Item frequency in probe-recognition memory search: Converging evidence for a role of item-response learning

Rui Cao¹ · Richard M. Shiffrin¹ · Robert M. Nosofsky¹

Published online: 20 December 2017
© Psychonomic Society, Inc. 2017

Abstract

In short-term probe-recognition tasks, observers make speeded old–new recognition judgments for items that are members of short lists. However, long-term memory (LTM) for items from previous lists influences current-list performance. The current experiment pursued the nature of these long-term influences—in particular, whether they emerged from item-familiarity or item-response-learning mechanisms. Subjects engaged in varied-mapping (VM) and consistent-mapping (CM) short-term probe-recognition tasks (e.g., Schneider & Shiffrin, *Psychological Review*, 84, 1–66, 1977). The key manipulation was to vary the frequency with which individual items were presented across trials. We observed a striking dissociation: Whereas increased presentation frequency led to benefits in performance for both old and new test probes in CM search, it resulted in interference effects for both old and new test probes in VM search. Formal modeling suggested that a form of item-response learning took place in both conditions: Each presentation of a test probe led to the storage of that test probe—along with its associated “old” or “new” response—as an exemplar in LTM. These item-response pairs were retrieved along with current-list items in driving observers’ old–new recognition judgments. We conclude that item-response learning is a core component of the LTM mechanisms that influence CM and VM memory search.

Keywords Short-term probe recognition · Memory search · Long term memory · Computational modeling · Response times

Probe-recognition memory-search tasks are among the most common paradigms for studying memory. In these tasks, a list of to-be-remembered items (the “memory set”) is presented, followed by a test probe that either is a member of the memory set (an “old” probe or “target”) or not a member of the memory set (a “new” probe or “foil”). Subjects aim to make the old–new judgment as quickly and accurately as possible. Both accuracy and response time (RT) are recorded to measure subjects’ performance. As observed in the original Sternberg (1966) studies, RT increases substantially as the size of the memory set increases, a result termed the set-size effect. The detailed processes that operate when participants engage in the memory-search task may vary with details of the experimental conditions (e.g., McElree & Doshier, 1989; Monsell, 1978;

Nosofsky, Little, Donkin, & Fific, 2011; Sternberg, 1966; for a comprehensive review and analysis, see Sternberg, 2016). For present purposes, however, the key point is that the presence of the memory set-size effect provides a clear indication that the observers’ engagement with the current set in short-term memory plays a fundamental role in determining performance.

In their studies that examined hybrid forms of memory/visual search, Schneider and Shiffrin (1977) discovered that the set-size effect could be greatly reduced or eliminated under consistent mapping (CM) conditions. In CM, the old (target) probes are chosen from one fixed set of items on every trial (termed the positive set), and the new (foil) probes are chosen from a separate fixed set of items on every trial (termed the negative set). Thus, the old targets and the new foils never switch roles across trials. As practice proceeded in Schneider and Shiffrin’s studies, performance improved dramatically: subjects were able to make their old–new judgments with shorter RT and fewer errors. Most importantly, the performance became largely invariant to the set-size manipulation, suggesting reliance on a process other than the retrieval of the list held in short-term memory (see also Logan & Stadler, 1991). In Schneider and Shiffrin (1977),

✉ Robert M. Nosofsky
nosofsky@indiana.edu

Rui Cao
caorui.beilia@gmail.com

¹ Department of Psychological and Brain Sciences, Indiana University,
1101 E. Tenth Street, Bloomington, IN 47405, USA

subjects were also tested in a varied-mapping (VM) condition, where the items that served as old probes on some trials were new probes on other trials, and vice versa. In contrast to CM, performance in the VM condition improved very little with practice, and the set-size effect persisted even after extensive practice. The researchers proposed that performance in the VM condition required an effortful, controlled process, regardless of the amount of practice, whereas practice in the CM condition allowed for the development of an extremely efficient, automatic form of information processing.

Although Shiffrin and Schneider (1977) developed a conceptual framework for understanding the nature of the controlled and automatic processes that developed in these tasks, their modeling did not delve deeply into the quantitative details of the processes at work. One influential model that aims to provide a quantitative account of the development of automaticity is Logan's (1988) instance theory. According to instance theory, highly efficient automatic performance arises by retrieving responses that are linked to the instances stored in long-term memory. With increased practice, more instances are stored in memory, which leads to a more efficient retrieval process (see Logan, 1988, 1990, for details). However, instance theory focuses only on how behavior changes in CM training and therefore does not provide a detailed account for the difference between CM and VM performance. Strayer and Kramer (1994) developed some descriptive accounts based on diffusion modeling (Ratcliff, 1978) to characterize the differences across CM and VM data patterns. They concluded that the difference reflects changes in both drift rates (i.e., rates of evidence accumulation) and response thresholds. Although the researchers further interpreted the diffusion-model parameters from the perspective of strategic versus learning factors, the aim of the paper was not to develop a mechanistic account of the cognitive processes that give rise to the different evidence accumulation rates in the CM and VM conditions. The main goal of the current work is to fill that gap and move toward the development of a process-level model that provides a quantitative account of performance in both CM and VM memory-search tasks. Because it is often difficult to derive precise predictions from theories that are specified at a purely verbal level, and because results from VM and CM memory-search tasks have been extremely influential in guiding thinking about the development of automaticity, this goal of developing a formal mathematical process model for VM and CM memory search is clearly a highly significant one.

Some progress towards this goal was made in recent work by Nosofsky, Cao, Cox, and Shiffrin (2014; Nosofsky, Cox, Cao, & Shiffrin, 2014). The formal model is an extended version of the exemplar-based random-walk (EBRW) model that has been successfully applied to various forms of categorization (Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005) and old–new recognition memory (Nosofsky et al., 2011). In the version of the model applied to probe-

recognition memory search, each item of the memory set is stored as an exemplar in short-term memory. Items from previous memory sets may also be stored in long-term memory. When the test probe is presented, it activates exemplars to which it is similar (both short-term and long-term), and the activated exemplars race to be retrieved (see Formal Models section for details). In an item-familiarity-only version of the model, each retrieved exemplar leads an information accumulator to move toward an old threshold, while failure to retrieve an old exemplar leads the information accumulator to move toward a new threshold. The retrieval process continues until one or the other response threshold is reached, at which time the observer emits the response that is associated with that threshold.

In initial tests, Nosofsky, Cox, et al. (2014) found that, with appropriate parameter settings, this item-familiarity version of the model provided excellent accounts of both VM and CM accuracy and RT patterns across conditions with a wide range of list lengths (memory set sizes of 1, 2, 4, 8, and 16). However, in subsequent work, Nosofsky, Cao, et al. (2014) obtained evidence that clearly challenged the item-familiarity-only account of CM performance. In particular, in this study, the researchers examined cases in which the same stimulus served as a test probe across two consecutive trials. A key result was that in VM, there was massive interference in responding “new” to “new” test probes if that probe had been presented on the previous trial (for similar previous findings, see, e.g., Monsell, 1978). Crucially, however, there was no such interference in CM: if anything, repeating a new test probe across two consecutive trials led to slight facilitation.

