Familiarity and categorization processes in memory search

Robert M. Nosofsky *, Rui Cao, Gregory E. Cox, Richard M. Shiffrin

Indiana University, Bloomington, United States

A R T I C L E   I N F O

Article history:
Accepted 22 August 2014

Keywords:
Memory search
Old–new recognition
Categorization
Familiarity
Automatic processing
Response times
Exemplar model

A B S T R A C T

A fundamental distinction in tasks of memory search is whether items receive varied mappings (targets and distractors switch roles across trials) or consistent mappings (targets and distractors never switch roles). The type of mapping often produces markedly different performance patterns, but formal memory-based models that account quantitatively for detailed aspects of the results have not yet been developed and evaluated. Experiments were conducted to test a modern exemplar-retrieval model on its ability to account for memory-search performance involving a wide range of memory-set sizes in both varied-mapping (VM) and consistent-mapping (CM) probe-recognition tasks. The model formalized the idea that both familiarity-based and categorization-based processes operate. The model was required to fit detailed response-time (RT) distributions of individual, highly practiced subjects. A key manipulation involved the repetition of negative probes across trials. This manipulation produced a dramatic dissociation: False-alarm rates increased and correct-rejection RTs got longer in VM, but not in CM. The qualitative pattern of results and modeling analyses provided evidence for a strong form of categorization-based processing in CM, in which observers made use of the membership of negative probes in the “new” category to make old–new recognition decisions.

© 2014 Elsevier Inc. All rights reserved.

* Corresponding author. Address: Department of Psychological and Brain Sciences, 1101 E. Tenth Street, Indiana University, Bloomington, IN 47405, United States.
E-mail address: nosofsky@indiana.edu (R.M. Nosofsky).

http://dx.doi.org/10.1016/j.cogpsych.2014.08.003
0010-0285/© 2014 Elsevier Inc. All rights reserved.
1. Introduction

Memory-search tasks (Sternberg, 1966) are among the major vehicles used for studying how different types of practice influence controlled and automatic processes in human cognition (Shiffrin & Schneider, 1977). In such tasks, subjects are presented with a list of to-be-remembered items (the “memory set”) and are then probed with a test item. The subjects’ aim is to respond as quickly as possible, while minimizing errors, whether the probe was a member of the memory set. Old probes are termed “targets” whereas new probes are termed “distractors”.

In their studies that examined hybrid forms of memory/visual search, Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) varied the types of practice in which subjects engaged. Under varied-mapping (VM) conditions, items that served as targets on some trials would serve as distractors on other trials, and vice versa. By contrast, under consistent-mapping (CM) conditions, targets and distractors never switched roles across trials. Shiffrin and Schneider observed dramatic differences in performance across the VM and CM conditions. For VM, response times (RTs) lengthened considerably with list length (“memory set size”) and this pattern remained even after extensive practice. By contrast, for CM, following sufficient practice, RTs were nearly invariant with memory set size. The general interpretation was that performance in VM tasks required effortful, controlled information processing, regardless of the amount of practice; whereas practice in CM tasks allowed for the development of automatic information processing.

Shiffrin and Schneider (1977) developed a conceptual framework for understanding the nature of these forms of controlled and automatic processing and provided support for that framework with an extensive set of experimental results involving diverse manipulations across tasks. We will review portions of that conceptual framework in the present article. A limitation of that original work, however, is that a formal quantitative model for accounting for the detailed performance patterns of the individual subjects was not provided. Townsend and Ashby (1983) reviewed and analyzed a wide variety of formal models that have been applied to the domain of memory search, but did not consider formal accounts of the differences between VM and CM performance. Strayer and Kramer (1994) used diffusion modeling (Ratcliff, 1978) to characterize differences in performance across VM and CM memory-search conditions. For example, under conditions in which VM and CM tasks were tested in separate blocks, they found that both drift rates (i.e., rates of evidence accumulation) and response-threshold settings differed across tasks. However, their aim was not to develop a deeper process-level model of the underlying memory and cognitive processes that give rise to different rates of evidence accumulation across VM and CM conditions. One main goal of the present work was to begin to fill these gaps and aim for the development of a unified, memory-based quantitative model of performance in VM and CM memory-search tasks.

Some progress towards that goal was recently made in a study reported by Nosofsky, Cox, Cao, and Shiffrin (2014). In that study, a modern formal model of probe recognition was used to account for performance in VM and CM memory-search tasks in cases involving a wide range of list lengths (memory set sizes of 1, 2, 4, 8 and 16). The formal model was an extended version of the exemplar-based random-walk (EBRW) model of categorization (Nosofsky & Palmeri, 1997) and old–new recognition (Nosofsky, Little, Donkin, & Fific, 2011). This model predicts both accuracy and RTs, thereby extending prior exemplar models of both categorization (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986) and recognition memory (e.g. Gillund & Shiffrin, 1984; Hintzman, 1988; Kahana & Sekuler, 2002; Nosofsky, 1991; Shiffrin & Steyvers, 1997) that predicted accuracy alone. Indeed, the model joins other modern approaches that aim to unravel the nature of memory through detailed modeling of the time course of old–new recognition decision making (e.g., Rae, Heathcote, Donkin, Averall, & Brown, in press; Starns, Ratcliff, & McKoon, 2012). We describe the EBRW model in formal detail later in this article. The basic idea is that each item of a study list is stored as a separate exemplar in memory. Presentation of a test probe leads to the probabilistic retrieval of these old exemplars. The probability of retrieval is greatest for old exemplars that are highly similar to the test probe and that have high “memory strengths”. In cases in which presentation of the test probe leads to the efficient retrieval of the old exemplars, information accumulates rapidly towards an “old” response threshold and the observer makes fast “old”
responses. By contrast, in cases in which the test probe is dissimilar to the exemplars stored in memory, information accumulates rapidly towards a “new” response threshold.

