Massive memory revisited: Limitations on storage capacity for object details in visual long-term memory

Corbin A. Cunningham,[1] Michael A. Yassa,[2] and Howard E. Egeth[1]

[1]*Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [2]*Department of Neurobiology and Behavior, University of California, Irvine, Irvine, California 92697, USA*

Previous work suggests that visual long-term memory (VLTM) is highly detailed and has a massive capacity. However, memory performance is subject to the effects of the type of testing procedure used. The current study examines detail memory performance by probing the same memories within the same subjects, but using divergent probing methods. The results reveal that while VLTM representations are typically sufficient to support performance when the procedure probes gist-based information, they are not sufficient in circumstances when the procedure requires more detail. We show that VLTM capacity, albeit large, is heavily reliant on gist as well as detail. Thus, the nature of the mnemonic representations stored in VLTM is important in understanding its capacity limitations.

Humans are surprisingly good at remembering thousands of items, which have been presented only once and for a limited time, in visual long-term memory (Shepard 1967; Standing et al. 1970; Standing 1973; Vogt and Magnussen 2007; Brady et al. 2008; Konkle et al. 2010a,b). Many of these findings have been taken as evidence that not only is visual long-term memory (VLTM) capable of supporting thousands of images, but that memory for these items is "highly detailed." Additionally, it has been suggested that when observers make gist-based false recognition errors (e.g., you mistake your friend's cell phone for yours), it may be related to insufficient recruitment of stored details during retrieval (Guerin et al. 2012). Altogether, results across many studies have supported the notion that long-term memory representations are highly detailed.

However, most of these studies utilized alternative forced-choice designs, e.g., two-alternative forced choice (2AFC), where detailed memory "may not be necessary" for successful retrieval. For example, it is not clear whether the respose to a 2AFC probe trial is based on familiarity and/or recollection. A wealth of research has suggested that these two components of recognition memory, familiarity (knowledge that an item has been previously seen) and recollection (detailed recall of the item along with its context i.e., where and when it was observed) (Yonelinas, 2001, 2002) may be supported by different neural mechanisms (Fortin et al. 2004; Ranganath et al. 2004; Diana et al. 2007; Vilberg and Rugg 2007). We suggest that since both of these processes likely contribute to retrieval judgments, the ability to harness one more than the other may affect performance on retrieval tasks. In a 2AFC design, it is likely that at least some subset of the decisions are guided by familiarity information alone, whereas old/new recognition designs make it very difficult to utilize representations based solely on familiarity. Thus, a more complete understanding of the nature of the mnemonic representations underlying retrieval judgments requires probing with both types of testing procedures.

To further illustrate the critical role that differential underlying mnemonic representations may play in retrieval judgments, consider that in a 2AFC trial, two images are compared side by side, which allows observers to solve the trial by relying on relative familiarity (or novelty) between the two choices. This logic is similar to the interpretation of multiple choice exam questions. For example, a student encounters a question such as: What is the capital of Kazakhstan? (A) Astana (B) Paris (C) London, and (D) Buenos Aires. The student then examines the answers and realizes that the correct response is not B, C, or D, so it must be A. To the instructor, the student's response may lead to the assumption that the student possessed adequate detailed knowledge. However, if this knowledge were probed using a short-answer question, it is possible that the student would not arrive at the correct answer. The same logic holds true for 2AFC versus a single item old/new recognition (ONR) probe. An ONR test allows us to limit the possibility that the observer arrives at the correct answer using a noisy or low-fidelity representation (although it is still a possibility on some number of "guesses" that observers could use a noisy representation to influence their guess). We suggest that an ONR probe limits the use of familiarity-based memory mechanisms, allowing it to reveal critical performance differences as a function of lure similarity. Thus, it significantly limits the possibility that observers will rely on a relative comparison process to solve the task. We hypothesized that probing performance on similar lures using an ONR test would result in degraded performance compared with 2AFC as it would be less affected by familiarity processes and would be more reliant on processes attributed to recollection. We argue that this degradation in performance is not a decrease along a single continuum, rather performance changes as a function of the utilization of different sources of memory information (Daselaar et al. 2006). Specifically, in the current study, we used two different types of probe tests, a 2AFC and an ONR test in a within-subjects design. We replicated the Brady et al. findings using the 2AFC test but additionally showed that ONR performance in the same subjects was degraded, which suggests that the underlying mnemonic representations are not sufficiently detailed in some circumstances. Critically, we reliably show that, much like the multiple choice versus short-answer anecdote, the two different probe types lead to different conclusions about the fidelity of

VLTM. The present study complements the previous "massive memory" results by demonstrating that when observers are able to utilize gist representations, mnemonic representations are sufficiently detailed for successful test performance. However, when observers can only access one source of information (i.e., ONR), memory representations are not sufficiently detailed for test performance. Overall, our results provide additional insight into the mnemonic representations that underlie VLTM capacity limitations.

