimated by the model in the form of 95% Posterior Quantile Intervals (95% PQI). For further explanation of 95% Posterior Quantile Intervals and how we report Bayesian statistics, see Appendix C, Bayesian statistics reports.

---

[*] Although we did not use truncated normal distributions to model target or misassociated responses due to computational efficiency, the small standard deviation of location memories should result in a negligible portion of the probability density extending outside of the environment, thus making the truncation correction unnecessary.

[†] It is possible that locations associated to incorrect objects would be remembered with different levels of precision. However, we assume that locations and associations are stored and decay separately such that whether a location is correctly associated with an object will be independent of its precision.

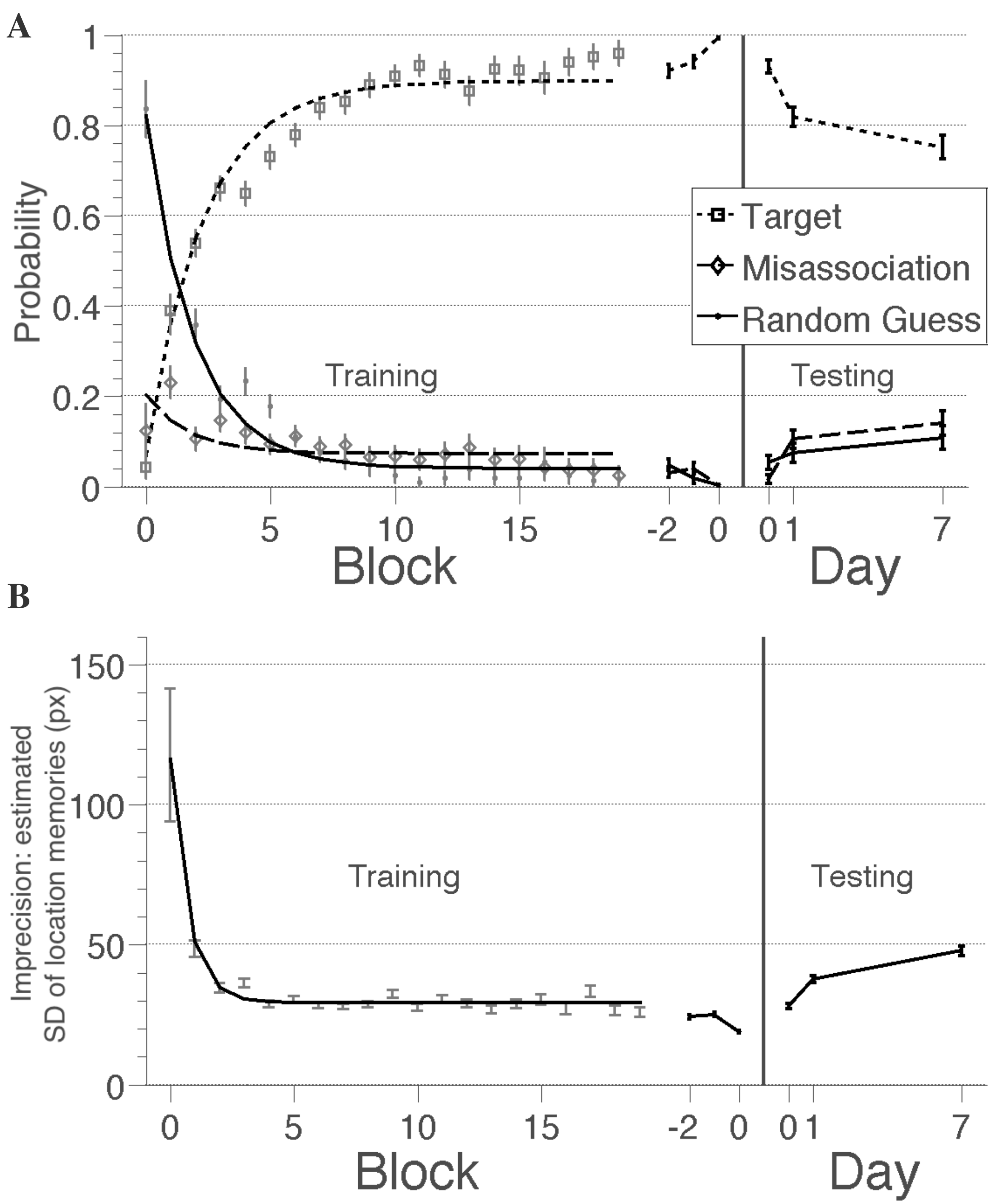


Figure 4. Estimated imprecision, misassociation and random guessing for Experiments 1 and 2. (A) The probability of selecting the target objects (dotted lines and squares),

misassociation (associating an object with the wrong location; dashed lines and diamonds) and randomly guessing (solid lines and dots) during the training blocks in Experiment 1, and testing blocks in Experiment 2. In the training blocks, the points show across-subject estimates of the different response types and the lines show exponential fits to those estimates. (B) The estimated imprecision (standard deviation) of remembered locations. Consistent with subjects' RMSE, the imprecision of locations, the probability of making misassociations and the rate of random guessing decreased during training and increased during testing. All error bars indicate posterior SD.

	Target	**Misassociation**	**Random guess**	**SD**
Initial ($\boldsymbol{A}$)	.07(.03–.11)	.20(.12–.28)	.82(.74–.90)	1116.2(78.2–170.9)
Asymptote ($\boldsymbol{B}$)	.90(.87–.93)	.07(.06–.09)	.04(.01–.06)	29.2(28.4–30.0)
Slope (τ)	2.3(1.7–3.0)	1.7(.49–3.0)	1.9(1.3–2.7)	.72(.50–1.06)

Table 1. Mean exponential fit parameters. SD indicates standard deviation. We fit the parameters using the exponential decay function $B + (A - B)e^{\frac{-t}{\tau}}$. A and B determine the initial and asymptotic values of the function and τ is the time constant (exponential slope) and t is time. Numbers in parentheses indicate the 95% PQI.

How did the sources of error change during learning? The imprecision of location memories, the probability of making a misassociation, and the probability of random guessing all decreased over the course of training (Figure 4, Training). To assess whether some aspects of memories were more quickly acquired, we quantified the speed with which these sources of error changed during learning by fitting exponential decay functions of the form $B + (A - B)e^{\frac{-t}{\tau}}$ to each parameter (Table 1). A and B indicate the initial and asymptotic values of the function (such that when A is greater than B the function will decrease over time), τ is a “time constant” and t is the block number. A larger time constant of the exponential decay function indicates a slower rate of change in a given parameter, and thus slower acquisition of this facet of memory during learning.

To estimate these parameters across subjects, we used a hierarchical model that assumes that the parameters for each subject are normally distributed around the population value, thus allowing us to efficiently pool estimates across subjects by using the statistics of the group to compensate for uncertainty in any one subject’s parameters. We fit the parameters using a Metropolis-Hastings algorithm (Metropolis, et al., 1953).

The time constant for the increasing rate of *correct* associations (*2.3, 95% PQI = 1.7–3.0*) was considerably larger than that for the decreasing imprecision of locations (*.72, 95% PQI = .50–1.06; 95% PQI on the difference between P(target) and SD time constants=.87–2.4*), indicating that subjects learned to associate objects to locations more slowly than they learned to accurately recall the exact positions of those locations. This pattern indicates that precise location memories are acquired quickly, but it takes some time to correctly associate them with their respective targets.

How did the sources of error change during forgetting? Although all sources of error increased during forgetting (Figure 4), misassociations, unlike noise and random guessing, increased abruptly after the first day. During testing, the standard deviation of location memories steadily increased from 28 to 38 to 48 pixels. The probability of random guessing remained constant over the first two days (*95% PQI on the difference between: test day-0 and test day-1 =-.076–.032*) and then increased somewhat by the final day of testing (*95% PQI on the difference between: test day-0 and test day-7 =-.11–.004; test day-1 and test day-7=-.10–.03*). In contrast, in the immediate post-training test, subjects made almost no misassociation errors (*1.6%*), but at the 1-day retention interval these jumped to 11%, and by day 7 had only increased slightly to 14% (*95% PQI on the difference between: test day-0 and test day-1 = .04–.14; test day-1 and test day-7 = -.10–.03*). When we directly compared changes in the rates of misassociation and random guessing, the probability of misassociations trended towards increasing more from test day-0 to test day-1 than the number of random guess (*95% PQI on the difference between: misassociations day-0 and day-1 and random guesses test day-0 and test day-1 = -.014–.15*), further suggesting that misassociations were exceptionally fragile early on during forgetting. While location memories steadily became less precise from the end of training and gradually became irretrievable, memories of associations were preserved in the immediate post-training test but deteriorated sharply after a single day.