Although the EBRW model accounted well for the VM and CM memory-search data reported by Nosofsky et al. (2014), there were several limitations associated with that study. First, the emphasis was on capturing only qualitative aspects of the summary patterns of performance seen in averaged subject data. For example, the focus was on modeling how group mean RTs and error rates varied across conditions. In the present article, the goal is to model detailed memory-search RT-distribution data observed at the individual-subject level (e.g., Ashby, Tein, & Balakrishnan, 1993; Donkin & Nosofsky, 2012a, 2012b; Hockley, 1984; McElree & Dosher, 1989; Ratcliff, 1978), thereby providing far more rigorous tests of the proposed model of VM and CM memory search. Second, in Nosofsky et al. (2014), each subject participated for only a single session of testing, so the performance of highly practiced subjects was not examined. In the present study, subjects each participate for between 32 and 40 100-trial sessions and so are highly practiced at the tasks.

Third, and perhaps most important, the aim in the present work is to unravel the cognitive processes involved in CM memory search to a greater extent than was possible in the previous study. As we explain more fully in the modeling section, the version of the EBRW model tested by Nosofsky et al. (2014) can be viewed as a “familiarity-only” model. As is the case for many global-familiarity models of old–new recognition (e.g., Clark & Gronlund, 1996; Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1982), the model presumes that the evidence that a test probe is “old” vs. “new” depends only on the extent to which it yields high “summed activation” of the traces stored in memory: The higher the summed activation, the greater the familiarity, and the greater is the speed and probability of an “old” response. This familiarity-only model was able to do a good job of accounting for both the VM and CM memory-search data reported by Nosofsky et al. (2014).

However, one of the key hypotheses advanced by Shiffrin and Schneider (1977) is that a major component in the development of “automatic” processing in CM memory-search tasks involves the use of categorization (for more extended discussion and debate, see, e.g., Cheng, 1985; Logan & Stadler, 1991; Schneider & Shiffrin, 1985). Note that under CM conditions, there is one fixed set of items (the “positive” category) that always receives an “old” response and a second fixed set of items (the “negative” category) that always receives a “new” response. In principle, the observer does not need to pay any attention to the current memory set in order to perform the task: If the test probe belongs to the positive category then the observer can respond “old”, and likewise for the negative category. It seems plausible that, following sufficient practice, the observer can develop these categories and use them as a basis for performing the task. In order to reduce possible confusion we should remark that Shiffrin and Schneider used tasks combining memory and visual search, and proposed two types of learning in CM conditions: One was specific to visual search and involved automatic attention drawn to the spatial position of consistently trained targets. The other, automatic categorization, is the focus of the present research: We use a paradigm that provides clear-cut evidence of the spontaneous development of such a categorization process in CM memory search and use the results to develop and assess several variants of an elaborated version of the exemplar-retrieval model.

In our main experiment, we attempt to decouple the predictions from a familiarity-only model and a familiarity-plus-categorization model of memory search. Besides testing both VM and CM performance of highly practiced subjects, the key manipulation is to include trials in which the test probe on the current list was just tested on the previous list. The use of recent negative probes is a well-known manipulation in VM memory-search tasks (e.g., Monsell, 1978) and has been used in modern work to help assess the nature of forgetting from short-term memory (e.g., Berman, Jonides, & Lewis, 2009; McKeown, Holt, Delvenne, Smith, & Griffiths, 2014); however, to our knowledge, this type of manipulation has not been used previously in CM memory search. Note that categorization is not a feasible strategy for VM, because the targets and distractors switch roles across trials, so there are no long-term “old” vs. “new” categories to learn. Thus, in VM only familiarity should operate – repeating a test probe from the previous list should increase its familiarity, thereby enhancing performance for repeated “old” test probes, but degrading performance for repeated “new” test probes, which should have longer correct-rejection RTs and higher error rates (Monsell, 1978).

The crucial question concerns performance for repeated probes in the CM condition. If only “familiarity” operates, then the qualitative pattern of effects for CM should be the same as for VM,
for the reasons outlined above. By contrast, if a categorization process intervenes, the expectation is that performance may be enhanced for both repeated “old” and “new” probes. In particular, if a new probe is tested on trial $n - 1$, then the observer’s memory of its assignment to the “new” category should be enhanced. If that probe is then repeated on trial $n$, it could result in a shorter “new” RT and a lower error rate than for non-repeated new probes.

To anticipate, the findings in our main experiment (along with converging evidence from some transfer experiments) provide both qualitative and model-based quantitative evidence in favor of the role of some form of categorization process in our CM condition. Moreover, we will argue that the form of categorization that is involved goes beyond other types of categorization demonstrated in previous related research. For example, Logan and Stadler (1991) demonstrated that CM performance is influenced by category relations between test items and members of the positive set: On “catch trials” in their experiments, observers often false-alarmed to lures that were members of the positive category but that were not in the specific memory set presented on a given trial. However, even “familiarity-only” models based on summed similarity can in principle predict such effects if, for example, CM training is presumed to increase psychological similarity relations among same-category exemplars (e.g., Nosofsky, 1986). In that case, a catch-trial lure that is a member of the positive category will have high psychological similarity to all members of the memory set in CM conditions. Such a lure will yield high familiarity, even though it did not appear on the study list. As will be seen, the present paradigm will yield evidence of categorization processes in CM that even this type of flexible-similarity familiarity model cannot handle. We will then formalize a combined familiarity-plus-categorization version of the exemplar model to account for both VM and CM memory-search performance.

2. Experiment 1

Observers were tested in both VM and CM memory-search tasks over multiple sessions. The stimuli were pictures of real-world objects of all different kinds. On each trial, memory-set size was 2, 4, 6, 8 or 16. At each set size, half the test probes were old and half were new. If the test probe was old, its serial position on the list was chosen randomly. In both the VM and CM conditions, with probability $0.20$, the test probe from trial $n - 1$ was repeated on trial $n$.

2.1. Method

2.1.1. Subjects

The subjects were four members of the Indiana University community who were paid at the rate of $12 per two sessions (every 2 sessions were completed within a one-hour time span). All subjects had normal or corrected-to-normal vision.

2.1.2. Stimuli

The stimuli were drawn from 2400 unique object images obtained from the website of Talia Konkle and described by Brady, Konkle, Alvarez, and Oliva (2008). Each image subtended a visual angle of approximately 7 deg and was displayed on a gray background. The experiment was conducted on a PC running MATLAB and the Psychophysics Toolbox (Brainard, 1997).