A group of 28 Johns Hopkins University undergraduate students and community members (mean age = 22.5, SD = 4.6, 16 female) participated in the experiment. The protocol was approved by the Johns Hopkins Homewood Institutional Review Board. Stimuli used in the study were photographic images of objects collected from Brady et al. (2008), Hemera Photo-Objects: Volumes I and II, and Google Image Search. We had participants complete a study phase containing 1054 images (910 new and 144 repeated images). During the study phase, participants were presented with each image for 3 sec followed by an 800-msec fixation cross (experimental setup similar to Brady et al. 2008). Participants performed a repetition-detection task (n-back) during the study phase where repeated images in the study stream varied in the number of n-backs: ranging from 1-back to 128-back, with more trials distributed to the shorter n-backs and fewer in the longer n-backs. Participants were told to respond (by pressing the spacebar) whenever they thought they saw a repeated image. They were given feedback only when they responded (i.e., "Hit" or "False Alarm"). Each participant saw a randomized order of the study objects and the specific items used for the repeated items were different for each of the participants.

Immediately after participants completed the study phase, they were given two memory tests (2AFC and ONR) to complete (see Fig. 1). Order of the tests was counterbalanced across all subjects. Objects from the study phase (old) and different new objects were used for each of the memory tests. Old items used in the test phase were drawn in an equally distributed manner from across the duration of the study task.

In the 2AFC test, we investigated memory for the previously seen objects by presenting two objects on the screen (side-by-side), one previously seen and the other new, and asked participants to make a judgment about which object they had previously seen. Positions of the old and new objects were counterbalanced. Responses were recorded via the left and right arrow keys. There were three types of trials: novel, exemplar, and state. Novel trials contained one previously seen object (old) and a different randomly chosen object that was never seen before (new).

Exemplar trials contained a previously seen old object and a new object that shared the same category (e.g., both objects might be bicycles). Finally, the State trials contained a previously seen old object and a new item that was the exact same object but appeared in a different position or state (e.g., the old item might be a blue mailbox closed and the new item (lure item) might be the same blue mailbox but with the door open). There was a total of 300 test trials (100 trials each for the trial types, all randomly intermixed).

In the ONR test, we presented participants with a single image in the center of the screen and instructed them to rate from 1–6 (using the number keys) how strongly they felt that the presented image was either a previously seen object (old) or a brand new object (new). Selecting 1 on the scale would indicate a strong rating that the item is old and that the observer had previously seen it, while selecting a 6 would correspond to a strong rating that the item was new. For the purposes of our analyses, we considered responses from 1–3 as "old" and responses 4–6 as "new." The numerical scale was reversed for half of the participants. The ONR memory test consisted of 300 trials; half of the items were old items from the study phase and half of the items were new items. Similar to the 2AFC memory test, new items consisted of brand new objects that had never been seen before, objects that shared similar exemplar category with a previously seen object in the study phase, and objects that were the same as one seen in the study phase, only it was in a different position or state. All trial types were randomly intermixed.

We found that performance was high for repetition detection during the study phase (89% correct) with few false alarms (<2%). Performance in the 2AFC memory test was high and showed a similar pattern to the results from Brady et al. (2008); performance was best in the novel condition (88%), followed by the exemplar condition (75%), and the state condition (73%) (Fig. 2, right side). A one-way repeated-measures analysis of variance (ANOVA) revealed that there was a significant effect of condition, $F_{(2,54)} = 99.06$, $P < 0.001$. Additional contrasts revealed that performance in the novel condition was significantly better than performance in the exemplar and state conditions (both $P < 0.001$, Bonferroni-corrected for multiple comparisons). However, the exemplar and state conditions were not significantly different from one another ($P > 0.1$).