Nosofsky, Cao, et al. (2014) suggested that in VM, the observer relied on an item-familiarity process: Recent past presentations of items boost their familiarity on the current trial, leading to greater tendencies to respond “old” to such items. Thus, one would observe interference across trials in which a new test probe was repeated. By contrast, the researchers interpreted the facilitation observed in the repeated trials of the CM condition as evidence that observers instead relied on remembered item-response mappings in that condition. In particular, the idea is that the old–new response associated with each test probe is stored along with that test probe on each trial of the experiment. In later trials, retrieval of exemplars with “old” response labels would drive the random walk toward the old response threshold, but retrieval of exemplars with “new” response labels would drive the random walk toward the new response threshold. Thus, in the CM condition, when a new test probe is repeated across trials, the item-response learning that took place would facilitate the “new” response on the current trial. Nosofsky, Cao, et al. (2014) developed a formal model to implement these ideas and it yielded good quantitative accounts of the full range of individual-subject performance across more than 30 sessions of CM and VM practice.

Although the model performed well, the crucial empirical result that motivated the model was derived from cases in which test probes repeated across consecutive trials. A concern that arises is that the observer might have access to the item-response label only for exemplars that have been presented very recently. In other words, the facilitation observed in CM for the repeated new stimuli may be a byproduct of more vivid memory traces from very recent presentations rather than arising from more durable long-term associations. Such a view is consistent with the finding that the lag with which items are presented on current lists often exerts a powerful effect on short-term probe recognition (e.g., McElree & Doshier, 1989; Nosofsky et al., 2011). In addition, participants may apply a special-purpose strategy to take advantage of the repeated-trials manipulation, but the strategy may have little generality across more usual conditions of CM training.

The present study addressed these concerns by varying the long-term frequency with which individual items were presented in both the CM and VM conditions. Clearly, the more frequently presented items would give rise to higher long-term familiarity. Thus, to the extent that item-familiarity mechanisms play the sole role, one would expect the high-frequency items to lead to greater tendencies to respond “old” (for both old and new test probes) in both the VM and CM conditions. Thus, one should observe *interference* effects for high-frequency new test probes in both VM and CM. By contrast, suppose instead that observers form long-term *item-response* associations in CM. Increasing the frequency of the consistent pairings should boost the strength of those item-response associations. Thus, even for the new test probes in the CM condition, one should see *facilitation* in performance for the high-frequency items (compared to the low-frequency ones), in direct contrast to the predictions from the item-familiarity model.

Finally, although most past accounts of the influence of LTM on VM recognition-memory performance involve familiarity-only mechanisms, we were also interested in exploring the extent to which item-response-learning mechanisms might play some role in VM as well. As will be seen, models that formalize the role of these item-familiarity versus item-response-learning factors also make very different predictions concerning the patterns of results that will be observed when presentation frequency is manipulated in the VM condition. We defer the precise statement of these predictions until after presentation of the formal model that guides the research.

Experiment

We tested subjects in both VM and CM probe-recognition memory-search tasks. In both tasks, we manipulated memory-set size. The key manipulation was to also vary the

frequency with which individual items were presented across trials in both the VM and CM tasks.

Method

Subjects

The subjects were 109 undergraduate students from Indiana University, who participated in partial fulfillment of an introductory psychology course requirement. Subjects were randomly assigned to either the CM condition (55 subjects) or the VM condition (54 subjects).

Stimuli and apparatus

The stimuli were drawn from a pool of 2,400 unique object images used and described by Brady, Konkle, Alvarez, and Oliva (2008). Each image subtended a visual angle of approximately 7 degrees and was displayed in the center of a gray background. The experiment was conducted on PCs using MATLAB and the Psychophysics Toolbox (Brainard, 1997). All subjects were tested individually in sound-attenuating cubicles.

Procedure

In all conditions, half the test probes were targets and half were foils, with type of test probe chosen randomly on each trial. The memory set sizes were 2, 4, and 6; memory set size was chosen randomly on each trial. For each subject, 32 stimuli were randomly sampled from the 2,400 images. On each trial in the VM condition, the memory set was randomly sampled from the 32-stimulus set, subject to the constraints of an item-frequency manipulation described below. Targets were randomly chosen from the memory set; foils were randomly chosen from the remaining items in the 32-item set. In the CM condition, for each subject, 16 stimuli were randomly drawn from the 32-stimulus set and served as the “positive set” on all trials; the remaining 16 stimuli served as the “negative set” on all trials. On each trial, the memory set was randomly sampled from the positive set, subject to the constraints of the item-frequency manipulation (see below). Target test probes were randomly chosen from the memory set; foil test probes were randomly selected from the negative set.

Item frequency was manipulated as follows. In both the VM and CM conditions, items in each subject’s stimulus set were randomly divided into high-frequency (HF), medium-frequency (MF), and low-frequency (LF) roles. HF items were assigned a “selection weight” of 10; MF items a selection weight of 5; and LF items a selection weight of 1. For each subject, the CM positive set contained two HF items, two MF items, and 12 LF items. The following sequential-selection

algorithm was used for constructing the memory set on each trial of the CM condition. Let w_i denote the selection weight associated with item i . Then, the probability that item i was the first item selected was given by $w_i / \sum w_k$. Next, the probability that item j was the second item selected (from among the remaining positive-set items) was given by $w_j / \sum_{k \neq i} w_k$, where $k \neq i$ denotes that the sum is across all items not including i . The item selections continued in analogous fashion until the memory set size was reached. (Note that although the memory-set items were selected using the just-described sequential algorithm, the serial positions of the selected items in the presentation sequence were chosen at random.) For target trials, the test probe was randomly drawn from the memory set; each memory-set item had an equal probability of serving as the test probe, regardless of the assigned weights. The CM negative set had the same structure as the CM positive set (i.e., two HF items, four MF items, and 12 LF items). Test items that were foils were selected from the CM negative set with probability equal to their relative selection weights (i.e., $w_i / \sum w_k$).

The item-frequency manipulation in the VM condition was analogous to the one just described for the CM condition. For each subject, the VM set contained four HF items, four MF items, and 24 LF items, with selection weights as described above. The probability that item i was the first item selected for inclusion in the memory set was given by $w_i / \sum w_k$; the probability that item j was then the second item selected was given by $w_j / \sum_{k \neq i} w_k$; and so forth until the memory set size was reached. For target trials, the test probe was randomly drawn from the selected memory set; each memory-set item had an equal probability of serving as the test probe. For foil trials, the test probe was randomly selected from among those items not in the memory set, with probability proportional to its assigned selection weight.