2.1.3. Procedure

In both the VM and CM conditions, for each subject, two sets of 32 stimuli were randomly sampled from the 2400 images and served as that subject’s stimulus sets for the entire experiment. There was no overlap between the VM and CM stimulus sets.

In the VM condition, on standard no-repeat trials, the memory set was randomly sampled from the 32 VM stimuli. If the test probe was a target (old), it was randomly chosen from this memory set. If the test probe was a distractor (new), it was randomly sampled from the remaining members of the 32-item stimulus set. On repeat trials, the test probe on trial $n$ was identical to the test probe on trial $n - 1$. On such trials, if the computer program determined that the test probe would be “old,” then
the memory set was constrained to include that test probe; whereas if the test probe was new, then the memory set was constrained to not include that test probe.

In the CM condition, for each subject, 16 stimuli were randomly chosen from the 32-item CM stimulus set to form a “positive set,” with the remaining 16 stimuli forming the “negative set.” On each standard (no-repeat) trial, the memory set was randomly chosen from the positive set. If the test probe was old, then it was randomly chosen from the memory set. If the test probe was new, then it was randomly chosen from the negative set. On repeat trials, the test probe on trial \( n \) was identical to the test probe on trial \( n - 1 \). Due to the constraints of the CM condition, the old–new status of the test probe on repeat trials was the same on trials \( n - 1 \) and \( n \). The memory set on trial \( n \) was chosen randomly from the positive set subject to this constraint.

The memory-set sizes were 2, 4, 6, 8 and 16. The size of the memory set was chosen randomly on each trial. Except for repeat trials in the CM condition, the old–new status of the test probe was chosen randomly on each trial, but the probability of “old” remained .50 overall.

In both the VM and CM conditions, the probability that the probe on trial \( n - 1 \) repeated on trial \( n \) was equal to .20. The repeat/no-repeat status of the test probe was chosen randomly on each trial.\(^1\)

Each trial began with the presentation of a fixation point (asterisk) in the center of the screen for \( .1 \) s, followed by the presentation of the memory set. Each memory-set item was presented for \( 1 \) s, with a \( .1 \)-s inter-stimulus interval. Following a \( 1 \)-s retention interval, a second fixation point (plus sign) was presented for \( .5 \) s, followed immediately by the test probe. The test probe remained on the screen until the subject responded. Feedback (“Correct!” vs. “Incorrect”) was then provided for \( 1 \) s. Subjects were instructed to respond as quickly as possible while minimizing errors. Subjects entered “old” responses by pressing the “J” key on the keyboard, and entered “new” responses by pressing the “F” key.

There were four blocks of 25 trials in each session. The computer reported to the subjects their overall percentage of correct responses at the end of each block. The VM and CM conditions were presented in alternating blocks. The order of alternation was rotated across sessions. Subjects typically completed 2 sessions per day. Subjects 1–4 completed 32, 37, 40 and 39 sessions, respectively.

2.2. Results

The first session was considered practice and was not included in the analyses. In addition, we deleted from the analysis any trial in which the RT was less than 180 ms or greater than 3000 ms (less than 1% of the data). Finally, for each subject, we computed the mean and standard deviation of correct RTs at each combination of condition (VM vs. CM), set size, old/new status of the probe, and repeat status of the probe, and then deleted from analysis any trial in which the correct RT was greater than 3 standard deviations above the mean for that combination (2.6% of the trials).

We visually inspected the patterns of mean correct RTs and the error proportions separately for Sessions 2–10 and Sessions 11–40. Although RTs were shorter and error rates were lower in the later sessions compared to the earlier ones, the qualitative pattern of results was very similar across the early and late sessions. Therefore, we decided to collapse across Sessions 2–40 in modeling the data. (We consider the practice effects in these tasks in more depth in Section 2.4 of Experiment 1.) Although we will model the data separately for each subject, all subjects showed the same qualitative patterns of results. Therefore, in describing the summary trends, we report the data averaged across the four subjects.

The mean correct RTs are plotted as a function of condition (VM vs. CM), memory set size, probe type (old vs. new), and repeat status of the probe in Fig. 1 (top panels). The mean proportions of errors are plotted as a function of these variables in the top panels of Fig. 2.

---

\(^1\) Because of the constraints involving the repeat trials in the CM condition, there is a slightly higher probability that the correct response on trial \( n \) will be the same as the correct response on trial \( n - 1 \). This same characteristic does not operate in the VM condition. We conducted detailed analyses that examined whether patterns of performance changed depending on whether the response was repeated rather than whether the previous test probe was repeated and found little if any effect of that variable in either condition.
We start by describing the results in the standard (no-repeat) conditions. In the standard VM condition (solid triangles), mean RTs for both the old and new probes get longer with increases in memory set size, and this lengthening is curvilinear in form. This pattern of results mirrors ones observed by other researchers in related designs for this range of set sizes (e.g., Burrows & Okada, 1975; Nosofsky et al., 2014; Wolfe, 2012). The error proportions in the standard VM condition show the same pattern.

---

Fig. 1. Mean correct response times (ms) plotted as a function of condition (VM vs. CM), old–new status of probe, set size, and repeat manipulation. Top panels: observed, bottom panels: predicted. VM = varied mapping, CM = consistent mapping.
In the standard CM condition (solid squares), the mean RTs for the new probes are a flat function of memory set size, whereas the mean RTs for the old probes lengthen curvilinearly with increases in set size. The set-size function in the CM condition is not as steep as in the VM condition. The error proportions (Fig. 2) are extremely low in the CM condition, although for the old probes there is a slight increase in errors as set size increases.

Fig. 2. Mean probability of error plotted as a function of condition (VM vs. CM), old–new status of probe, set size, and repeat manipulation. Top panels: observed, bottom panels: predicted. VM = varied mapping, CM = consistent mapping.
The data from the VM-repeat condition are symbolized by X’s. Perhaps the most dramatic results are that, for the new probes, compared to the standard VM condition (solid triangles), there is a major lengthening in mean RTs and a major increase in error rates in the VM-repeat condition (cf. Monsell, 1978). The mean RTs for the new probes in the VM-repeat condition are not monotonic with set size, but this pattern varied considerably across the different subjects and the irregular plot probably reflects noise due to the smaller sample sizes in the VM-repeat condition. Note that the error proportions for the new probes in the VM-repeat condition do increase in highly regular fashion as set size increases. Regarding the old probes, there is little change in mean RTs and a slight decrease in error rates (except at set-size 16) when the probe repeats from the previous trial.