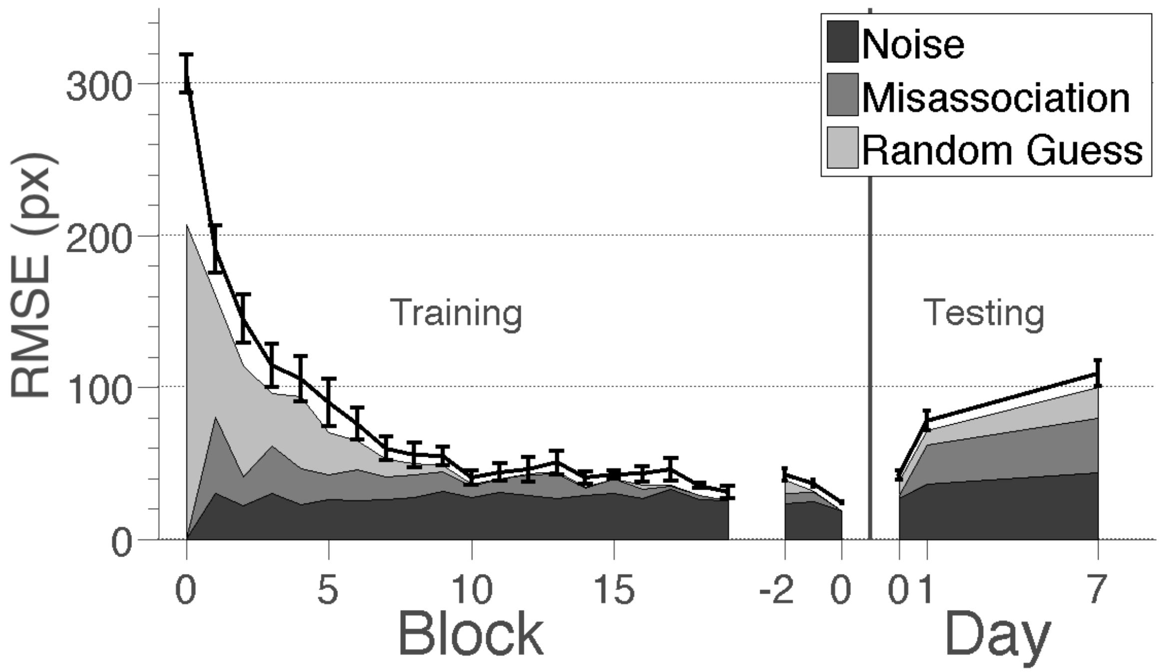


Figure 5. Learning curves from Experiment 1 and the forgetting curve from Experiment 2 with errors partitioned based on their estimated source. The black line indicates subjects' raw root mean square error (RMSE) (Identical to Figure 2). Shading indicates the estimated errors due to noise from recalled locations, misassociations and random guessing. Decreasing errors due to random guessing characterized learning while increasing errors due to misassociation drove forgetting.

How did errors contribute to performance during learning and forgetting? Based on the estimated probabilities of random guessing and misassociations, and the imprecision of location memories, we can infer how much each of these sources of error contributes to the overall RMSE at different points in time. To do so we use maximum likelihood estimation (MLE) to classify responses as noisy correct responses, noisy misassociations or random guesses. We then calculate the model's expected RMSE for each type of error given the parameter estimates (Figure 5). Specifically, for each response we calculated the error due to precision as the estimated standard deviation of location memories, the error due to misassociations as the distance between the target location and the location of the misassociated item (if applicable), and the error due to random guessing as the distance between the target location and the center of the environment (if applicable); we then aggregated these across items and subjects. The bulk of error reduction during learning arises from decreasing rates of random guessing as people learn the locations of objects, but the increased error during forgetting seems to arise from increasing misassociations as people retain the locations, but fail to map them onto the correct objects.

Experiment 3

In Experiments 1 and 2, we found that the precision of locations, and the ability to retrieve and correctly associate locations improved during learning and deteriorated during forgetting. Although all sources of error decreased with training and increased with forgetting, memories for associations were exceptionally unstable and contributed disproportionately to overall error during learning and especially forgetting.

One shortcoming of the cued recall task we used in Experiments 1 & 2 is that it can only reveal latent knowledge of locations that subjects have associated (either correctly or as an incorrect misassociation) with a cue. If a subject learned a location, but failed to match it with any of the potential retrieval cues, they may never produce that location in a cued response. Consequently, this latent knowledge might not be detectable, even in a model that can detect misassociations.

In Experiment 3 we aimed to directly measure knowledge of locations by asking subjects to report the locations in a two-step procedure: first in a free recall portion they reported all the locations they remembered, and then matched these locations to objects. Thus, like verbal paired associates tasks that aim to distinguish object and associative information (Tulving & Wiseman, 1975) this design removes the demand for correct associations during location recall, and might reveal latent location knowledge that was obscured in Experiments 1 and 2.

Subjects. A new set of subjects from the UCSD Psychology Department's online subject pool who did not overlap with the subjects from Experiment 2 participated in this 3-session experiment for payment. 74 subjects finished at least session 1, and 25 completed all three sessions. Subjects who completed all three sessions received a monetary bonus based on their performance.

Design. Experiment 3, like Experiment 2, was comprised of three sessions. In the first session subjects were trained to criterion. They were tested (without feedback) immediately after training (testing day-0), one day after training (test day-1) and seven days after training (test day-7).

The critical change introduced in Experiment 3 is the use of a free recall task that occurred after every two blocks (starting after block 1) during training and that replaced cued recall during testing. In this free recall task subjects reported all the locations they remembered, and then matched objects to those locations (see Procedure).

In further contrast to Experiment 2, we omitted the math distractor task between training and the immediate day-0 test. Additionally, rather than drop out individual objects during training (as in Experiment 2), subjects recalled the locations of all ten objects in each block until all were reported correctly (as in Experiment 1).

Stimuli. The objects were identical to those used in Experiments 1 and 2. We made minor aesthetic changes to the framing of the task: omitting the island cover story, replacing the central island with a fixation cross and changing the color of the background to white. To prevent locations from overlapping during the free recall task, we required the centers of objects to be located 120 px (2 objects) from each other. We also decreased the size of the environment to a radius of 275 px to allow room for the free recall task.

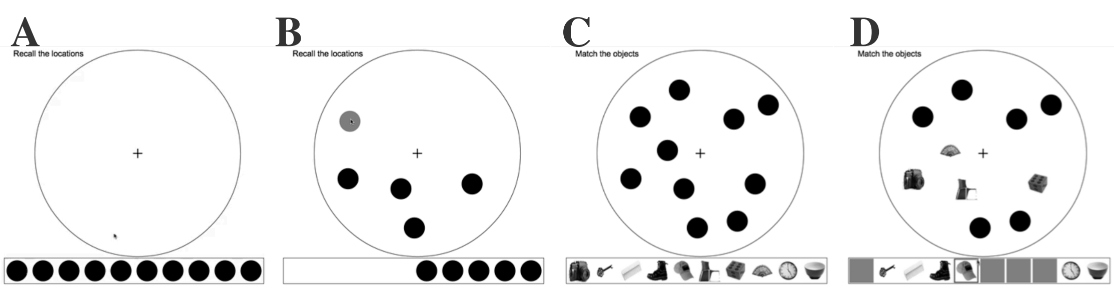


Figure 6. Example free recall trial from Experiment 3. (A) Subjects saw 10 black circles that would mark the locations of objects and (B) placed the circles wherever they recalled the location of an object. (C) Once subjects placed all ten circles, they saw all 10 objects in a random order and (D) matched the objects to the locations. Subjects had unlimited time to do the location recall and object matching phases and could rearrange the locations and object-to-location assignments as much as they wanted.

Procedure. In session 1, subjects recalled the location of a cued object and received feedback, as in Experiment 1. We interleaved these training blocks with a free recall phase (Figure 6). Free recall occurred after the first block and every two blocks afterwards. During the free recall phase, subjects saw 10 black circles at the bottom of the screen, and were instructed to place those (by clicking and dragging) at the locations of the ten objects. They could rearrange the placed circles as much as they desired. Once subjects indicated that they were done placing the circles, they saw all 10 objects on the bottom of the screen, and matched the objects to their locations by clicking on an object and then a location. They had unlimited time to perform the location recall and object matching subtasks, and they received no feedback at the end of free recall and matching. During testing, subjects reported the locations of the objects using the free recall task instead of the cued recall task.