The relative proportion of trials with which the different item types served as memory-set items, old test probes, and new test probes in the actual experiment is reported in Table 1.

Inspection of the table confirms that, in both the CM and VM conditions, and at both test and study, individual HF items occurred with the highest frequency, followed by individual MF items and finally individual LF items. In addition, the total probability with which each of the individual item types appeared as test probes was roughly equated across the CM and VM conditions. Of course, the relative frequency with which individual item types were assigned to specific responses differed across CM and VM; for example, in CM, an HF item from the positive set would always appear as an old test probe; in VM, an HF item would appear roughly half the time as an old test probe and half the time as a new test probe.

Each trial began with the presentation of a fixation point (asterisk) in the center of the screen for 0.1 second, followed by the presentation of the memory set. Each memory-set item

Table 1 Relative frequency for individual HF, MF, and LF items

	Prob. OLD test probe	Prob. NEW test probe	Prob. tested	Prob. studied
CM-positive HF	0.095	0	0.095	0.708
CM-positive MF	0.060	0	0.060	0.496
CM-positive LF	0.016	0	0.016	0.133
CM-negative HF	0	0.118	0.118	0
CM-negative MF	0	0.060	0.060	0
CM-negative LF	0	0.012	0.012	0
VM HF	0.053	0.044	0.097	0.43
VM MF	0.031	0.029	0.060	0.25
VM LF	0.007	0.008	0.017	0.05

CM = consistent mapping; VM = varied mapping; HF = high frequency; MF = medium frequency; LF = low frequency; Prob. = probability

was presented in the center of the screen for 1 s with a 0.1-s interstimulus interval. After a 1-s retention interval, a second fixation point (plus sign) was presented for 0.5 s, followed by the presentation of the test probe. The test probe remained on the screen until subjects responded (by pressing the ‘F’ or ‘J’ key on the computer keyboard). Feedback (“Correct!” or “Incorrect”) was then provided for 1 s. Each subject completed five blocks of testing with 25 trials per block. The computer reported to the subjects their overall percentage of correct responses at the end of each block. Each block took approximately 5 minutes to complete, with the entire session lasting approximately 30 minutes. Subjects were instructed to make their responses as quickly and accurately as possible. Subjects were not alerted to the possibility that some items would repeat frequently across trials; nor were they alerted to the differing structures of the VM versus CM conditions.

Results

We considered the first block to be a practice block, so did not include the data from the first block in our analyses. Although our inclusion of both MF and HF items was originally intended to yield stronger parametric constraints for model fitting, inspection of the data indicated similar results for the HF and MF items. To reduce noise in the data, we combined the results from the HF and MF trials (and refer to both as “HF”). (Combining HF and MF can be theoretically justified by assuming that strength in memory increases in negatively accelerated fashion with item repetition, so the MF items are much closer in strength to the HF items than to the LF ones.) Also, because our investigation was intended to investigate the effects of long-term frequency on CM and VM performance, we considered trials in which the same test probe was repeated from an immediately previous trial to be a special case and removed the few such trials from analysis

(~0.16% trials). Trials with response time (RT) greater than 5,000 ms or less than 180 ms were also eliminated (~0.7% trials). We then calculated the mean and standard deviation of RT for each Condition (CM vs. VM) $\times$ Set Size $\times$ Probe Type (old vs. new) $\times$ Frequency (HF vs. LF) $\times$ Lag combination and discarded trials that were greater than 2.5 standard deviations away from the mean (~3% trials). Finally, we eliminated the data from three outlier subjects in the CM condition who performed significantly worse than the remaining subjects in the group (overall median RT greater than 1,500 ms or overall proportion correct less than 0.8); and eliminated the data of three outlier subjects in the more difficult VM condition (overall median RT greater than 2,000 ms or overall proportion correct less than 0.6).

The main results of the experiment are displayed in the left panels of Figs. 1 and 2. In Fig. 1, we plot the mean RTs for correct responses as a function of tasks (CM vs. VM), set size, type of test probe (old vs. new) and item frequency. The error probabilities are plotted as a function of these variables in Fig. 2. The error bars indicate between-subjects standard errors. Ignoring for a moment the effects of the item-frequency manipulation, the overall data pattern is highly consistent with that of recent studies using a similar paradigm and set of materials (e.g., Nosofsky, Cox, et al. 2014): Performance in the CM condition (top row of each figure) is better than in the VM condition (bottom row of each figure), with both lower error rates and shorter RTs. Moreover, there is little if any effect of set size for new items in the CM condition but a big effect of set size for new items in the VM condition. Both conditions show set-size effects for old items, although the effects tend to be smaller in the CM condition than in the VM condition. A more detailed breakdown of the old-item data is shown in Figs. 3 and 4, which plots performance on the old items as a joint function of set size and lag of presentation (where lag is measured backwards from the end of the study list). Although the plots are noisy due to small sample sizes, they basically replicate patterns we have observed in closely related experiments (Nosofsky, Cao, et al., 2014, Nosofsky, Cox, et al., 2014; Nosofsky et al., 2011): Overall performance on old items gets worse with increases in lag, with little if any additional effect of set size once one condition is on lag. The main basis for the set-size dependence seen in Figs. 1 and 2 is the fact that larger set sizes include items with longer lags.

The key new results of interest involve the effects of the item-frequency manipulation. As can be seen in Figs. 1 and 2, in the CM condition, overall performance tends to be better for the HF items than for the LF items, although the locus of the effect varies across test-probe types and performance measures. Specifically, for the new test probes, mean RTs are shorter for HF than LF (with little difference in error probability, which is near floor for both HF and LF). For the old test probes, error probability is lower for HF than for LF (with

little difference in the mean RTs). Because the locus of the effect differs for the old and new probes, we conducted separate statistical analyses for them. We analyzed the CM data using a 3 (set size: 2 vs. 4 vs. 6) $\times$ 2 (HF vs. LF) repeated-measures ANOVA. The effect of the frequency manipulation on mean correct RTs for the new test probes was significant, $F(1, 51) = 12.96$, $p < .001$. In addition, overall mean correct RTs for the old HF probes were significantly shorter than for the old LF probes, $F(1, 51) = 4.59$, $p = .037$ (although the effect appears to be restricted to only the largest set size).¹ The effect of the frequency manipulation on error probability for the old test probes was also significant: $F(1, 51) = 8.26$, $p = .006$. There is no evidence for an effect of the frequency manipulation on error probability for the new test probes, $F(1, 51) < 1$; however, error probability is near floor for both HF and LF, so the lack of an effect is not surprising.