The data from the CM-repeat condition are symbolized by open circles. Whereas there was a dramatic slowdown for repeat-new probes in the VM condition, there was no change in RT for the repeat-new probes in the CM condition (and error rates remain essentially at zero). In addition, mean RTs got shorter for old probes in the CM-repeat condition and error rates got even lower than in the CM no-repeat condition.

These qualitative patterns of results seem consonant with the predictions from a familiarity-only model in the VM condition but appear to challenge the predictions from a familiarity-only model in the CM condition. As we will discuss, the exemplar-based familiarity-only model naturally predicts that as set size increases, there will be a curvilinear lengthening in mean RTs and an increase in error rates for both old and new probes in the VM condition. Furthermore, as will be seen, it naturally predicts that when a new probe repeats from the previous trial, there will be a major lengthening in correct mean RTs and an increase in error rates in the VM condition: When the new probe repeats from the previous trial, its familiarity is increased, making it more difficult for the observer to correctly reject the new probe. When an old probe repeats from the previous trial, its already-high familiarity is further enhanced, leading the model to predict slight speedups in processing and reductions in error rates compared to the standard no-repeat condition.

The problem for the familiarity-only model arises mainly for new probes in the CM condition. Whereas repeating the new probe in the VM condition led to dramatic lengthening in correct RT and to increases in error rate, there is no such interference in the CM condition. Because any boost in familiarity in the VM condition should be paralleled by a similar boost in familiarity in the CM condition, the differing qualitative patterns of results across the VM and CM conditions appear to challenge the familiarity-only model. We will develop this argument in more rigorous fashion in our modeling analysis section. To anticipate, the formal modeling will suggest that performance in the CM condition is mediated by both current-list memory and by longer-term categorization processes: When a probe is repeated from the previous list, the observer's memory for the membership of that item in either the new or old category is enhanced, and the observer can use this category-membership information as a basis for making her old–new recognition judgments.

Finally, more detailed breakdowns of the patterns of performance for the old probes across conditions are displayed in Figs. 3 and 4 (top panels), which plot mean RTs and error rates as a joint function of set size and lag. “Lag” is defined as the number of items back in the study list from which a positive test probe was presented. For example, when set size is four, the item in serial position 4 has lag 1, the item in serial position 3 has lag 2, and so forth. Because of small sample sizes, to remove noise in the data, for set-size 8 the data are averaged across lags – 1–2, 3–4, 5–6, and 7–8; and for set-size 16, the data are averaged across lags 1–4, 5–8, 9–12, and 13–16. The plots reveal that, for the old probes, mean RTs and error rates depend virtually only on lag. That is, once one takes into account lag, there is little if any remaining effect of set size per se (cf., McElree & Dosher, 1989; Monsell, 1978; Nosofsky et al., 2011). Items with short lags are responded to more rapidly and with fewer errors, whereas items with long lags have longer mean RTs and higher error rates. Thus, the reason why the set-size functions in Figs. 1 and 2 show a lengthening in old mean RTs and an increase in error rates is mainly because lists with larger set sizes tend to include items with longer lags. Note that the slowdown with lag is observed not only in the VM condition but in the CM condition as well. Thus, performance in the CM condition appears to be governed not solely by long-term category representations but by memory for the current list items as well.
2.3. Formal modeling analyses

2.3.1. Outline of formal model

The EBRW model is a discrete-step random walk model. Although the model is naturally applied for predicting mean RTs and choice probabilities, in the present work our goal is to fit data at the level of individual-trial choices and RTs. To meet this goal, it is more convenient to apply continuous information-accumulation models. Thus, following Donkin and Nosofsky (2012a, 2012b), in the present work we formulate the exemplar model within a linear-ballistic-accumulator (LBA) framework (Brown & Heathcote, 2008).

Fig. 3. Mean correct response times (ms) for old test probes plotted as a function of condition (VM vs. CM), repeat manipulation, set size, and lag. Top panels: observed, bottom panels: predicted. VM = varied mapping, CM = consistent mapping.
2.3.1.1. Familiarity-only model. A schematic illustration of some of the main components of the exemplar-based LBA (EB-LBA) model is provided in Fig. 5. We start by describing the “familiarity-only” component of the model. According to the model, each item of a study list is stored as a separate exemplar in memory. Under the present conditions, the “memory strength” of each exemplar is presumed to depend solely on the lag with which it was presented on the study list. Based on evidence
reported by Donkin and Nosofsky (2012a; see also Anderson and Schooler, 1991; Wickelgren, 1974; Wixted & Ebbesen, 1991), we assume more specifically that memory strength is a decreasing power function of lag $j$,

$$m_j = a + \frac{j^{-\beta}}{C_0};$$

(1)

where $a$ is asymptotic strength and $\beta$ reflects the rate at which memory strength decreases with lag. This component of the model converges with other theoretical approaches that place emphasis on the role of the retention interval in influencing list-length effects (e.g., Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Murdock, 1985). The differential memory strengths are represented schematically in Fig. 5 (Panel A) in terms of the larger sizes of the circles that surround exemplars with shorter lags.

In the general version of the model (Nosofsky & Palmeri, 1997; Nosofsky et al., 2011), exemplars are represented as points in a multidimensional space, and similarity is a decreasing function of distance between points in the space (Fig. 5A). For the present types of stimuli, however, we apply a highly simplified model of similarity. In particular, the similarity of an exemplar to itself is set at one; whereas the similarity between any pair of distinct exemplars is given by a free parameter $s$ ($0 < s < 1$).²

The degree to which exemplar $j(e_j)$ from the study list is “activated” when test-item $i(t_i)$ is presented is a joint function of exemplar $j$’s memory strength and its similarity to test item $i$:

$$a_{ij} = m_j, \quad \text{if } t_i = e_j,$$

(2a)

$$a_{ij} = m_j s, \quad \text{if } t_i \neq e_j.$$

(2b)

Thus, the study-list exemplars that are most highly activated are those that match the test probe and that have short lags. We also presume that when a test probe is presented, there may be residual “background” activation of exemplars from previous lists (and pre-experimental experience), given by free parameter $B$. As illustrated schematically in Fig. 5B, when a test probe is presented, the exemplars stored

² A more general and conceptually important extension of the model would make allowance for within-category similarity relations to exceed between-category ones due to the operation of selective attention to category-relevant features (Nosofsky, 1986). However, this extension was not needed for modeling the present data.
in memory “race” to be retrieved, with rates that are proportional to their activations (cf. Logan, 1988). As described more fully below, this retrieval of old exemplars leads an information–accumulation process to move in the direction of an “old” response threshold in the LBA architecture.