Results

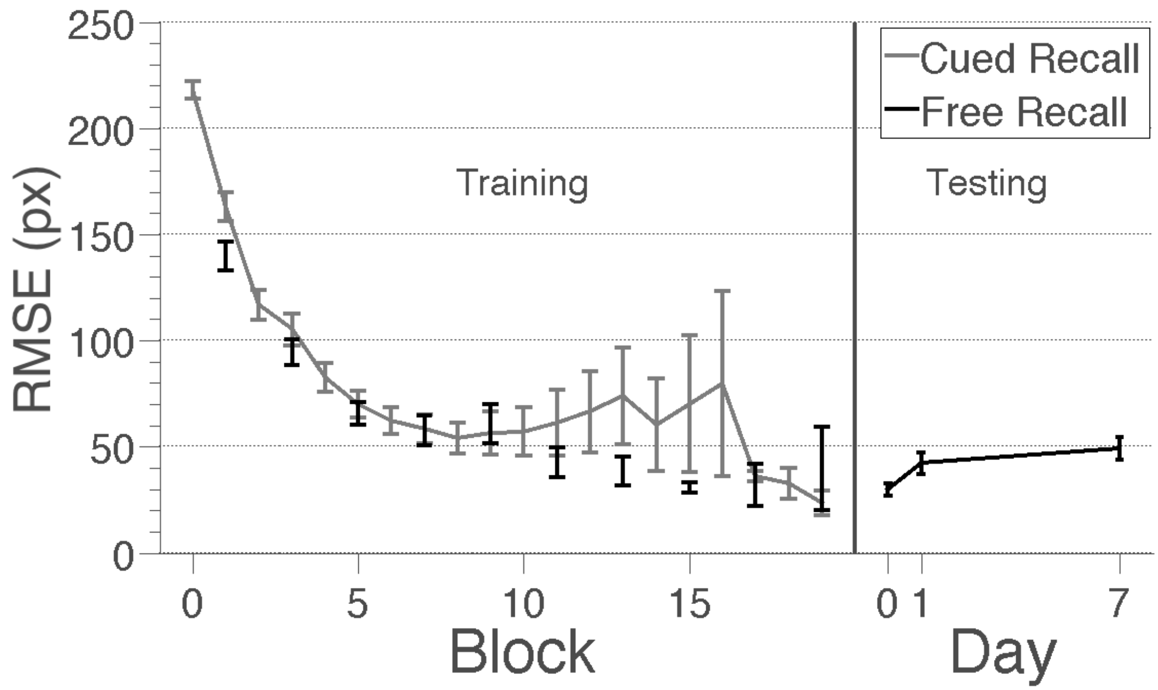


Figure 7. Learning and forgetting in Experiment 3 measured in root mean square error (RMSE) for the cued recall task during training and the free recall task during training and testing. The grey line indicates cued recall performance and the black points and line indicate free recall performance. Training results reflect all 74 subjects that finished session 1 and testing results reflect the 25 subjects that finished all three sessions. Cued and free recall performance during training were very similar. Error bars indicate ±1 SEM across subjects.

Did subjects learn and forget the locations of objects? As in Experiments 1 and 2, subjects learned the locations of objects during training and forgot them during testing (Figure 7). During training, the cued and free recall performance of each subject in each block was strongly correlated ($r=.72, p<.001$), indicating that both tasks adequately evaluate memory. We used a mixed effects model to test whether subjects performed better in the free recall vs. cued recall task, treating task type, block and their interaction as fixed effects and subjects as random effects. Subjects performed better in the free recall task (*t(600)=3.84, p<.001*), perhaps because this task discourages random guessing and encourages misassociations or allows subjects to choose the order in which they recall the objects (e.g., strongest items first). Additionally, this improvement significantly interacted with block number (*t(600)=2.98, p=.002*), reflecting subjects learning associations for the cued recall task over time.

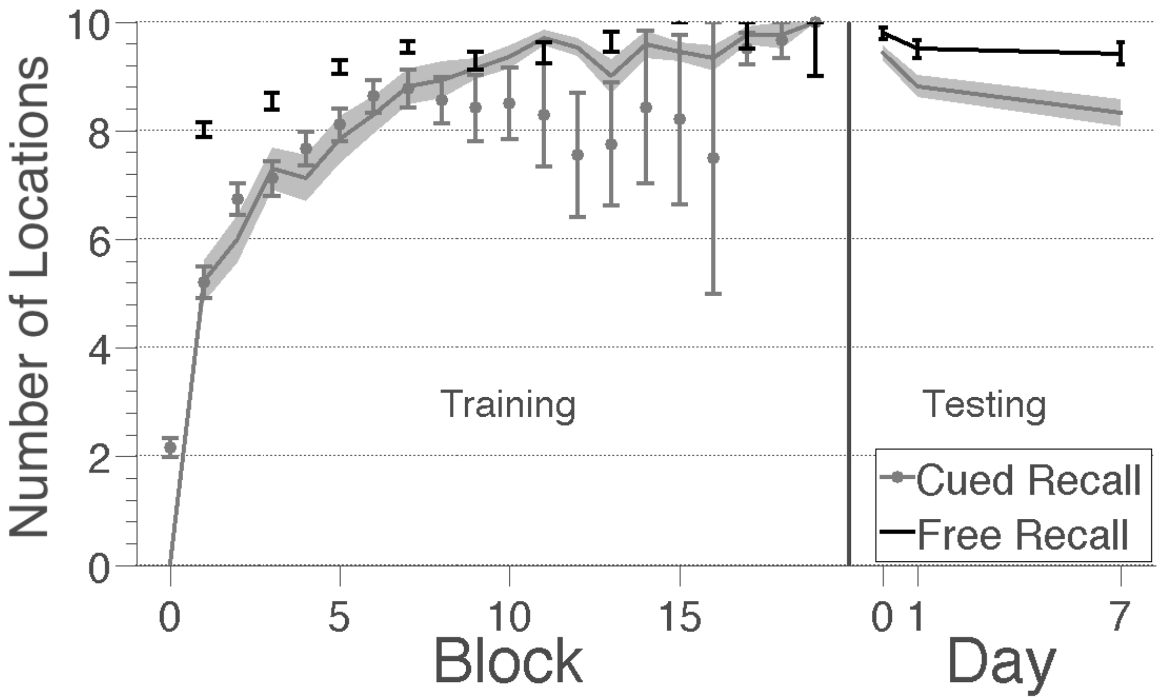


Figure 8. Estimated number of unique object locations recalled as either correctly associated responses or misassociations during training and testing. Grey points are the number of locations recalled in blocks of the cued recall task matched to the free recall blocks. Black points indicate the number of locations reported in the free recall task. For comparison, the shaded grey areas show the number of locations recalled in single blocks from Experiments 1 (Training) and 2 (Testing). Subjects correctly recalled more locations

in the free recall task than in cued recall tasks. Error bars indicate ±1 SEM across subjects.

How did subjects learn and forget locations separate from associations? We used our error model to obtain MLE estimates of the number of unique locations recalled (i.e., locations that were classified as either correctly associated or as misassociations) during the cued recall and the free recall task (Figure 8). For comparison, we also determined the locations recalled during cued recall in Experiments 1 and 2. The training results reflect all 74 subjects who completed session 1 and the testing results reflect the 25 subjects who completed all three sessions. To compare the number of locations recalled across tasks, we again used mixed effect models, treating task type, block and their interaction as fixed effects and subjects as random effects. The number of locations recalled during cued recall was similar to Experiment 1, suggesting that including the free recall task did not change how subjects learned the locations of objects.

During training, subjects recalled more locations when using free recall than when using cued recall ($t(600)=9.67, p<.001$). For instance, after the first block, subjects recalled on average 8.0 (*SEM=.13*) of the 10 locations during free recall compared to 5.2 (*SEM=.28*) during cued recall. There was also a significant interaction between task type and block number ($t(600)=7.52, p<.001$), reflecting subjects learning the associations between objects and locations and consequently recalling locations increasingly accurately during cued recall.

By comparing the number of locations recalled during free recall in Experiment 3 to cued recall performance in Experiment 2, we could directly assess the contribution of lost associations to apparent forgetting. In the immediate post-training test (testing day-0), we found that the number of locations recalled during free recall trended towards being greater than the number of locations recalled during cued recall; nevertheless they did not significantly differ ($t(60)=1.89, p=.06$). However, there was a significant interaction between task type and block number ($t(179)=2.06, p=.041$), indicating that subjects performing free recall increasingly recalled more locations than subjects performing the cued recall task. Altogether, over delays up to a week subjects appear to remember the locations they learned, but forget the objects to which those locations correspond. During recall, this loss of associations can result in subjects either making misassociations or randomly guessing.

Experiment 4

In the previous experiments, we found that forming and maintaining associations were the main factors limiting long-term visuospatial memory for locations. Is this also true for verbal memory? On one hand, both visual and verbal memory exhibit classic memory phenomena like a benefit to retention from spaced practice (visual: Paivio, 1974; verbal: Ebbinghaus, 1913) as well as advantages from primacy and recency (visual: Hollingworth, 2004; verbal: Ebbinghaus, 1913). So we might expect that forgetting operates similarly for both types of memory. On the other hand, visual and verbal working memory seem to rely on mechanisms dissociable with interference tasks (Baddeley & Hitch, 1978) and there are discrepancies in the magnitude of recency effects for auditory and visual information (Murdock & Walker, 1969; Madigan, 1971), so perhaps forgetting would also operate differently. In Experiment 4 we assess the contributions of imprecision, misassociation, and wholesale forgetting to long-term memory errors during learning and forgetting for verbally presented qualities. Specifically, we aimed to assess whether verbal memory follows a similar pattern of deterioration as visuospatial memory by training subjects on numerical values: the "great circle" distance between pairs of cities. Furthermore, we extended the delay period to examine forgetting over even longer periods of time.