In contrast, hit rate on the ONR memory test (see Fig. 2, left side, "Repetition" bar) was much lower than the proportion correct in any of the 2AFC memory test conditions (see Fig. 2, right side, "Novel," "Exemplar," and "State" bars. As a first pass, raw performance for both tests (hit rate for the ONR [hit rate = 0.604]
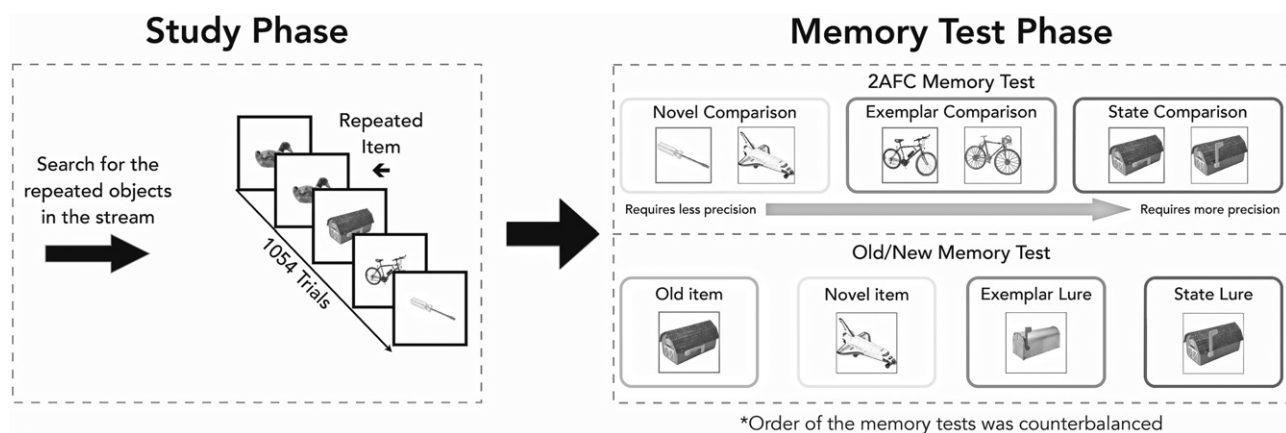
Figure 1. Examples of the study phase and memory test phases.
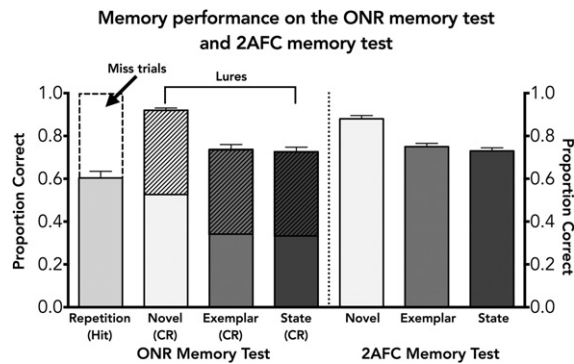
Figure 2. Memory performance on the ONR memory test (*left*) and 2AFC memory test (*right*) that were given after the study phase. The dashed box *above* the "Repetition" bar indicates the proportion of miss trials. The raw correct rejection score is equal to the total bar height for each of the lure conditions. The lure discrimination index (Yassa et al. 2011) accounts for response bias by subtracting the number of "new" responses attributed to misses from each of the lure conditions. This proportion is shown in the hashed portion on the *top* of each lure condition (novel, exemplar, and state).