In sum, combining the patterns of accuracy and correct mean-RT data, overall performance on both the old *and* new test probes is better for the HF items than for the LF items in the CM condition. These results are consistent with the hypothesis that item-response-learning governed performance in the CM condition: For both old and new items, performance is benefited by increases in the frequency of consistent item-to-response training. The results are inconsistent with the hypothesis of a pure item-familiarity hypothesis for the CM condition: HF new items are far more familiar than are LF new items; and according to the item-familiarity hypothesis, increased familiarity should lead to increases in “old” responding. For new items in the CM condition, however, the RT results point decidedly in the opposite direction.

Our results for CM are reminiscent of results from a hybrid memory/visual-search paradigm reported recently by Wolfe, Boettcher, Josephs, Cunningham, and Drew (2015). In these studies, subjects repeatedly searched visual displays for the presence of targets from a single memorized list. The key manipulation across experiments was to vary the familiarity of foils that appeared in the visual displays. The general finding was that foil familiarity exerted little if any impact on visual-search performance (either in terms of false-alarm rates or slowed RTs), leading Wolfe et al. to conclude that, under CM conditions, item-familiarity mechanisms do not cause observers to confuse foils with targets. (Wolfe et al. did not test VM versions of their task.) Our present results for CM in pure memory-search tasks converge with those observed by Wolfe et al. in their hybrid memory-visual search tasks. Indeed, our

¹ Although there are some exceptions, it is generally the case that the frequency effects are larger at the large memory set sizes than at the smaller ones. This pattern of results is not surprising because there will tend to be floor effects for small-size memory sets due to the ease of those conditions. Because our main concern is with overall item-frequency effects across the CM and VM conditions, in order to facilitate reading of our Results section, we report the outcome of the more detailed tests of set-size effects and tests of interactions between set size and item frequency in Table 4 in the [appendix](#).

Fig. 1 Mean correct response times plotted as a function of conditions (CM, VM), set size, test-probe type (new, old), and item frequency (HF, LF). Left panel = observed data; middle panel = predictions from full

version of item-familiarity model; right panel = predictions from core version of item-response-learning model. Error bars show the between-subject standard error of the mean

results suggest that increased foil frequency can *benefit* the process of rejecting CM foils.

In direct contrast to the CM task, in the VM task overall performance is *worse* for the HF items than for the LF items (see Figs. 1 and 2). Furthermore, this results holds for both performance measures (error probabilities and RTs) for both new *and* old probes. To analyze the data, we conducted a 2 (test-probe type: old vs. new) $\times$ 3 (set size: 2 vs. 4 vs. 6) $\times$ 2 (frequency: HF vs. LF) repeated-measures ANOVA. The analysis yielded a significant main effect of item frequency on both error probability, $F(1, 50)$

$= 18.66$, $p < .001$, and correct RT, $F(1, 50) = 8.75$, $p = .005$, reflecting the worse overall performance associated with the HF items. There was also a significant interaction between test-probe type and frequency, $F(1, 50) = 11.74$, $p = .001$, for the error data; $F(1, 50) = 11.54$, $p = .001$ for the RTs. The interaction reflects the finding that, whereas there was a big effect of item frequency for the new probes, there was only a trend for the old probes.

In the VM task, it is not surprising that HF new items are classified more slowly and with greater error probabilities than are LF new items. Such an effect is predicted by the item-

Fig. 2 Mean error probabilities plotted as a function of conditions (CM, VM), set size, test-probe type (new, old), and item frequency (HF, LF). Left panel = observed data; middle panel = predictions from full version

of item-familiarity model; right panel = predictions from core version of item-response-learning model. Error bars show the between-subject standard error of the mean

Fig. 3 Mean correct response times for the old test probes plotted as a function of conditions (CM, VM), set size, lag, and item frequency. Left panel = observed data; right panel = predictions from core version of item-response-learning model

familiarity hypothesis of VM performance. In particular, HF new items will tend to have far greater long-term familiarity than LF new items, which should interfere with observers' ability to classify such test items as "new." More interesting is that there was a trend for the HF *old* items to show a performance deficit compared to LF old items. This pattern of results is the opposite of what would be predicted by a simple item-familiarity hypothesis. Because HF old items have greater familiarity than LF old items, observers should show performance benefits in classifying them as "old," but the results point in the opposite direction.

As we demonstrate in our ensuing Theoretical Analysis section, the overall pattern of results is instead consistent with the idea that an item response-learning mechanism operates not only in the CM condition but may operate to some extent in the VM condition as well. The key factor is that high-frequency VM items have served as both old and new test probes in numerous previous test trials. There are several ways to implement a mechanism by which this factor could cause interference. The specific approach that we follow is to implement a mechanism that leads the high-frequency inconsistent mappings to result in lowered evidence-accumulation rates in

Fig. 4 Mean error probabilities for the old test probes plotted as a function of conditions (CM, VM), set size, lag, and item frequency. Left panel = observed data; right panel = predictions from core version of item-response-learning model

the EBRW memory-search model, resulting in lowered accuracy and longer RTs for the HF items.

Theoretical analysis

The formal models

A schematic illustration of the main components of the EBRW memory-search model² is presented in Fig. 5. We start by describing the components that are sensitive to the contents of the current study list (“short-term memory”). Then, we expand our description to include contributions from long-term memory as well.

Short-term memory components According to the model, each of the study items from the current list is stored as an individual exemplar in memory. The memory strength of each exemplar decreases with the lag with which it was presented on the study list. (Lag is measured backwards from the end of the study list.) More specifically, based on evidence reported by Donkin and Nosofsky (2012a; see also Anderson & Schooler, 1991; Wixted & Ebbesen, 1991), it is assumed that memory strength decreases as a power function of lag j :

$$m_j = \alpha + j^{-\beta}, \quad (1)$$

where α is asymptotic strength and β describes the rate of decrease in strength with lag. The differential memory strengths are represented schematically in Fig. 5a, where the larger circles represent exemplars with greater memory strength.

When the test probe is presented, the exemplars stored in memory are “activated” and “race” to be retrieved, with rates that are proportional to their activations (cf. Logan, 1988)—see Fig. 5b. The degree to which exemplar j (e_j) is activated is a joint function of exemplar j ’s memory strength and its similarity (s) to test-probe i (t_i):

$$a_{ij} = m_j, \quad \text{if } t_i = e_j \quad (2a)$$

$$a_{ij} = m_j s, \quad \text{if } t_i \neq e_j, \quad (2b)$$

where s ($0 < s < 1$) is a freely estimated similarity parameter. Thus, the study-list exemplars that are most highly activated are those that match the test probe and that have short lags.