Methods

Subjects. 24 subjects recruited through our online subject pool participated in this 4-session experiment for payment with an additional monetary reward for good performance.

Design. Subjects participated in one training session followed by three testing sessions. In the first session, subjects were trained on 24 facts. Like in Experiment 2, within each block the order of the facts was randomized and facts dropped out when they were recalled accurately. At the end of the first session, subjects recalled all 24 facts (testing week-0). Subsequent testing sessions occurred 1 week (testing week-1), 2 weeks (testing week-2) and 4 weeks (testing week-4) following the training session. To control for testing effects, of the 24 facts, 6 were presented on all three testing sessions, while the other 18 appeared in only one testing session (6 in each of the three testing sessions). Thus, in each testing session participants were probed on 12 facts: 6 that were tested in every session, and 6 unique to that session.

Stimuli. Subjects learned 24 distances[‡] between pairs of cities. The distances were the great circle distances (the shortest distance between two points on a sphere). For example, subjects would learn that the distance between Amsterdam, Netherlands and Athens, Greece is 1343 miles. Henceforth, we report the $\log_{10}$ distances[§]. The mean log distance was 3.6, with a standard deviation of .35.

Procedure. In session 1, subjects trained on 24 city-distance pairs over multiple blocks. On every trial, subjects saw two city names and reported the great circle distance between those cities; subjects then received feedback with the correct distance. Thus, in the first block, every response was a guess informed only by subjects' prior geography knowledge, but in subsequent blocks, subjects would have learned from the feedback. As in Experiment 2, subjects were trained to criterion with dropout; specifically, after subjects reported the distance for a particular city-pair correctly (within 1%) once, that item was excluded from subsequent training blocks.

In each test session, subjects recalled 12 of the distances (see Design) but did not receive feedback.

Results

[‡] Subjects chose whether the distances were in miles or kilometers. Here all distances are presented in miles.

[§] Analysis in log space respects the Weber-law like noise pattern common to magnitude, number and length estimation.

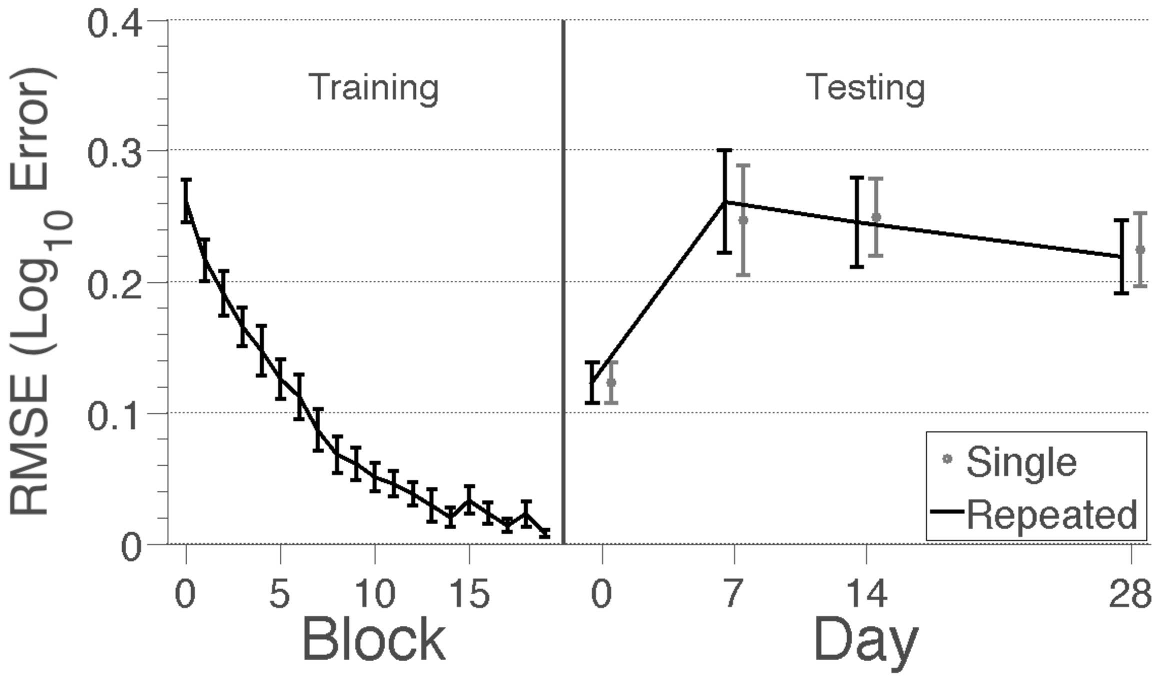
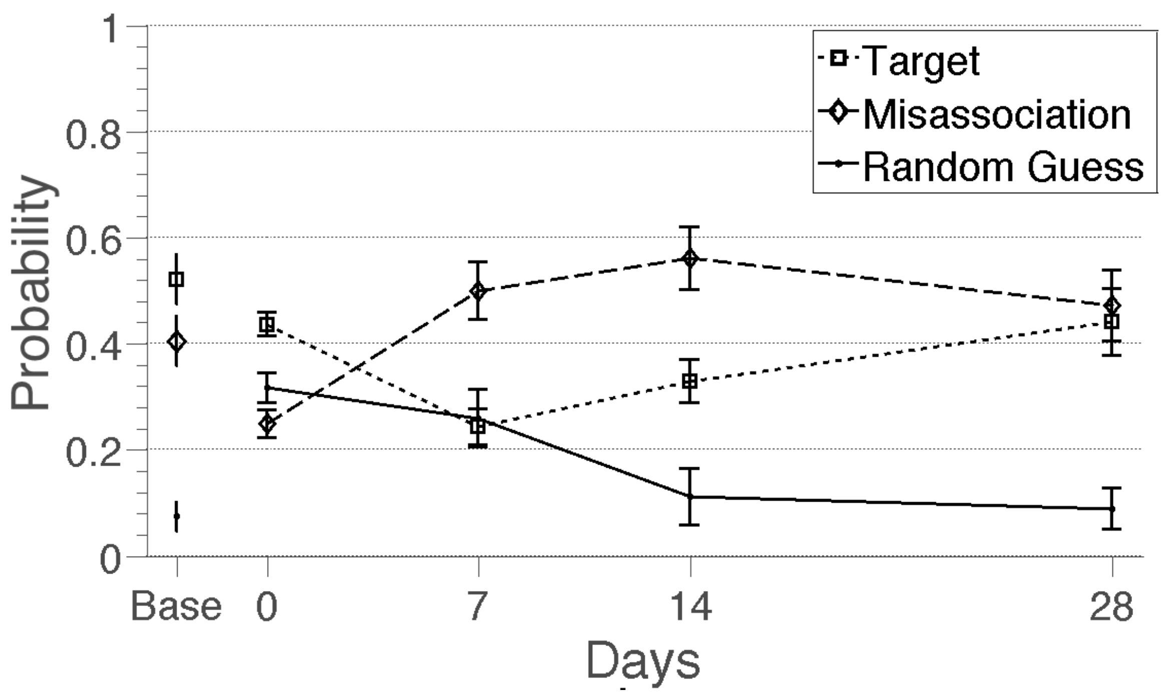


Figure 9. Learning and forgetting curves for Experiment 4. Error was measured in log_{10} root mean square error (RMSE). The first 20 blocks of training is left (in Blocks) and testing is right (in Days). Because subjects completed training in different numbers of blocks, we imputed their results for subsequent blocks in the learning curve to avoid misrepresenting errors in later blocks (our analyses do not rely on these imputed values). For testing, the continuous black lines indicate facts that were tested every session and the grey points indicate facts that were only tested in that session. Subjects learned the locations during training and appeared to return to baseline after one week. Error bars indicate ±1 SEM across subjects.

Did subjects learn and forget the facts? Subjects' raw performance (as measured by the RMSE of their log-transformed responses) improved throughout training and deteriorated during the testing sessions (Figure 9). Training took on average 17.21 blocks (*SEM=.24*).

Subjects forgot the facts quickly such that RMSEs in testing weeks 1, 2 and 4 were indistinguishable from the first training block (*mixed effect model treating block as a fixed effect and subject as a random effect, main effect of block: t(94)=1.58, p=.11*)[**]. There were also no discernible differences in RMSE for facts recalled during repeated testing sessions vs. only during individual testing sessions (*mixed effect model treating task, block and their interaction as fixed effects and subject as a random effect, main effect of task: t(140)=.037, p=.97; interaction between task and block: t(140)=.64, p=.52*): thus we pool them in subsequent analyses.

A



B **C**

[**] This rapid forgetting compared to the previous experiments may reflect any of a number of differences between the experiments: e.g., the larger number of items, the lower training criterion, or differences in associating city pairs with continuous numbers.