Recognition RT models require not only a process that accumulates evidence for an item being old but also a process that provides evidence for a new decision. In most cases, these processes are instantiated in the models by assuming that observers set evidence criteria. If the evidence for “old” exceeds the criterion then the information–accumulation process moves in the direction of an “old” response threshold; whereas if the evidence fails to exceed the criterion then the information accumulates in the direction of a “new” response threshold. In Ratcliff’s (1985) terminology this type of criterion is termed the “drift-rate criterion” and it is generally conceptualized in terms of the setting of a boundary value along an evidence axis (e.g., Ratcliff, 1985, p. 215). To apply the present exemplar-retrieval model to the domain of old–new recognition RTs, we make an analogous assumption, except the notion of “criterion” is conceptualized in terms of a retrieval process instead of the setting of a boundary value. In particular, we assume that the observer establishes what we term “criterion elements” in the memory system. Just as is the case for the stored exemplars, upon presentation of a test probe the criterion elements (labeled “c” in Fig. 5B) race to be retrieved. Opposite to the old exemplars, retrieval of the criterion elements leads the information–accumulation process to move in the direction of a “new” response threshold in the LBA architecture. Thus, the criterion elements provide a mechanism that allows the system to respond “new” by competing with the retrieval of the old exemplars. As an analogy, one might imagine a neural mechanism in which the system establishes a separate negative channel of baseline firing that competes with a positive channel activated by memory-based familiarity.

Whereas the retrieval rates of the stored exemplars vary with their lag-dependent memory strengths and their similarity to the test probe, the retrieval rates of the criterion elements are independent of these factors. Instead, the criterion elements race with some fixed rate $k$, independent of the test probe that is presented. As discussed more fully below, the setting of $k$ is presumed to be, at least in part, under the control of the observer.

Finally, the retrieved exemplars and criterion elements drive the LBA process that governs old–new recognition decisions (Fig. 5C). On separate accumulators, the observer sets response thresholds $OLD$ and $NEW$ that establish the amount of evidence needed for making an “old” or a “new” response. Information accumulates in linear and ballistic fashion (i.e., without momentary noise) toward each response threshold with the rate on each accumulator determined by the retrieved exemplars and criterion elements (see below). The evidence-accumulation process has a random starting point on each accumulator that is uniformly distributed in the interval $[0, h]$. The exemplar retrieval/evidence-accumulation process continues until one of the response thresholds is reached, at which time the observer emits the appropriate response.

By way of analogy with the rate equations derived by Nosofsky and Palmeri (1997) for the discrete random walk, it is assumed that the mean rate at which information accumulates on the OLD accumulator is given by

$$ p_i = \frac{A_i}{(A_i + k)}, \quad (3) $$

where $A_i$ gives the summed activation of the test probe to all old exemplars (including the background items from previous lists):

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

and $k$ is the level of criterion–element activation. (The mean rate of accumulation to the $NEW$ response threshold is given by $q_i = 1 - p_i$.) On each trial, the “momentary” rate of accumulation, $p_i'$, is presumed to be normally distributed around the mean drift rates, i.e.,

$$ p_i' = p_i + \varepsilon, \quad (5) $$

where $\varepsilon$ is a normally distributed random variable with mean zero and standard deviation $\sigma$.

Note that test probes that match recently presented exemplars (with high memory strengths) will cause high summed activations ($A_i$ in Eqs. (3) and (4)), leading to rapid accumulation to the $+OLD$
threshold and fast old RTs. By contrast, test probes that are highly dissimilar to the memory-set items will not activate the stored exemplars, so only criterion elements will be retrieved. In this case, information will accumulate rapidly towards the NEW threshold, resulting in fast new RTs.

Through experience in the task, the observer is presumed to learn an appropriate setting of the criterion-element activation \( k \), such that 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. In this way, the LBA process will tend to move toward the appropriate response threshold on trials in which old vs. new probes are presented.

As an approximation to implementing this form of criterion adjustment, we assume that the criterion setting varies linearly with memory set-size \( M \):

\[
    k(M) = u + v \cdot M, \tag{6}
\]

where \( u \) and \( v \) are free parameters. The idea is that as set size increases, summed activation of study exemplars \( (A_i) \) will also tend to increase, so the observer may need to set a stricter criterion for responding “old”.

Finally, in cases in which the test probe repeats from the previous trial, we presume that there is a boost in the summed-activation \( (A_i) \) familiarity term, which we denote \( \text{FAM}_b \). In principle, the boost in familiarity may depend on the set size of the current list. For example, if the current list has a long length, then the previous test probe would have been presented in the more distant past. However, as we will report and discuss in our modeling analysis section, we found that allowing \( \text{FAM}_b \) to vary with set size led to minuscule improvements in absolute fit for the range of set sizes tested here. Therefore, we focus our presentation on a simple version of the model in which a single \( \text{FAM}_b \) parameter is estimated for repeated probes regardless of the set size of the current list.

This familiarity-only version of the exemplar model makes use of 11 free parameters for the VM condition and 11 free parameters for the CM condition (plus the \( \text{FAM}_b \) parameter for the repeat trials). The parameters include \( x \) and \( \beta \) in the memory-strength power function; the similarity parameter \( s \); the background-activation \( B \); criterion-setting parameters \( u \) and \( v \); response-thresholds OLD and NEW; the between-trials standard-deviation-of-drift parameter \( \sigma \); the start-point variability parameter \( h \); and a mean residual-time parameter \( T_r \) corresponding to factors not associated with recognition decision making (e.g., encoding and motor-execution times). \(^3\) A glossary of the free parameters (along with best-fitting estimates from a version of the model to be described below) is provided in Table 1.