compared with the mean proportion correct for all three conditions of the 2AFC for each subject [mean hit rate = 0.788]) were significantly different, $t_{(27)} = 13.2$, $P < 0.0001$. However, to appropriately compare these two memory tests, we corrected for response bias across the conditions in the ONR memory test. This involved correcting the raw correct rejection (CR) performance (i.e., the total bar height for each of the three lure conditions in Fig. 2) for the novel, exemplar, and state lure conditions to account for the number of "miss" responses, which indicate a bias to respond to items as "new." Specifically, lure performance was corrected for any response bias using a lure discrimination index (LDI) operationalized as p("new"|similar item)—p("new"|old item), which controls for response bias (Yassa et al. 2011). This correction has been indicated in the hashed region in Figure 2 on the lure bars. LDI performance was higher in the novel lure condition (53%) compared with the exemplar lure (35%) and state lure (34%) conditions. A one-way repeated-measures ANOVA revealed that there was a significant effect of condition, $F_{(2,54)} = 54.86$, $P < 0.001$. Additional contrasts revealed that performance in the novel condition was significantly better than performance in the exemplar condition ($P < 0.001$) and state condition ($P < 0.001$); contrasts were corrected for multiple comparisons. However, the exemplar and state conditions were not significantly different from one another ($P > 0.1$) similar to the 2AFC results.
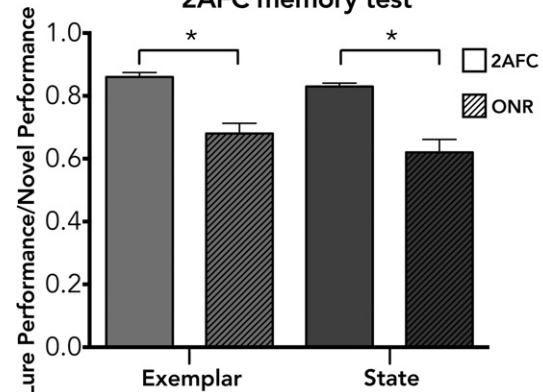
We next conducted a repeated-measures ANOVA to compare performance across both recognition memory tests. Results revealed that there was a main effect of test type (2AFC versus ONR), $F_{(1,27)} = 538.8$, $P < 0.001$, Geisser–Greenhouse corrected for nonsphericity, indicating that performance in the 2AFC recognition memory test was significantly better than performance in the ONR memory test. Additionally, there was a main effect of lure condition, $F_{(2,54)} = 117.4$, $P < 0.001$, Geisser–Greenhouse corrected for nonsphericity. Finally, the interaction was significant, $F_{(2,54)} = 3.7$, $P < 0.05$, Geisser–Greenhouse corrected for nonsphericity. Additional contrasts revealed that performance in the novel condition differed significantly compared with the state and exemplar conditions for the 2AFC and ONR memory tests ($P < 0.01$). Finally, using a d′ table from Hacker and Ratcliff (1979) we found that the d′ for the novel condition in the 2AFC task was 1.66. Additionally, we calculated the d′ for the novel condition in the ONR memory test (mean hit rate = 0.604, mean false

alarm rate for the Novel condition = 0.084), which was 1.64, suggesting that performance for detecting an old item and knowing a random new item is "new" was comparable in both tasks.

We examined the differences in the memory representations that are probed with 2AFC memory tests and ONR memory tests by comparing the change in performance as a function of the level of detail that was probed. Specifically, we measured the change in performance from the novel (baseline) condition compared with performance in the exemplar condition and the state condition for each memory test (2AFC and ONR), allowing us to investigate the change in memory performance when more object details are required to answer correctly. By using a ratio, we can account for the overall differences in the baseline performance and only examine how performance changes within each memory test type as a function of the level of detail that was probed. For the 2AFC test, we found that performance decreased 14% for the exemplar condition compared with the novel condition and 17% for the state condition compared with the novel condition. In the ONR test we found that performance, using the LDI measures for each lure condition, decreased 32% for the exemplar condition compared with the novel condition and 38% for the state condition compared with the novel condition. A repeated-measures ANOVA revealed that there was no significant difference in the lure condition (exemplar versus state), $P > 0.1$ across the memory tests (2AFC versus ONR). However, there was a significant difference in the memory test type (2AFC versus ONR), $F_{(1,27)} = 47.75$, $P < 0.001$. When performance was tested using ONR, memory for object details is ∼20% lower than what was previously found in Brady et al. (2008) (see Fig. 3).

We found that memory performance is high when probed using a 2AFC memory test as previous results suggested, however, when different objects from the same study set were probed within the same subjects using an ONR memory test, performance was significantly lower. The higher performance in the 2AFC probe is attributed to the fact that observers could arrive at the correct answer using two very different strategies: (1) the observer knows which image is the old item (this is the strategy that is typically assumed to underlie performance), or (2) the observer may have a

Figure 3. Detailed object memory for the 2AFC memory test compared with the ONR memory test. The *y*-axis is the ratio of performance in a lure condition divided by performance in the novel (baseline) condition. This ratio is a measure of the change in performance as a function of the degree of detailed memory that is probed. The *x*-axis represents the different lure conditions. Performance in the ONR memory test is significantly lower than performance in the 2AFC memory test across both exemplar and state conditions.

noisy representation of the old item, but knows he/she did not see the new item, therefore the other item must be the old item he/she previously saw. Thus, it is possible that observers utilize low-fidelity gist memory representation to infer the correct answer, at least on a subset of trials.