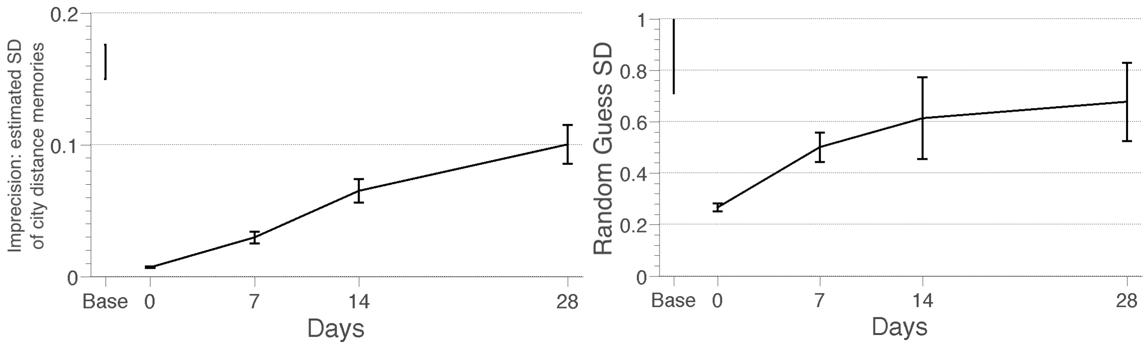


Figure 10. Estimated noise, misassociation and random guessing for Experiment 4. (A) Estimated probability of selecting the target (dotted lines and squares), making a misassociation (dashed lines and diamonds) and randomly guessing (solid lines and dots). (B) The standard deviation (SD) of recalled facts. (C) The standard deviation of random guesses around the mean distance. In these graphs, the first block of training acts as baseline performance (Base). The continuous lines indicate performance during the four testing blocks. During forgetting, subjects remembered many distances precisely but associated them to the wrong city pairs. Error bars indicate posterior SD.

How did the sources of error change during learning and forgetting? We fit the error model to subjects' responses to estimate the sources of errors in the first training block and the four testing blocks (Figure 10). Objects dropping out during training prevented us from analyzing the other training blocks.

In the first training block, a combination of imprecise prior knowledge, and mutual information across items (e.g., learning the distance between Amsterdam and Greece may bias estimates of the distance between Berlin and Ankara) precluded any decisive analyses of error contributions. Specifically: responses were frequently characterized as recalled target distances or as misassociations, despite this being the first training block. These responses may have reflected subjects' imprecise prior knowledge of geography since these apparently informed responses had very low precision (*.16, 95% PQI = .14–.19*), or may correspond to subjects making responses based on feedback they received in previous trials of the same block. In short, people started out training with vague ideas about city-pair distances and their relationships.

In the immediate post-training test (testing week-0), subjects recalled the locations precisely (*.0069, 95% PQI = .0060–.0079),* and made few misassociations (*.25, 95% PQI = .20–.30*), consistent with their overall low RMSE in this immediate test. RMSE in testing sessions at 1-4 week delays suggests that subjects returned to their baseline pre-training performance after just a one-week delay. At face value, this could indicate that subjects forgot everything they learned and reverted to randomly guessing based on their prior knowledge. On the other hand, the high RMSE might instead reflect subjects making many misassociations, which would indicate that subjects actually retained accurate memories of facts, but not associations between city pairs and distances.

Indeed, the high RMSEs in testing weeks 1, 2 and 4 seem to be caused by very high rates of precisely reported, but incorrectly associated, distances. For instance, testing week 1, distance imprecision was just .030 (*95% PQI = .023–.039*), compared to .16 (*95% PQI = .14–.19*), in the first training block (*95% PQI on the difference between: baseline and day-7=.11–.16*) demonstrating that facts are being remembered precisely. Overall RMSE is indistinguishable, however, due to a 49% (*95% PQI = 38–60%)* misassociation rate. Similarly, the precision of correctly and incorrectly associated distances after 2 weeks (*.064, 95% PQI = .049–.083*) and 4 weeks (*.10, 95% PQI = .072–.13*) is better than baseline (*95% PQI on the difference between: baseline and* *day-14 = .067–.13; baseline and day-28 = .024–.10*), but this latent knowledge is not evident in RMSE due to high misassociation rates (*day-14: 56%, 95% PQI = 44 to 67%; day-28: 47%, 95% PQI = 34–60%)*. Thus, it seems that verbal numerical memory for city-pair distances–like memory for object locations–is primarily hampered by misassociations, so much so that they obscure relatively precise, and stable, latent knowledge of learned distances when considering overall measures of error.

General Discussion

Previous work has primarily evaluated the acquisition and loss of information in long-term memory by using binary measures such as "recalled versus not-recalled". These studies have documented long-term memory's large capacity and temporal stability. Here, we examined the mechanisms of forgetting in a finer grained manner, asking how noise, misassociations and complete loss of memory traces contributed to declines in memory performance over time. Consistent with previous characterizations of long-term memory, we found that verbal and visual long-term memory representations were extremely robust over long delays and that visual long-term memories formed very quickly. The chief limitation on long-term memory—apparent in both acquisition and forgetting—was a difficulty forming the correct associations and maintaining those associations over time. Accordingly, our comparison of performance in cued and free recall tasks suggests that the free recall task helped disentangle memories of locations and associations, allowing us to more accurately assess the contents of visual memory.

Learning and forgetting in long-term memory

We show that although long-term memory is impressive in its ability to retain precise facts, it is strikingly limited in its ability to form and recall associations between memories. These results are consistent with earlier investigations of verbal long-term memory demonstrating that the recency effect deteriorates much more rapidly for paired associates (Murdock, 1967) than for individual items (Murdock & Kahana, 1993). This may reflect associative information being fragile or memories interfering with each other (Briggs, 1954; Barnes & Underwood, 1959; Underwood, 1957).

We find that misassociations drive forgetting in long-term memory and, to a lesser extent, these memories become less precise over time. In contrast, Brady et al. (2013) found that long-term memories exist in a constant, low-fidelity state and spontaneously give way to random guesses. Although seemingly in conflict, these two sets of results may actually be quite consistent. Our subjects were trained to criterion, while the subjects trained by Brady et al. saw stimuli only briefly. Consequently, long-term memories in Brady et al. may have never gained enough precision to yield detectable losses. Moreover, because Brady et al. could not estimate misassociations, such responses would have appeared as random guesses in their data. Thus, both sets of results are consistent with misassociations being the primary cause of forgetting.

Comparison to visual working memory

Our finding that during learning and forgetting subjects often knew locations but did not associate them is somewhat similar to previous findings that visual working memory represents (Vul & Rich, 2010) and forgets (Fougnie & Alvarez, 2011) the features of objects independently, and that the appropriate binding (association) of these features is fragile over time (Gorgoraptis, et al. 2011). The difficulty of binding features together in visual working memory and the associative limits of visual long-term memory may reflect a common limitation on our ability to correctly associate features together.

When we removed the need to associate locations with objects in the free recall task, we found subjects recalled many more locations than during parallel cued recall tasks. Similarly, using different stimuli and memory probes in working memory experiments can affect the difficulty of recalling associative information. Stimuli with dependent integral features (Fougnie & Alvarez, 2011; Bae & Flombaum, 2013) or that do not suffer from proactive interference (Endress & Potter, 2014) result in larger estimates of visual short-term memory capacity. Likewise, probing memory using a two-alternative forced-choice task instead of a same-different task can make it more difficult to keep track of associations (Makovski, et al., 2010). Varying the distinguishability of stimuli and the method of recall may help determine when visual working memory is limited by observers' ability to recall features vs. the associations between them.

Limitations

We treated the free recall and cued recall tasks in Experiment 3 as comparable tasks, differing only in how subjects recalled locations. However, the tasks may have encouraged subjects to encode the objects differently. Simultaneous report (as in the free recall task) compared to sequential report (as in the cued recall task) may have encouraged subjects' to encode objects based on their "ensemble statistics" (Chong & Treisman, 2005; Brady & Alvarez, 2011). Using such statistics may have even helped subjects remember the objects more accurately (Orhan, et al., 2014). Although free recall helped us assess subjects' memories of unassociated and/or incorrectly associated locations, whether the free recall task introduced differences in performance requires further investigation.

Additionally, recall performance may have been hindered by the lack of natural structure in our task. Memory relies on prior expectations (Bartlett, 1932) and using real-world priors can impair recall when those priors are inconsistent with structure in the experiment (Orhan & Jacobs, 2014). In Experiments 1-3, for example, subjects could have expected the hat and boot to be close together (because both are articles of clothing), conflicting with the actual randomness of locations in the experiment. In contrast, using stimuli that are structured consistently with subjects' prior expectations improves the fidelity of memories (Orhan, et al., 2014). If the structure of the stimuli in our task was consistent with subjects' prior expectations, subjects may have exhibited different patterns of learning and forgetting.

Implications

Instead of passively observing stimuli during training, in our study subjects reported locations/distances and received feedback. Many studies have shown that different training manipulations such as spacing presentations (see Cepeda, et al., 2008, for a review), review through testing rather than restudy (Bjork & Bjork, 1992; Roediger & Karpicke, 2006) and allowing self-directed learning (Markant & Gureckis, 2014) can aid the formation and long-term survival of memories. Asking how these different training techniques affect the sources of people's error may help reveal the mechanisms that these techniques rely upon and the associative limitations of long-term memory.