2.3.1.2. Familiarity-plus-categorization model. The familiarity-plus-categorization model is a straightforward extension of the familiarity-only model and applies only to the CM condition. We presume that the observer forms long-term category representations for the positive and negative sets and that these representations are activated and retrieved along with the current memory-set items when the test probe is presented. (We consider the detailed nature of these category representations later in our article.) The long-term category activation for old probes is given by \( \text{CAT}_{\text{old}} \) and for new probes is given by \( \text{CAT}_{\text{new}} \). Thus, the mean correct drift rate for old probes is given by

\[
    p_{\text{old},i} = (A_i + \text{CAT}_{\text{old}})/[(A_i + \text{CAT}_{\text{old}}) + k], \tag{7a}
\]

whereas the mean correct drift rate for new probes is given by

\[
    q_{\text{new},i} = (k + \text{CAT}_{\text{new}})/[(k + \text{CAT}_{\text{new}}) + A_i]. \tag{7b}
\]

In cases in which a probe is repeated from the previous trial, these category-activation values are boosted by a parameter \( \text{CAT}_b \). Although the long-term \( \text{CAT}_{\text{old}} \) and \( \text{CAT}_{\text{new}} \) parameters are conceptually important, their values cannot be reliably identified from fits of the model to the data from the present paradigm. (The value of \( \text{CAT}_{\text{new}} \) is non-identifiable with the criterion-activation \( k \), and the value of \( \text{CAT}_{\text{old}} \) is difficult to estimate separately from the parameters that contribute to \( A_i \).) However, the hypothesized boost in category activation \( \text{CAT}_b \) that results from repeated presentations of the test probe is identifiable, and it is that boost that is the focus of the analyses involving the extended model.

\(^3\) Applications of the EBRW model to probe recognition also sometimes make allowance for a primacy-boost parameter (e.g., Nosofsky et al., 2011), but that parameter contributed negligibly to the fits to the present data.
Conceptually, there is no very strong \textit{a priori} reason to set certain of the parameters in the model at default values or to hold certain parameters constant across conditions, so we consider the version of the familiarity-plus-categorization model described above to be the “core” model. However, as will be seen, it is possible to produce far more parsimonious fits to the present data by introducing various parameter constraints and we describe these constraints when we report the detailed fits.

Before proceeding, we should clarify further the relation between the familiarity-only model and the familiarity-plus-categorization model. The familiarity-only model itself already incorporates a highly significant form of category-based processing. In particular, in their previous applications of the EBRW model to predicting VM and CM performance, Nosofsky et al. (2014) found that a crucial difference in parameter settings across tasks involved the similarity parameter ($s$ in Eq. 2). Whereas the best-fitting value of $s$ was approximately .10 in the VM condition, it was essentially zero in the CM condition. The psychological interpretation was that whereas targets and distractors were somewhat confusable in the VM condition, highly distinct memory representations were formed for members of the positive vs. negative sets in the CM condition. Nosofsky et al. (2014) argued that subjects have the opportunity to learn to attend to any category-level features that may be useful for discriminating between the members of the fixed positive and negative sets in the CM condition. This selective-attention process may increase within-category psychological similarity relations among members of the positive set, and decrease similarity relations between members of the positive and negative sets (Nosofsky, 1986). Thus, it seems reasonable that highly distinct memory representations between positive- vs. negative-set members may be formed.

Although selective attention to category-level features may operate, it is still the case that the “familiarity-only” version of the model presumes that old–new recognition is based solely on the extent to which the test probe activates memory representations of “old” exemplars (either on the current list or past lists). The model was able to capture the flat set-size functions for new probes in the CM condition of Nosofsky et al. (2014) because, regardless of set size, the summed activation of old exemplars (on the current list) yielded by the new probes was essentially zero. The present paradigm, however, allows one to distinguish between the predictions of the zero-similarity familiarity-only model and the familiarity-plus-categorization model. In particular, in cases in which a new probe on the current trial was repeated from the previous trial, it provides a perfect match to that previous test probe. The zero-similarity mechanism applies only to cases of mismatching items, not matching ones. Thus, any boost in familiarity that arises from repeating probes in the VM condition

### Table 1
Glossary of parameters and best-fitting parameter values for the varied-mapping (VM) and consistent-mapping (CM) conditions of Experiment 1.

<table>
<thead>
<tr>
<th>Parameter Brief description</th>
<th>VM</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$ Power-decay rate</td>
<td>2.564</td>
<td>5.000</td>
</tr>
<tr>
<td>$\alpha$ Strength asymptote</td>
<td>0.092</td>
<td>0.065</td>
</tr>
<tr>
<td>$s$ Similarity</td>
<td>0.202</td>
<td>0.017</td>
</tr>
<tr>
<td>$B$ Background activation</td>
<td>[0]</td>
<td>[0]</td>
</tr>
<tr>
<td>$u$ Criterion intercept</td>
<td>0.132</td>
<td>0.032</td>
</tr>
<tr>
<td>$v$ Criterion slope</td>
<td>[0]</td>
<td>[0]</td>
</tr>
<tr>
<td>$OLD$ Old response threshold</td>
<td>0.309</td>
<td>0.232</td>
</tr>
<tr>
<td>$NEW$ New response threshold</td>
<td>0.375</td>
<td>0.253</td>
</tr>
<tr>
<td>$\sigma$ Drift-rate variability</td>
<td>0.198</td>
<td>0.198</td>
</tr>
<tr>
<td>$h$ Start-point variability</td>
<td>0.026</td>
<td>0.081</td>
</tr>
<tr>
<td>$T_r$ Residual time</td>
<td>0.100</td>
<td>0.023</td>
</tr>
<tr>
<td>$CAT_{old}$ Old-item category activation</td>
<td>[0]</td>
<td>[0]</td>
</tr>
<tr>
<td>$CAT_{new}$ New-item category activation</td>
<td>[0]</td>
<td>[0]</td>
</tr>
<tr>
<td>$FAM_b$ Familiarity boost</td>
<td>0.049</td>
<td>0.047</td>
</tr>
<tr>
<td>$CAT_b$ Boost in category activation</td>
<td>0.026</td>
<td>0.162</td>
</tr>
</tbody>
</table>

Notes: Parameter values in brackets were held fixed \textit{a priori}. VM = varied mapping, CM = consistent mapping. $Sn$ = Subject $n$. For reasons of psychological plausibility, the value $T_r$ had a lower limit of .100. To prevent computer-underflow issues, the value $\beta$ had an upper limit of 5.0. Time-related parameters are measured in seconds.
should also arise in the CM condition, allowing one to decouple the predictions from the familiarity-only and familiarity-plus-categorization versions of the model. In a nutshell, the familiarity-only model predicts interference in correctly rejecting repeated new probes, whereas, depending on parameter settings, the familiarity-plus-categorization model can predict enhancement.