Guerin et al. (2012) showed that in situations where observers make gist-based false recognition errors on a forced-choice memory test, they will overwhelmingly choose the target item if they are provided the right test conditions. Guerin et al. (2012) argue that observers have highly detailed VLTM representations available in memory because they are able to choose the appropriate target under the right circumstances. Thus, they suggest that differences in performance are attributed to whether observers have access to that information during retrieval. Our results may also be interpreted within the accessibility framework. Different types of information are available for observers during retrieval in 2AFC versus ONR probes. Critically, the 2AFC probe provides two different types of information that can aid retrieval: information about the item previously seen (familiarity), information about new items not previously seen (novelty). The combination of those sources of information may influence observers to perform better on the 2AFC probe compared with the ONR probe.

Finally, our results are also consistent with findings from Holdstock et al. (2002), which showed that a patient with hippocampal damage differed in her performance in forced-choice memory tests compared with ONR memory tests. Specifically, patient Y.R. was tested on her ability to recognize previously seen objects using a 4AFC memory test and an ONR memory test. The results revealed that patient Y.R. was impaired for the ONR task, however her forced-choice task performance was not significantly different from normal controls (even when tested using similar lures, similar to the state comparison from Brady et al. (2008)). Additionally, it has been shown that the contributions of familiarity and recollection memory sources can differ depending on the type of items that are presented in forced-choice memory tests (Migo et al. 2014). These results suggest that regions outside the hippocampus are capable of supporting high performance on forced-choice tasks.

We conclude that while VLTM capacity may be large, it is reliant on gist as well as detailed representations. These contributions can be identified and isolated using different probe strategies, allowing us to make advances in understanding the nature of the mnemonic representations stored in VLTM.

Acknowledgments

References

Brady TF, Konkle T, Alvarez GA, Oliva A. 2008. Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci* **105:** 14325–14329.

Daselaar SM, Fleck MS, Cabeza R. 2006. Triple dissociation in the medial temporal lobes: recollection, familiarity, and novelty. *J Neurophysiol* **96:** 1902–1911.

Diana RA, Yonelinas AP, Ranganath C. 2007. Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends Cogn Sci* **11:** 379–386.

Fortin NJ, Wright SP, Eichenbaum H. 2004. Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature* **431:** 188–191.

Guerin SA, Robbins CA, Gilmore AW, Schacter DL. 2012. Retrieval failure contributes to gist-based false recognition. *J Mem Lang* **66:** 68–78.

Hacker MJ, Ratcliff R. 1979. A revised table of d′ for M-alternative forced choice. *Percept Psychophys* **26:** 168–170.

Holdstock JS, Mayes AR, Roberts N, Cezayirli E, Isaac CL, O'Reilly RC, Norman KA. 2002. Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus* **12:** 341–351.

Konkle T, Brady TF, Alvarez GA, Oliva A. 2010a. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J Exp Psychol Gen* **139:** 558–578.

Konkle T, Brady TF, Alvarez GA, Oliva A. 2010b. Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol Sci* **21:** 1551–1556.

Migo EM, Quamme JR, Holmes S, Bendell A, Norman KA, Mayes AR, Montaldi D. 2014. Individual differences in forced-choice recognition memory: partitioning contributions of recollection and familiarity. *Q J Exp Psychol (Hove)* **67:** 2189–2206.

Ranganath C, Yonelinas AP, Cohen MX, Dy CJ, Tom SM, D'Esposito M. 2004. Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia* **42:** 2–13.

Shepard RN. 1967. Recognition memory for words, sentences, and pictures. *J Verbal Learn Verbal Behav* **6:** 156–163.

Standing L. 1973. Learning 10,000 pictures. *Q J Exp Psychol* **25:** 207–222.

Standing L, Conezio J, Haber RN. 1970. Perception and memory for pictures: single-trial learning of 2500 visual stimuli. *Psychon Sci* **19:** 73–74.

Vilberg KL, Rugg MD. 2007. Dissociation of the neural correlates of recognition memory according to familiarity, recollection, and amount of recollected information. *Neuropsychologia* **45:** 2216–2225.

Vogt S, Magnussen S. 2007. Long-term memory for 400 pictures on a common theme. *Exp Psychol* **54:** 298–303.

Yassa MA, Lacy JW, Stark SM, Albert MS, Gallagher M, Stark CE. 2011. Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus* **21:** 968–979.

Yonelinas AP. 2001. Components of episodic memory: the contribution of recollection and familiarity. *Philos Trans R Soc Lond B Biol Sci* **356:** 1363–1374.

Yonelinas AP. 2002. The nature of recollection and familiarity: a review of 30 years of research. *J Mem Lang* **46:** 441–517.