As explained in detail in previous articles (e.g., Nosofsky, Cao, et al., 2014; Nosofsky et al., 2011), the EBRW-recognition model presumes that the observer establishes “criterion elements” in the memory system. Upon presentation of the test probe, the criterion elements (labeled “c” in Fig. 5b) race to be retrieved (along with the stored exemplars). The criterion elements race at a constant rate k , independent of the specific test probe that is presented.

Finally, the retrieved exemplars and criterion elements drive a random-walk process that leads to “old” versus “new” decisions (Fig. 5c). The observer sets response boundaries +*OLD* and −*NEW* that establish the amount of evidence needed for making an “old” or a “new” response. On each step of the random-walk process, if an old exemplar is retrieved, a random-walk counter takes a step toward the “old” response boundary; whereas if a criterion element is retrieved, the random-walk counter steps toward the “new” response boundary. The retrieval process continues until one of the response boundaries is reached, at which point the observer emits the appropriate response.

Given further technical assumptions concerning the distribution of exemplar race times (see Nosofsky & Palmeri, 1997, p. 268), it turns out that, on each step of the random walk, the probability that the random-walk counter steps toward the +*OLD* response boundary (p_i) is given by:

$$p_i = A_i / (A_i + k), \quad (3)$$

where A_i is the summed activation of the test probe to all the study-list items:

$$A_i = \sum a_{ij}, \quad (4)$$

and k is the level of criterion-element activation. (The probability that the random walk steps toward the new boundary is simply $q_i = 1 - p_i$.)

Through experience in the task, the observer is presumed to learn an appropriate setting of criterion-element activation k , such that the summed activation (A_i) tends to exceed k when the test probe is old, but tends to be less than k when the test probe is new. Because A_i tends to increase with set size (for “new” test probes), we presume that the observer may adjust the criterion-element activation with changes in set size. As an approximation to implementing possible criterion adjustment, it is assumed that the criterion setting varies linearly with memory set-size M :

$$k(M) = u + vM. \quad (5)$$

Long-term memory components Recent extensions of the EBRW memory-search model (Nosofsky, 2016; Nosofsky, Cao et al., 2014) implement the influence of previous study-test trials (beyond the current study list) with a set of long-term memory (LTM) components (Fig. 5b). Specifically, study and

² As noted in our introduction, different strategies may operate in the short-term probe-recognition task, depending on details of the experimental conditions (Sternberg, 2016). The strategy formalized in the EBRW model is hypothesized to operate under conditions involving fairly rapid presentations of memory-set items, short intervals between study and test, and no requirement that the participants report the ordering of the memory-set items subsequent to their old–new judgment on each trial.

Fig. 5 Application of the exemplar-based random-walk model to the short-term probe-recognition task. *Note.* O_k is the old item on the current study list that is presented in serial-position k

test items from the previous trials are presumed to be stored as exemplars in LTM and race to be retrieved along with the current study-list exemplars. We distinguish between two processes that may mediate the influence of the retrieved LTM exemplars. First, retrievals of LTM exemplars may always drive the random walk towards the +*OLD* response boundary, regardless of the exemplars' status as "old" versus "new" test probes on the previous trials. We denote such a process as an "item-familiarity" (FAM) model and formalize the model with a set of *FAM* parameters. Alternatively, the observer may store along with the previously tested items their associated "old" versus "new" response labels and retrieve item-response pairs. Retrieval of exemplars with "old" response labels would drive the random walk towards the *OLD* response boundary, whereas retrieval of exemplars with "new" response labels would drive the random walk towards the *NEW* response boundary. We denote such a process as an "item-response-learning" (IR) model and formalize it with a set of *IR* parameters. The details of both models are described below.

LTM-FAM In the FAM model, we presume that the activation and retrieval of LTM exemplars always drives the random-walk counter towards the +*OLD* boundary. For simplicity, we account for the boost in the summed activation (A_i) with a free parameter *FAM*:

$$p_i = (A_i + FAM) / [(A_i + FAM) + k]. \quad (6)$$

It is natural to assume that HF test probes receive a greater familiarity boost than do LF test probes. Therefore, we estimate separate *FAM* parameters for the HF items versus the LF items (with the constraint that $FAM_{HF} \geq FAM_{LF}$). As discussed in more detail in the model-fitting section, although the model supposes that the same basic process applies across the CM and VM conditions, we allow the *FAM* parameter values to vary across these conditions.

LTM-IR In the IR model, we presume that the retrieved "item-plus-response-label" exemplars direct the random walk counter to the response threshold that corresponds with the stored

response label. We denote by *IR-OLD* the boost toward the +*OLD* boundary and by *IR-NEW* the boost toward the −*NEW* boundary. Given the structure of the CM condition, old test probes will activate many exemplars with "old" response labels but no exemplars with "new" response labels. Thus, in the CM condition, on trials in which old test probes are presented, the probability that the random walk steps toward +*OLD* is given by:

$$p_i(\text{old}) = (A_i + IR-OLD) / [(A_i + IR-OLD) + k]. \quad (7a)$$

Analogously, because new test probes will retrieve only exemplars with "new" response labels, the probability that the random walk steps toward the −*NEW* boundary (if tested with a "new" probe) is given by:

$$q_i(\text{new}) = (k + IR-NEW) / [(k + IR-NEW) + A_i]. \quad (7b)$$

As is the case for the FAM model, we presume that HF items have strengths in LTM at least as great as LF items, so we introduce the parameter constraints that $IR-OLD_{HF} \geq IR-OLD_{LF}$ and that $IR-NEW_{HF} \geq IR-NEW_{LF}$. Thus, from inspection of Equations 7a and 7b, it can be seen that for CM, the IR model predicts increased evidence-accumulation rates to the correct response boundaries with increases in the frequency of the consistent response mappings.

Unlike in the CM condition, in the VM condition, a test probe will activate previous-trial exemplars with both "old" response labels and "new" response labels, regardless if it serves as an old or a new test probe on the current trial. (The reason is that in VM each item serves randomly as an old test probe and as a new test probe throughout the experiment.) Therefore, for both old and new test probes, the probability that the random walks steps toward the +*OLD* boundary is given by:

$$p_i = (A_i + IR-OLD) / [(A_i + IR-OLD) + (k + IR-NEW)]. \quad (8)$$

As was the case for CM, we again presume that $IR-OLD_{HF} \geq IR-OLD_{LF}$ and that $IR-NEW_{HF} \geq IR-NEW_{LF}$. Thus, from

inspection of Equation 8, it can be seen that for VM, the IR model predicts *decreased* rates of evidence accumulation to the correct response boundaries with increases in item frequency.

We should note that if the *IR-OLD* and *IR-NEW* parameters grow indefinitely with frequency and training, then they would come to dominate VM responding, and the current list would not even matter. In a subsequent discussion of the model-fitting results, we provide reasons why the IR parameters are *not* expected to grow indefinitely in VM; this subsequent discussion will also explain why the magnitude of the IR parameters is expected to be lower in VM than in CM.