2.3.2. Model-fitting method

We fitted the models to the data from each subject by conducting computer searches for the values of the free parameters that yielded maximum-likelihood fits to the individual-trials choices and RTs. In other words, in this model-fitting method, the likelihood of the choice and RT on each individual trial is assessed, and the overall likelihood is the joint likelihood of all of the individual-trial likelihoods. The analytic equations for expressing the individual-trial likelihoods from the LBA model are presented by Brown and Heathcote (2008, Eqs. 1–3). The fit of each model was then assessed by using the Bayesian Information Criterion (BIC; Schwarz, 1978),

\[
\text{BIC} = -2 \ln(L) + P \ln(N),
\]

where \(L\) is the (maximum) likelihood, \(P\) is the number of free parameters in the model, and \(N\) is the total number of trials on which the fit is based. The term \(P \ln(N)\) is a penalty term that penalizes a model for its number of free parameters. The model that yields the smallest BIC is considered to provide the most parsimonious account of the data. In an attempt to avoid local minima, we used 40 different random starting configurations of the parameters in conducting the computer searches. We should emphasize that the goals for the modeling are ambitious: There are 2 mapping conditions (VM vs. CM) crossed with 2 types of test probes (old vs. new) crossed with 5 set sizes (with the set sizes for old items involving anywhere from 1 to 16 lags) crossed with the “repeat” manipulation. For each combination of conditions, the aim is to predict, at the level of individual subjects, the complete distribution of the correct RTs and also the probabilities of correct and error responses. (The modeling is also constrained by the RTs for the error responses, although these distributions are based on far less data than are the correct responses.)

2.3.3. Model-fitting results

2.3.3.1. Fits and main qualitative predictions. The model-fitting results are reported in Table 2. The table reports the BIC fits of different versions of the model to the data of each of the four subjects. Perhaps the most important comparison is between the fits of the familiarity-only model (Row 1) and those of the familiarity-plus-categorization model (Row 3). For all four subjects, the familiarity-plus-categorization model yields appreciably better BIC fits than does the familiarity-only model. In the Row-1 model, the familiarity-boost parameter (\(FAM_{b}\)) was constrained to be equal across the VM and CM conditions, for reasons outlined earlier in our article. However, even when this parameter is allowed to vary freely.
across the VM and CM conditions (Row 2), the familiarity-plus-categorization model still yields better BIC fits for all four subjects.

In the core version of the model (Row 3), we allowed all parameters to vary freely (except \( CAT_{new} = 0 \), because it is non-identifiable with \( v \)). However, inspection of the best-fitting parameters revealed that many were superfluous. Based on this inspection, we also fitted a constrained version of the familiarity-plus-categorization model in which: (i) the background-noise (\( B \)) and criterion-adjustment (\( v \)) parameters were held fixed at zero in both the VM and CM conditions; (ii) the long-term categorization-activation parameter \( CAT_{old} \) was held fixed at zero; and (iii) the standard-deviation-of-drift parameter (\( \sigma \)) was held fixed across the VM and CM conditions. We found that the absolute log-likelihood fit for this constrained model was virtually as good as that of the core model for all four subjects, so its BIC fit was appreciably better (Row 4 of Table 2). We emphasize that, in our view, the processes represented by all of these constrained parameters are of conceptual importance. The point of the analysis is simply to make clear that either: (i) we are unable to estimate some of their precise values under the present conditions of testing (e.g., the values of \( CAT_{old} \) and \( CAT_{new} \)); or (ii) some of the hypothesized processes did not play major roles in the present experiment (e.g., criterion adjustment with changes in list length [\( v \)].

Before describing other more detailed model-based analyses, we now consider the extent to which the core model captures the observed patterns of performance. The predictions of the mean correct RTs and error rates as a function of set size are shown together with the observed data in Figs. 1 and 2 (bottom panels), and the predictions of the lag by set-size functions are shown in Figs. 3 and 4 (bottom panels). Although there are some quantitative departures, the model accounts for most of the main qualitative effects very well. First, consider the standard (no-repeat) trials (solid symbols in Figs. 1 and 2). For the old items (right panels in Figs. 1 and 2), the model fits reasonably well the curvilinear increase in mean “old” RTs and error rates as a function of set size in both the VM and CM conditions. Also, as can be seen in Figs. 3 and 4 (bottom panels), the predicted lag by set-size functions match reasonably well the observed functions, with the different set-size results essentially overlapping one another within a single curvilinear lag function. These lag effects underlie the collapsed old-item set-size predictions from the model that are displayed in Figs. 1 and 2. As lag increases, memory strength of the studied exemplars decreases, so there is lowered activation of the old exemplars. Thus, there is a lowered rate of evidence accumulation towards the old response threshold, resulting in longer RTs and increased error rates. The collapsed set-size predictions for the old items (Figs. 1 and 2) emerge from the lag predictions because as set size increases, there is an increased proportion of cases in which items have long lags.

Regarding the new items in the no-repeat conditions (solid symbols in left panels of Figs. 1 and 2), the model predicts the lengthening in mean RTs and increase in error rates that are observed as a function of set size in the VM condition (although it underestimates the steepness of the RT curve). The model-based explanation is that as set size increases, the summed activation (\( A_i \)) yielded by the new items in the VM condition tends to increase. Thus, evidence accumulates less efficiently in the direction of the new response threshold (Eq. (7b)). The model also accounts correctly for the result that the mean new RTs are flat in the CM condition. Part of the explanation is that estimated similarity between members of the positive and negative sets is near zero in the CM condition (see Section 2.3.3.4). Thus, regardless of set size, summed activation for new items is near zero in the CM condition.