Conclusions

We described a number of experiments designed to assess the contributions of imprecision, misassociation, and the absence of relevant memory traces in memory to limited performance in learning and forgetting. When remembering visual and verbal stimuli, people quickly formed fairly accurate memories for scalar quantities (locations and distance), with this precision decaying only minimally over time. In both cases, however, associations between those memories were learned slowly and were readily lost over time.

References

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *The Journal of Neuroscience*, *31*(3), 1128-1138.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, *8*, 47-89.

Bae, G. Y., & Flombaum, J. I. (2013). Two Items Remembered as Precisely as One How Integral Features Can Improve Visual Working Memory. *Psychological Science*, *24*(10), 2038-2047.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. New York, NY: Cambridge University Press.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*(2), 97-105.

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10), 6-21.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851-854.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622-1631.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35-67.

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.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325-14329.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*(6), 981-990.

Briggs, G. E. (1954). Acquisition, extinction, and recovery functions in retroactive inhibition. *Journal of Experimental Psychology*, *47*(5), 285-293.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning a temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095-1102.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1-13.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (No. 3). New York, NY: Teachers College, Columbia University.

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General*, *143*(2), 548-565.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12):3, 1-12.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721-741.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: the roles of short-and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 519-537.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347-356.

Madigan, S. A. (1971). Modality and recall order interactions in short-term memory for serial order. *Journal of Experimental Psychology*, *87*(2), 294-296.

Makovski, T., Watson, L. M., Koutstaal, W., & Jiang, Y. V. (2010). Method matters: systematic effects of testing procedure on visual working memory sensitivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1466-1479.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94-122.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092.

Murdock, B. B. (1967). Recent developments in short-term memory. *British Journal of Psychology*, *58*(3-4), 421-433.

Murdock, B. B., & Kahana, M. J. (1993). List-strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 1450-1453.

Murdock, B. B., & Walker, K. D. (1969). Modality effects in free recall. *Journal of Verbal Learning and Verbal Behavior*, *8*(5), 665-676.

Orhan, A.E., & Jacobs, R.A. (2014). Toward ecologically realistic theories in visual short-term memory research. *Attention, Perception, & Psychophysics*, *76*(7), 2158-2170.

Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The adaptive nature of visual working memory. *Current Directions in Psychological Science*, *23*(3), 164-170.

Paivio, A. (1974). Spacing of repetitions in the incidental and intentional free recall of pictures and words. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 497-511.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Tulving, E., & Wiseman, S. (1975). Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society*, *6*(1), 79-82.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49-60.

Vul, E., & Rich, A. N. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science*, *21*(8), 1168-1175.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology*, *3*(2), 316-347.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233-235.

Acknowledgements

This work was also supported by the Office of Naval Research (MURI Grant #N00014-10-1-0072), by a collaborative activity grant from the James S. McDonnell Foundation, and by the National Science Foundation (Grant SBE-582 0542013 to the UCSD Temporal Dynamics of Learning Center).

Appendix

Appendix A: Model overview

We used a finite mixture model similar to that used in Bayes, et al. (2011) to estimate the precision of memories and the probability of responses reflecting misassociations and random guessing. Formally, we are interested in estimating three parameters: the probability of selecting the target object (p_T), the probability of making a misassociation (p_M), and the imprecision of correct responses and misassociations around remembered features (σ). The probabilities of selecting the target object and making a misassociation determined the probability of random guesses ($p_R = 1 - p_T - p_M$). Thus, the basic mixture-model likelihood of reporting a particular feature, y, for a particular item t out of n items total is:

$$P\left(y\middle|t\right) = p_T N\left(y\middle|x_t, \sigma\right) + p_M \left(\frac{1}{n-1}\right) \sum_{i \neq t} N(y|x_i, \sigma) + (1 - p_T - p_M) R\left(y\right)$$

where x_i is the feature value for item i and $\sum_{i \neq t}^{a}$ |denotes a sum over all the non-target items (candidate misassociations). Thus, the probability of making a misassociation is evenly split among all the items that are candidate misassociations. $N(y|m, s)$ denotes the density at a of a normal distribution with mean m and standard deviation s and $R(y)$ indicates the likelihood of randomly guessing y.

We modified the likelihood of random guessing ($R(y)$) in two ways to reflect the specific structure of our tasks. First, for Experiments 1-3 we modeled the distribution of random guesses as a two-dimensional Gaussian distribution around the mean feature value (the center of the environment) truncated by the borders of the environment. In Experiment 4, we used a one-dimensional Gaussian distribution centered on the mean feature value (the average $\log_{10}$ distance between cities) but because log distances are unbounded we did not truncate the distribution. In contrast, many prior studies using mixture models use a uniform distribution for random guesses (e.g., Zhang & Luck, 2008). In those cases, the feature values are often circular (e.g., hue angle) and thus have no natural "center". However, in both of our tasks, there is a natural center (either the center of the display, or the average distance) to which random guesses may be drawn to minimize expected errors. The truncated two-dimensional Gaussian and unbounded one-dimensional Gaussian likelihood functions offer a convenient way to parameterize between these random guessing strategies. With large standard deviations, these distributions will behave like a uniform distribution and with a small standard deviation will resemble responses around the central value.

In Experiments 1-3, we set the standard deviation of random guesses (σ_R) to the empirical standard deviation of all responses (we discuss this decision in later in Appendix D, Dispersion of random guesses). In Experiment 4, we estimated the standard deviation of random guesses (σ_R), just as we estimated the standard deviation of recalled locations (σ). We fit these parameters differently across experiments because the range of possible locations in Experiments 1-3 was constrained by the border of the environment but in Experiment 4 subjects' estimates of the range of possible distances changed over time.

Second, in Experiments 1-3, in addition to subjects selecting random values around the mean, we accounted for two other types of random guessing. When first learning the locations of the objects, subjects often either clicked the same location repeatedly or clicked the location of the preceding object. The first clearly does not reflect an attempt to recall the cued object's location. The second could indicate an attempt to correctly recall the cued object's location. However, given that the order of presentation was block randomized and that it is unlikely subjects forgot the correct object-location association over the course of a single trial, in these trials subjects most likely reported the wrong location intentionally. Our decision to account for these additional types of random guessing was supported by alternate forms of random guessing having a shorter response time than randomly guessing around the center of the environment (*mixed effect model treating error type as a fixed effect and subject as a random effect, main effect of error type:* $t(1667)=3.4$, $p<.001$). Consequently, we account for both types of responses and classify them as random guessing.

We extend our random guessing process to account for responses based on the previous response or feedback by treating them as responses centered on the previous response or previous object, respectively, with small standard deviations (σ_o). This introduces one additional parameter that describes the probability of random guesses broadly distributed around the center (p_{R1}) and the probability of structured random guesses ($1-p_{R1}$). ($1-p_{R1}$) is evenly split between the two types of structured random guessing. Thus the probability that responses are broadly distributed random clicks around the environment will be $(1-p_T-p_M)p_{R1}$; the guesses that are repeated clicks of the previous response, or repetitions of the previously presented location, will both be $\frac{(1-p_T-p_M)(1-p_{R1})}{2}$. In the main paper, we report the probability of random guesses as $(1-p_T-p_M)=p_R$.

We vary the random guessing parameters based on the constraints of the different tasks in our experiments. In Experiment 1 and cued recall in Experiment 3, when structured forms of random guessing were most likely to occur, we estimate p_{R2}. In Experiment 2 (where subjects know the locations), free recall in Experiment 3 (where subjects cannot use a structured form random guessing) we set p_{R2} to zero.

For Experiments 1-3, we modified random guessing to use a truncated two-dimensional Gaussian distribution and to account for additional forms of random guessing results in the likelihood of random guessing, $R(y)$ becoming:

$$R(y) = (1 - p_T - p_M)\, p_{R1} \Phi\left(y \middle| \mu_R, \sigma_R, r\right) + \frac{(1 - p_T - p_M)(1 - p_{R1})}{2} N\left(y \middle| x_{resp}, \sigma_o\right) + \frac{(1 - p_T - p_M)(1 - p_{R1})}{2} N\left(y \middle| x_{obj}, \sigma_o\right)$$

where $\Phi(a|m, s, b)$ indicates the density at a of a truncated normal distribution with mean m, standard deviation s and bound b. μ_{R}, σ_{R} and r indicate the center of the environment, the empirical standard deviation of responses and the radius of the environment, respectively. x_{resp} indicates the previous response, x_{obj} indicates the previously presented stimuli (in the first trial, the previous response/stimuli was substituted with the mean value) and σ_{o} is the standard deviation of responses around repeated responses/locations which we set to be very small ($\sigma_{\mathrm{o}} = 5\ px$).