The full version of the FAM model makes use of 11 free parameters for fitting the data of each condition (for a listing, see Table 3): the lag-related memory-strength parameters α and β (Equation 1); criterion-element parameters u and v ; similarity parameter s ; response-boundary parameters $+OLD$ and $-NEW$; a scaling parameter κ that measures the time of each step in the random walk; a residual-time parameter T_R that reflects non-decision-time processes; and the LTM parameters FAM_{HF} and FAM_{LF} . The IR model has 13 free parameters: the same ones as the FAM model, except instead of using the familiarity-based LTM parameters, it uses the set of item-response LTM parameters: $IR-OLD_{HF}$, $IR-OLD_{LF}$, $IR-NEW_{HF}$, and $IR-NEW_{LF}$.

Fits of the models to the group data

Because our main goal was to assess the extent to which the alternative models could account for broad, qualitative aspects of the data, we fitted both the FAM and IR models to the averaged group data by minimizing a weighted sum-of-squared deviations (WSSD) criterion. In particular, we required the models to simultaneously fit the mean-correct RT data and the error proportions data of (a) the new items as a function of set size and (b) the old items as a joint function of set size and lag. To jointly fit all these data sources, we need to apply different weights to the data points (because they are measured on different scales and based on differing sample sizes). We found that a good overall match to both the RT and accuracy data was achieved by minimizing the WSSD with the deviations from the accuracy data (measured in proportions) given twice the weight of the deviations from the RT data (measured in seconds); and the individual data points for new probes given 4 times the weight of the individual data points for the old probes. (Sample sizes for the new-item data points are much greater than for the old-item data points because they are not broken down by lag.)

Based on the theoretical considerations that we described earlier in the Formal Models section, we constrained the LTM parameters such that the boosts for the HF items were at least as great as for the LF items (for both the FAM and the IR models). Before imposing any other constraints, we started by fitting the “full” version of both models to the data, with

all parameters allowed to vary freely across the VM and CM conditions. The fits of the full models provide baselines for comparison with more constrained versions of the models that we examine subsequently. Because different processes may mediate performance across the CM and VM conditions, we reasoned that it was important to get started by fitting the models separately to the two conditions (i.e., with all free parameters allowed to vary).

The WSSD fits for different versions of the models are reported in Table 2. The best-fitting parameters from the full version of the FAM model are reported in Table 3 (along with the best-fitting parameters from a constrained version of the IR model that we describe below). Inspection of Table 2 reveals that the WSSD fit for the FAM model is worse than for the IR model for both the CM and VM conditions. Indeed, we will see that even highly constrained versions of the IR model fit the data from both conditions better than does the full version of the FAM model.

To see the reason for the poor fits yielded by the FAM model, we display its predictions of the mean-correct RTs and error probabilities in Figs. 1b and 2b, in the same fashion as for the observed data. As can be seen, the FAM model displays various qualitative shortcomings. First, it failed to predict any frequency effect for new test probes in the CM condition: the predictions for the HF items lie virtually on top of the predictions for the LF items for both the RT and accuracy data. By comparison, in the observed CM data, the RTs for the HF new items are much shorter than for the LF new items. Because the familiarity boost from HF items should be greater than for LF items, if anything the FAM model would predict that RTs for HF new items should be *longer* than for LF items, not shorter. (Its prediction of equality is achieved

Table 2 Weighted sum of squared deviation (WSSD) fits of different versions of the FAM and IR models to the mean correct RTs and error-probability data

Model	VM	CM	Total
FAM full model	0.181	0.113	0.294
IR full model	0.159	0.071	0.230
IR κ , T_r	0.160	0.072	0.224
IR κ , T_r , s	0.161	0.075	0.236
IR κ , T_r , s , u , v	0.164	0.076	0.240
IR κ , T_r , s , u , v , +OLD, -NEW	0.165	0.081	0.246
IR core	0.163	0.085	0.248

Note. FAM = item-familiarity model; IR = item-response learning model; VM = varied-mapping; CM = consistent-mapping. The parameter listings next to the IR model denote the parameters that were held fixed across the VM and CM conditions in fitting the special-case versions of the model to the data. (See Table 3 for a listing and description of the parameters.) The IR-core model was the most highly constrained of the special-case IR models and held fixed across VM and CM the parameters κ , T_r , s , u , v , +OLD, -NEW, and β

Table 3 The best-fitting parameter values for full FAM model and core IR model

FAM full			IR core		
Parameters		CM	VM	CM	VM
α		5.420	0.468	0.174	0.456
β		0.934	1.870	2.057	–
u		27.813	0.324	0.394	–
v		0.334	0.010	0.015	–
s		0.059	0.043	0.065	–
LTM	*FAM _{HF}	22.816	0.043	*IR-OLD _{HF}	0.557
				IR-OLD _{LF}	0.433
	FAM _{LF}	22.725	0.001	*IR-NEW _{HF}	0.088
				IR-NEW _{LF}	0.001
OLD		22.854	3.125	4.109	–
NEW		107.325	2.874	–	–
κ		0.349	84.907	36.037	–
Tr		349.695	378.527	507.697	–

Note. FAM = item-familiarity model; IR = item-response learning model; CM = consistent-mapping; VM = varied-mapping. Parameter values replaced with “–” were set equal to one another across the CM and VM conditions; all response thresholds were held fixed at a single value in the IR core model.

Parameters marked with “*” were constrained to be greater than or equal to the parameter immediately below. α = power-decay asymptote; β = power-decay rate; u = criterion intercept; v = criterion slope; s = similarity; OLD = old threshold; NEW = new threshold; κ = scale; Tr = residual time. See text for an explanation of the LTM parameters

only by setting the FAM_{HF} and FAM_{LF} parameters equal to one another; see Table 3.)

In addition, the FAM model struggles to account for the data from the VM condition. Although it correctly predicts the HF disadvantage for the new test probes, it failed to predict the trend of an HF disadvantage for the old test probes in the observed data (for both the RTs and the error probabilities). According to the model, when an HF item serves as a test probe, it will receive a higher familiarity boost from LTM (compared to LF items). If anything, this boost should *facilitate* responding “old”; thus, the model predicts somewhat shorter RTs and increased accuracy for HF old items than LF old items. By contrast, the observed data tend to show higher error rates and slightly longer RTs for the HF old items than for the LF old items. Because the FAM model failed to account for the data even with all its free parameters allowed to vary across conditions, we did not explore more constrained versions of the model.