Consequently, the full likelihood of reporting a particular feature, y, is:

$$P\left(y \middle| t\right) = p_T N\left(y \middle| x_t, \sigma\right) + p_M \left(\frac{1}{n-1}\right) \sum_{i \neq t} N(y | x_i, \sigma) + (1 - p_T - p_M)\, p_{R1} \Phi\left(y \middle| \mu_R, \sigma_R, r\right) + \frac{(1 - p_T - p_M)(1 - p_{R1})}{2} N\left(y \middle| x_{resp}, \sigma_o\right) + \frac{(1 - p_T - p_M)(1 - p_{R1})}{2} N\left(y \middle| x_{obj}, \sigma_o\right)$$

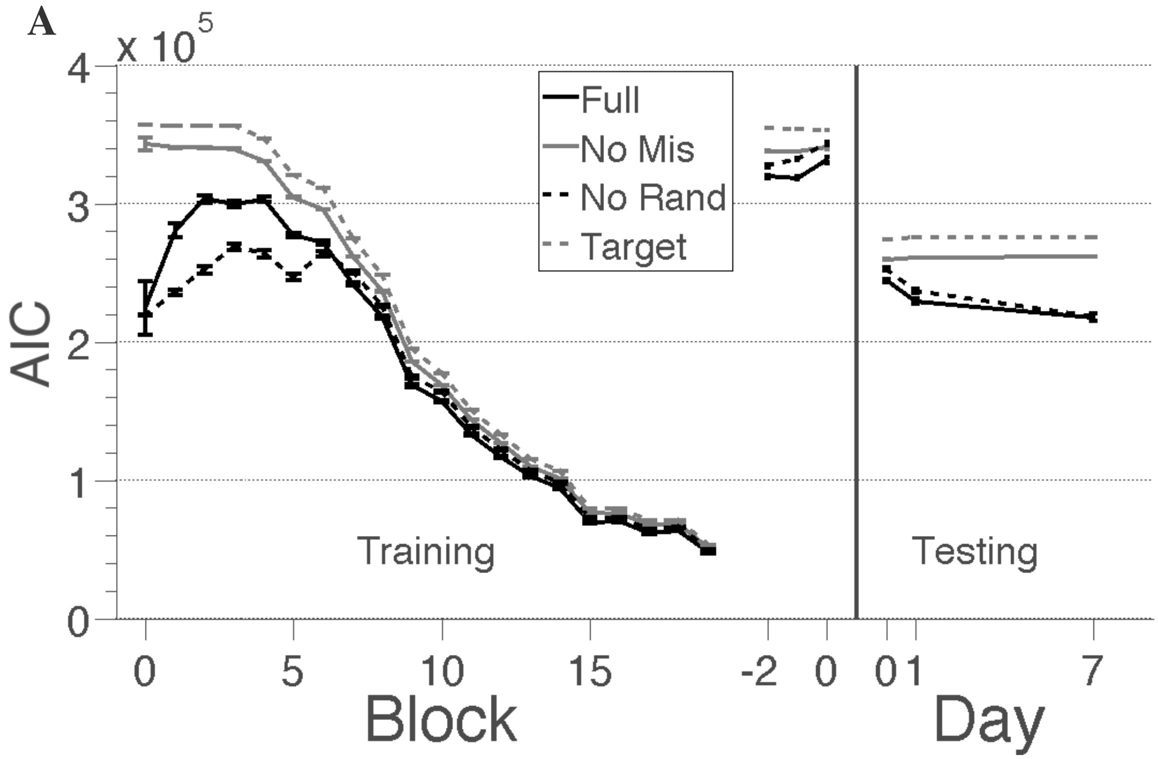
For Experiment 4, the random guessing likelihood is just a normal distribution; thus the complete likelihood function is:

$$P(y|t) = p_T N(y|x_t,\sigma) + p_M\left(\frac{1}{n-1}\right)\sum_{i\neq t} N(y|x_i,\sigma) + p_R N(y|\mu_R,\sigma_R)$$

where μ_R and σ_R indicate the mean distance between cities and the estimated standard deviation of random guesses (in log units), respectively.

For each block we fit the model across subjects using a Gibbs sampler (Geman & Geman, 1984). Our analyses of the parameter fits use 700 samples from the posterior (without thinning).

A



B

C

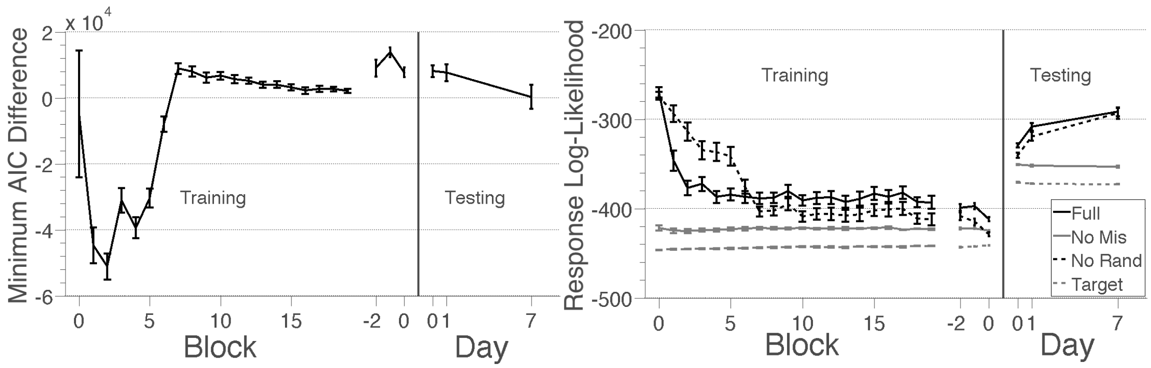


Figure A1. Model comparison of the mixture model with and without different types of errors for Experiments 1 and 2. Full is the full model (solid black line), No Mis is the model without misassociations (solid grey line), No Rand is the model without random guessing (dotted black line), and Target is the model with neither misassociations nor random guessing, making solely noisy guesses around the target object (dotted grey line). (A) Model fits as measured by Akaike Information Criterion (AIC). Smaller AIC values indicate better fits. Decreasing AICs during training reflect subjects completing the experiment and dropping out. (B) The difference in AIC between the full model and best fitting model that wasn't the full model. Differences greater than zero indicate the full model fit best (C) Model fits as measured by average log-likelihoods. Less negative log-likelihoods indicate better fits. Although the model without random guessing performs best during early training, the full model captures subjects' performance best in the rest of the study. AIC error bars indicate posterior SD, likelihood error bars indicate SEM.

Appendix B: Model comparison

In our analyses, we used a finite mixture model that captures errors due to noise, misassociations and random guessing. However, it is possible that the model falsely interpreted locations recalled very noisily as misassociations or random guesses. To examine whether subjects indeed made misassociations and random guesses, for Experiments 1 and 2 we tested how well mixture models without misassociations, without random guessing and without either type of error predicted subjects' responses. For each model, we calculated how well the model fit subjects' responses in each block or session, as measured by their Akaike information criterion (AIC) (Figure A1A). Smaller AICs reflect better model fits.

The full model fit subjects' responses well during training in Experiment 1 and testing during Experiment 2. To test when the full model provided the best fit, for each block/session we found the difference between the model with the smallest AIC (that wasn't the full model) and the full model (Figure A1B). Differences greater than zero indicate that the full model had a smaller AIC and was a better fit. The full model provided the best fit during the last 13 blocks of training in Experiment 1 and the first two sessions of testing in Experiment 2, indicating that subjects did indeed make misassociations and random guesses throughout our studies. Additionally, the model without random guessing but with misassociations performed much better than the full model during training early on and comparably during the final block of testing. The good fit of the model without random guessing demonstrates that possessing the correct associations was an important part of learning and forgetting.

Appendix C: Bayesian statistic reports

Several of our analyses report the posterior distributions of the parameters. Consider this example-"The time constant for the increasing rate of *correct* associations (*2.3, 95% PQI = 1.7–3.0*)". Here, 2.3 indicates the mean time constant. 95% PQI denotes the 95% Posterior Quantile Interval, such that 1.7 is the time constant at the 0.025 posterior quantile and 3.0 is the time constant at the 0.975 posterior quantile; and the posterior probability that the time constant falls within that interval is 95%. Because 95% of the sampled time constants fell above 0, this 95% PQI demonstrates that we can be confident that the time constant was positive.