In contrast to the FAM model, the full version of the IR model successfully captured most of the data patterns (and provided a better fit than the FAM model to both the CM and VM data; see Table 2). However, because a large number of free parameters were used, we decided to explore a series of more constrained versions of the IR model that might still achieve good accounts of the data. In each case, we held fixed across the CM and VM conditions additional parameters (rather than allowing the parameters to vary freely across the conditions). As can be seen in Table 2, with each additional constraint, there was a relatively small increase in the total WSSD. Here we focus on the most constrained version (the

“core” IR model). In this version, we held fixed across the CM and VM conditions: the scale κ and residual time T_r parameters; the similarity parameter s ; the lag-decay parameter β ; and the criterion parameters u and v . In addition, although one might expect the response-boundary parameters to vary in magnitude across both conditions and response types, we found that reasonable fits could be achieved with all response-boundary boundary parameters set at a single value. Thus, the key parameters that vary across conditions are the various LTM parameters (*IR-OLD_{HF}*, *IR-OLD_{LF}*, *IR-NEW_{HF}*, and *IR-NEW_{LF}*); see Table 3.

The predictions of the mean RTs and error probabilities from the core IR model are presented in Figs. 1c and 2c. (In addition, we show the predictions for the more fine-grained set-size by lag curves in Figs. 3b and 4b.) Although there are some minor exceptions, the model successfully captures most of the main qualitative effects in the data.³ In this discussion, we focus on the effects of the frequency manipulation, the

³ One minor exception is that the current version does not predict the decreases in RTs and error rates that are often observed at the greatest lag of each set-size condition. This decrease constitutes a primacy effect: The item with the greatest lag occupies the first serial position of the memory set. In past applications of the EBRW-recognition model (e.g., Nosofsky et al., 2011), a primacy parameter was added to capture the effect, but the effect seems tangential to the main issues under investigation in the present study. A second possible limitation is that the present version of the model predicts little, if any, additional effect of memory set size once one conditions on lag. In past applications, any small residual effects of set size on RT have been modeled in terms of increases in response-threshold settings (e.g., Donkin & Nosofsky, 2012b; Nosofsky et al., 2011; see also Ratcliff, 1978). Again, however, this more detailed issue is not central to the present investigation.

central theme of the present investigation. To begin, the model captures the overall HF advantage in the CM condition—for both old *and* new probes. According to the model, the test probe will receive a boost toward the correct response boundary when it activates the “item-plus-response-label” exemplars from past trials, because the response label is consistent with the correct response on the current trial. The magnitude of the boost should be positively correlated with the frequency of the consistent mapping (see also Schneider & Fisk, 1982, who manipulated *proportion* of CM trials associated with individual items in visual-search paradigms). This boost leads to the better performance for HF items than for LF items for both “old” and “new” test probes.

More surprising is that, compared to the FAM model, the IR model also provided a better account of the VM data. In particular, the IR model successfully predicted that performance on the old HF test probes would tend to be worse than on the old LF test probes. According to the model, although an old HF test probe will receive a strong boost toward +*OLD* from past trials, it will also receive a strong boost toward −*NEW* (from the many trials in which it served as a new test probe). As we explained previously, this strong LTM-based interference from past trials adds noise to the random-walk process (as formalized in the Equation-8 step-probability equation). There is less noise from LTM added for the LF items, so performance is governed more by the contents of the current study list.

The best-fitting parameters from the core IR model are reported in Table 3, and the pattern of parameter estimates seems reasonably interpretable. Naturally, for each response type (OLD vs. NEW), the LTM parameter values (IR-OLD and IR-NEW) associated with the HF items are greater in magnitude than those associated with the LF items: we imposed this relation as a constraint in our model fitting, but it emerged clearly in any case. In addition, for CM, the LTM parameters associated with old items were greater in magnitude than the LTM parameters associated with new items. A natural explanation is that, in addition to their consistent response mappings throughout the experiment, the old items had the advantage of often appearing on the study lists, whereas the new items would never appear on the study lists. The very low-magnitude estimate for IR-NEW_{LF} also seems reasonable, because individual LF items rarely appeared as new test probes, so there was little opportunity for item-response learning to occur for these items.

Another general result of the parameter estimates for the IR model is that the magnitude of the LTM parameter values for CM tended to be greater than the magnitude of the LTM parameters for VM. There are several reasonable explanations. Perhaps most important, in a fully specified model, provision would be made for the observer to differentially weight the STM and LTM sources of information in making old-new recognition judgments. An effective observer would give far

more weight to LTM in CM compared to VM. The reason is that the LTM item-response mappings are perfectly valid for the CM paradigm, whereas they provide zero information in the VM paradigm. Indeed, in VM, an effective observer would attempt to ignore the previous-lists history and focus solely on the contents of the current list, perhaps through the use of recency-based context cuing (e.g., Howard & Kahana, 2002). Such context cuing would tend to zero out the influence of memory traces laid down in the distant past. In sum, the differing magnitudes of the LTM parameters across CM and VM may be reflecting the greater weight that observers give to LTM in the CM condition than in the VM condition.⁴

A second issue involving the differing magnitudes of the LTM parameters across CM and VM is that the detailed mechanisms for the development of the item-response strengths remain to be delineated. For example, although one possibility is that observers simply accumulate individual item-response exemplars as they are experienced, an alternative is that in VM the “old” and “new” trials compete with one another to some extent, weakening the learned item-response strengths of both. Third, although the absolute frequency of individual item types was roughly balanced across the CM and VM conditions, it was of course the case that the frequency of assignment of items to specific responses differed: In CM, an HF item was either always assigned to an old response or always assigned to a new response, whereas in VM, an HF item was assigned roughly half the time to each response. Future research will be needed to disentangle these alternative explanations of the differing magnitude of the CM and VM long-term-memory parameters.

Finally, to gain additional information bearing on the nature of the LTM mechanisms in these tasks, we created summary plots of the data patterns shown in Figs. 1 and 2, separately for Blocks 2–3 and Blocks 4–5 of the experiment. Visual inspection indicated very similar patterns of results across these early and late blocks, suggesting that the detailed mechanisms that give rise to the LTM parameter values across CM and VM have their influence at fairly early stages of practice.

⁴ We should note that this differential-weighting hypothesis is reminiscent of the type of dual-process theory developed by Shiffrin and Schneider (1977) to account for the contrasting results across the CM and VM conditions of their hybrid visual/memory-search studies. For example, consider a simple form of dual-process model: CM, after enough training, would be accomplished by memorized and/or automatic stimulus–response pairings without retrieval or use of recently stored study or test traces. VM would be carried out with the use of context cuing that would activate recent traces, with responses based on some combination of item information and item-response information retrieved from those traces. (Of course, such a simple model for CM in our paradigm is unlikely, because low-frequency items do not receive much training, and subjects likely cannot help attending to the study lists, so that even from a dual-route perspective, CM performance is likely a mixture of both routes.) Our present studies were not designed to distinguish between the type of single-process model presented in this article and the type of dual-process theory described by Shiffrin and Schneider (1977), and future research would be needed to distinguish between such possibilities.

Discussion

Our results suggest that item-response learning is a core mechanism through which long-term memory (LTM) influences performance in tasks of both CM and VM short-term probe-recognition. According to this view, each presentation of a test probe leads to the storage of that test probe—along with its associated “old” or “new” response—as an exemplar in LTM. These item-response pairs are retrieved along with current-list items in driving observers’ short-term probe-recognition judgments.

The key manipulation used to diagnose the influence of this LTM mechanism in the present experiment was to vary long-term item frequencies across trials. In CM, increased item frequency led to better performance for both old and new items. Although the improved performance on old items is predicted by both item-response-learning and item-familiarity mechanisms, the improvement on the new items suggests a crucial role of item-response learning. In particular, item-familiarity mechanisms predict that increased frequency should have interfered with responding “new” to new items, but the results pointed decidedly in the opposite direction. Instead, our theoretical interpretation is that retrieval of the high-frequency new-item-response pairs from LTM boosted the evidence that such test probes were new. These results provide converging evidence for the significant role of item-response learning in CM memory search reported in the earlier study of Nosofsky, Cao, et al. (2014), in which the key manipulation was to repeat test probes across successive trials. However, whereas the evidence from that previous study may have reflected paradigm-specific strategies in which observers became sensitized toward the repeating-probe manipulation, the present results provide evidence of a much more general long-term item-response-learning influence on CM memory-search performance.

Whereas increased item frequency led to facilitation in CM memory search, it led to interference (for both new and old items) in VM memory search. Although the interference for new test probes is as predicted by familiarity-only models, the interference for old items contradicts the predictions from such models. In particular, because increasing frequency boosts familiarity, a familiarity-only model predicts that there should have been benefits in responding “old” for the high-frequency items, not interference. Instead, the overall pattern of results, even for VM, was again well explained by an item-response learning model. The basic idea is that in the VM condition, there were numerous previous trials in which the high-frequency items were assigned both old and new responses. Retrieval of these numerous conflicting item-response pairs adds significant noise to

the decision process, regardless of whether the correct response on the current trial is old or new. A formal implementation of this mechanism within an exemplar-based random-walk model yielded good qualitative accounts of the overall pattern of results across both the CM and VM conditions.

Thus, the present empirical and formal-modeling results suggest the possibility of considerable generality for the role of an item-response-learning component in wide varieties of memory-search performance. We suspect that a complete model of how memory-search performance evolves with practice will involve the contribution of multiple mechanisms, and we certainly do not conclude that long-term item-familiarity mechanisms do not also play a role. In addition, future work will also need to examine the extent to which the performance patterns for the high-frequency items reflect explicit strategies developed for those items or reflect processes that are more hard-wired into the memory system. Regardless of the outcome of these future investigations, the present evidence suggests that, at the least, item-response-learning mechanisms will form a core component of a fully satisfactory unified model of CM and VM memory search.

Author note This work was supported by AFOSR Grant FA9550-14-1-0307 to Robert Nosofsky.

Appendix

Table 4 Statistical test results examining main effects of set size, main effects of item frequency, and interactions between set size and item frequency for both RT and error probability for CM-old, CM-new, VM-old, and VM-new

	Set Size	Frequency	Interaction
CM old RT	$F(2, 102) = 13.76$ $p < .001$	$F(1, 51) = 4.59$ $p = .037$	$F(2, 102) = 0.22$ $p = .801$
CM old p (error)	$F(2, 102) = 0.29$ $p = .750^*$	$F(1, 51) = 8.26$ $p = .006$	$F(2, 102) = 1.03$ $p = .359^*$
CM new RT	$F(2, 102) = 1.30$ $p = .276$	$F(1, 51) = 12.96$ $p < .001$	$F(2, 102) = 2.69$ $p = .073$
CM new p (error)	$F(2, 102) = 1.02$ $p = .364$	$F(1, 51) = 0.07$ $p = .793$	$F(2, 102) = 2.43$ $p = .093$
VM old RT	$F(2, 100) = 40.70$ $p < .001$	$F(1, 50) = 0.02$ $p = .884$	$F(2, 100) = 0.17$ $p = .842$
VM old p (error)	$F(2, 100) = 6.31$ $p = .003$	$F(1, 50) = 1.84$ $p = .182$	$F(2, 100) = 3.86$ $p = .024$
VM new RT	$F(2, 100) = 32.71$ $p < .001$	$F(1, 50) = 18.43$ $p < .001$	$F(2, 100) = 4.84$ $p = .010^*$
VM new p (error)	$F(2, 100) = 24.48$ $p < .001^*$	$F(1, 50) = 29.69$ $p < .001$	$F(2, 100) = 7.14$ $p = .001$

Note. Asterisks denote cases in which there were violations of sphericity. Boldface entries denote cases that are statistically significant at least the $p = .05$ level. All analyses conducted in JASP

References

- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396–408.
- 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 of the United States of America*, 105, 14325–14329.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.
- Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short-and long-term recognition. *Psychological Science*, 23, 625–634.
- Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: An investigation using response time distribution models. *Psychonomic Bulletin & Review*, 19, 363–394.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269–299.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95, 492–527.
- Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms?. *Cognitive Psychology*, 22, 1–35.
- Logan, G. D., & Stadler, M. A. (1991). Mechanisms of performance improvement in consistent mapping memory search: Automaticity or strategy shift. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(3), 478–496.
- McElree, B., & Doshier, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General*, 118, 346–373.
- Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology*, 10, 465–501.
- Nosofsky, R. M. (2016). An exemplar-retrieval model of short-term memory search: Linking categorization and probe recognition. *Psychology of Learning and Motivation*, 65, 47–84.
- Nosofsky, R. M., Cao, R., Cox, G. E., & Shiffrin, R. M. (2014). Familiarity and categorization processes in memory search. *Cognitive Psychology*, 75, 97–129.
- Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1524.
- Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, 118, 280–315.
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104, 266–300.
- Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance*, 31, 608–629.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception & Psychophysics*, 31, 160–168.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84, 1–66.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127–190.
- Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153, 652–654.
- Sternberg, S. (2016). In defence of high-speed memory scanning. *The Quarterly Journal of Experimental Psychology*, 69, 2020–2075.
- Strayer, D. L., & Kramer, A. F. (1994). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 318–341.
- Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2, 409–415.
- Wolfe, J. M., Boettcher, S. E., Josephs, E. L., Cunningham, C. A., & Drew, T. (2015). You look familiar, but I don't care: Lure rejection in hybrid visual and memory search is not based on familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 41, 1576.