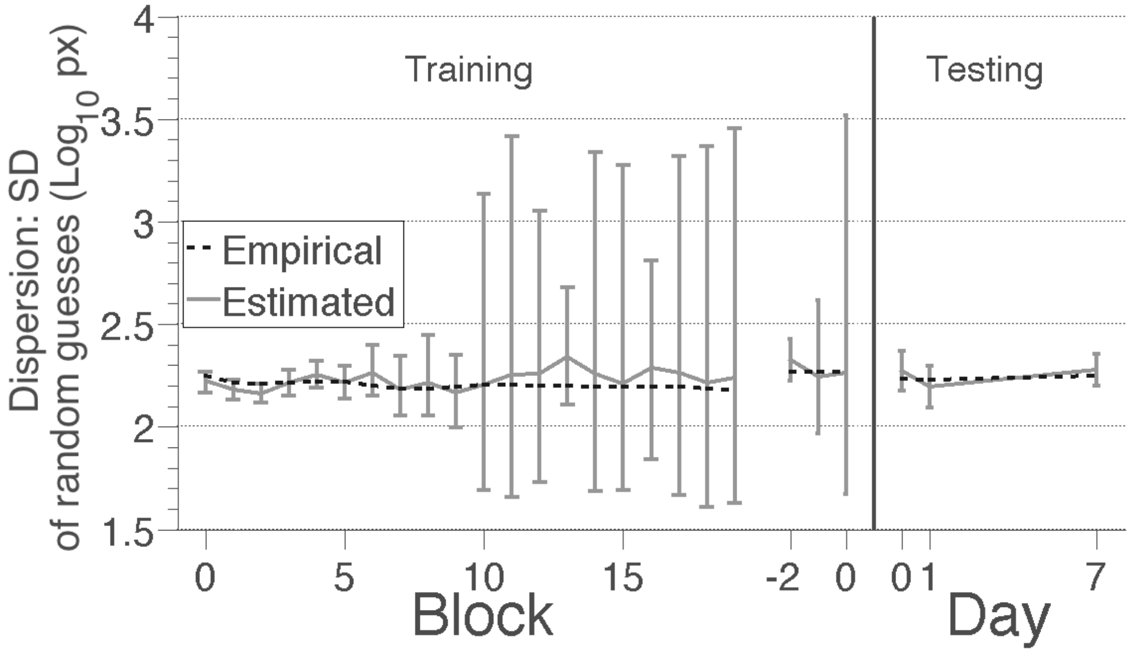


Figure A2. Empirical vs. estimated dispersion of random guessing. Empirical (black, dotted line) indicates the standard deviation of all of the subjects' responses around the center of the environment in each block/session. Estimated (grey, solid lines) indicates the model estimated standard deviation of random guesses around the center of the environment. The empirical standard deviation was generally a good approximation for the standard deviation of random guesses, and was far more stable, given that some blocks contained very few random guesses. Error bars indicate 95% PQI.

Appendix D: Dispersion of random guesses

In Experiments 1-3, we used the empirical standard deviation of subjects' responses around the center of the environment as the standard deviation of the truncated two-dimensional random guessing Gaussian distributions. However, because this calculation includes all reported locations (including those classified as correct reports and misassociations), it may have systematically overestimated the dispersion of random guesses. To examine whether the empirical standard deviation of responses was an accurate measure of the dispersion of random guesses, we modified our mixture model to estimate the standard deviation of random guesses and then compared the empirical and model estimated dispersion.

In every block and session, the empirical standard deviation fell within the 95% PQI of the dispersion estimated by the model (Figure A2), demonstrating that the empirical standard deviation was an accurate substitute for estimating the dispersion of random guesses explicitly. Moreover, since random guessing was relatively rare in later training blocks, explicit estimates of random guessing dispersion were highly unstable (as reflected by the very large 95% posterior intervals). In contrast, using the empirical standard deviation of responses yields a consistent, and stable estimate throughout the training session.

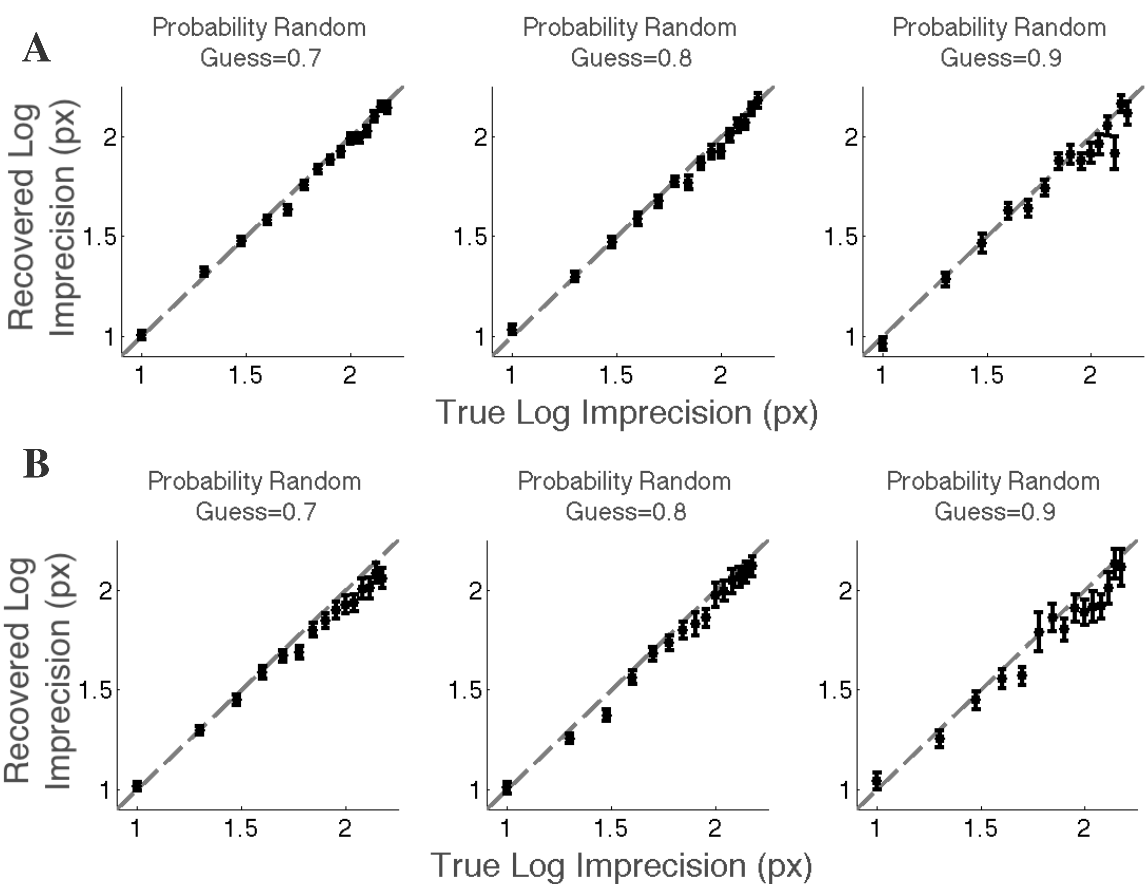


Figure A3. Imprecision parameter recovery when responses are A) a mixture of correct target selections and random guesses and B) a mixture of misassociations and random guesses. From left to right, each panel indicates mixtures with increasing proportions of random guesses. For example, in A3A, the panel "Probability Random Guess=.9" indicates that the probability of selecting the correct object was .1 and the probability of randomly guessing was .9. Each black point indicates the true log imprecision used to generate responses (X-axis) and the log imprecision estimated by the model (Y-axis). Dashed grey lines indicate equality. The model was able to consistently recover the imprecision of responses, even under very high levels of random guessing.

Appendix E: Imprecision parameter recovery with high levels of random guessing

We used our model to estimate the probability of selecting the target object, making a misassociation, randomly guessing, and the imprecision of recalled locations. However, in early training blocks the small number of locations recalled as targets or misassociations may have undermined our ability to estimate the imprecision of locations. Furthermore, in such situations high levels of random guesses may have been interpreted as very noisy correct responses or misassociations, inflating estimates of imprecision.

To examine whether the model could accurately estimate the imprecision of responses, we generated artificial data by drawing samples from our model with different parameter values. We focused on parameter values with high levels of random guessing to best capture conditions during early training blocks. Half of our parameters sets had a high probability of random guesses and a small probability of correct target selections (Figure A3A). The second half had a high probability of random guesses and a small probability of making misassociations (Figure A3B). We then used the model to recover the parameter values used to generate the data. For simplicity, we kept p_{R2} to zero when generating samples and estimating parameters.

The model was able to successfully recover the parameters used to generate the artificial data. The true and recovered imprecision were highly correlated (*smallest r: r=.99, p<.001*), and deviated only slightly from the identity line, reflecting a slight tendency to underestimate imprecision when random guessing was common (*most* *regression slopes in the range [.95—.99], most deviant slope from 1 was =0.96, 95% CI=.95—.98*). Rather than inflate noise estimates, the model slightly underestimated the imprecision of responses (*largest slope: .987, 95% CI=.981—.994*); this underestimation may reflect exceptionally noisy responses being more likely to be interpreted as random guesses, when the base rate of random guessing is high. Together, these results suggest that the model was able to adequately recover the imprecision of responses even under high levels of random guessing.

Appendix F: Reaction times and response type

In Experiments 1 and 2, we examined how reaction times varied for selecting the target item, making a misassociation and randomly guessing. We used mixed effect models that treat error type as a fixed effect and subject as a random effect to test whether different types of errors had different response times. We found no effect of error type on reaction time in Experiment 1($t(4678)=1.2, p=.23$) and Experiment 2 ($t(1108)=.22, p=.84$).