We now turn to the repeat trials. Recall that the results from the repeat trials are based on far fewer observations than the no-repeat trials. Even so, the model accounts for the large increase in error rates that are observed when new items repeat in the VM condition (X’s in left panels of Fig. 2) and accounts for the lengthening in mean RTs for these items (X’s in left panels of Fig. 1). It fails to predict the non-monotonic relation between mean RTs and set-size for the new probes in the VM-repeat trials. However, as discussed previously, that pattern was very inconsistent across the four subjects. In addition, there is no hint of this type of non-monotonic relation in the error-proportions data. The model also accounts successfully for the finding that mean RTs for the new items are essentially flat in the CM-repeat condition and predicts only a minuscule change for these items across the no-repeat and repeat trials (left panels of Fig. 1). Finally, it predicts the essentially zero error rates observed for the new probes in both the CM-no repeat and CM-repeat conditions. In sum, the familiarity-plus-categorization version of the exemplar model accounts successfully for the main qualitative pattern of results involving the repeat trials.
It is instructive to also consider the best-fitting predictions from the familiarity-only model, which are displayed in Fig. 6. Comparing to the observed data in Figs. 1 and 2, it can be seen that the familiarity-only model displays various qualitative shortcomings. First, it fails to predict an enhancement effect for the repeat-old items in the CM condition (open circles in right panels of Figs. 1 and 2). Second, it predicts incorrectly a small but systematic interference effect for the repeat-new items in the

**FAMILIARITY-ONLY MODEL PREDICTIONS**

![Fig. 6](image_url) Predictions from the familiarity-only version of the exemplar-based linear-ballistic-accumulator model plotted as a function of condition (VM vs. CM), old–new status of probe, set size, and repeat manipulation. Top panels: mean correct response-time predictions, bottom panels: mean probability of error predictions.
CM condition (open circles in left panels of Figs. 1 and 2). According to the model, when a new test probe is repeated, its familiarity is boosted by value $FAM_b$. If the boost is big, then the model will predict incorrectly a large interference effect for repeated new probes in the CM condition. Thus, the model settles on a smaller value of $FAM_b$, but this choice then prevents the model from predicting the degree of facilitation that is observed for repeated-old probes in the CM condition. Furthermore, because the value of $FAM_b$ is constrained to be equal across the VM and CM conditions, the model also fails to predict the magnitude of the interference that is observed for repeated new probes in the VM condition (X’s in left panel of Fig. 1).

In further analyses, we fitted other versions of the familiarity-plus-categorization model to the data from each subject. First, for completeness, we fitted a “categorization-only” version of the model that did not allow the familiarity process to play a role in the CM condition. In particular, it was presumed that the observer did not compare the test probe to the exemplars of the current list and relied solely on the longer-term category assignments of the test probes.

As reported in Table 2, this categorization-only model yielded substantially worse BIC fits than did the familiarity-plus-categorization model for all four subjects. The reason why the categorization-only model fits poorly is because it fails to predict the clear effects of current-list study-test lag that were observed for all subjects in the CM condition.

In a second special-case model, we constrained the similarity-mismatch parameter ($s$) to have fixed value across the VM and CM conditions. Recall that although this fixed-similarity model fared poorly in the tests conducted by Nosofsky et al. (2014), those analyses were restricted to the familiarity-only version of the model. As reported in Table 2, the BIC fits of the fixed-similarity model are still appreciably worse than those of the core model for Subjects 1 and 4, although they are slightly better for Subjects 2 and 3. It appears that even when making allowance for the role of a long-term category-memory process, there is still some evidence that the memory representations for items from the positive and negative sets are more distinct from one another in the CM condition than in the VM condition.

Finally, we considered various extended models that might account for the non-monotonic pattern of mean RTs observed for the new items in the VM-repeat condition. We reasoned that various competing factors might play a role. As the size of a current list increases, a repeated probe from the previous list will have occurred in the more distant past. Thus, on the one hand, the memory strength of that previously presented probe should be greater when the current list is short than when it is long, which could result in greater interference. On the other hand, if the current list is short, the observer may be able to differentiate the context of the current list from the previous one: the intuition is that the observer “knows” that the repeated probe was from the previous trial and should be ignored. The bottom line is that we tried adding a variety of parameters to capture these potential influences, but did not discover any extended models that yielded improved BIC fits compared to the core model. Thus, disentangling these complex factors pertaining to interactions between current list length and memory for previous trials remains as an issue for further research.

2.3.3.2. RT distributions. Thus far, we have focused on the pattern of mean RTs and error rates observed across the different conditions of testing. Another goal of Experiment 1 was testing the ability of the exemplar model to account for the detailed RT distributions. In Fig. 7 we display the observed correct RT distributions for Subject 1 as a function of condition (VM vs. CM), probe type (old vs. new) and set size (collapsed across the different lags). This subject’s results were representative of the results of the other subjects as well. (Because of small sample sizes, we do not conduct RT-distribution analyses for the repeat trials or for the data broken down by lags.) The observed RT distributions (solid dots) are displayed in terms of their 10%, 30%, 50%, 70% and 90% quantiles (i.e., the RT values below which each percentage of observations in the distribution falls). We also display the predicted RT quantiles (lines) from the familiarity-plus-categorization model. It is important to remember that we fitted the model to the individual-trials choices and RTs and not directly to the quantile data displayed in the figure. Nevertheless, as can be observed, with the exception of a few points at the high-variability 90% quantiles, the model appears to do an excellent job of capturing the RT-distribution data. Thus, not only does the model account for how mean correct RTs and error rates vary across conditions, it also provides a good overall account of the detailed shapes of the correct RT distributions.
2.3.3.3. Error RTs. Another challenge for models of memory search is whether they can capture patterns of error RTs along with the correct RTs. In the present situation, there is only a subset of conditions that allow for a meaningful analysis because errors were rare in various conditions. In Table 3 (top panel) we report the observed mean correct RTs and error RTs (averaged across the four subjects) as a function of set size for both old and new items in all of the VM conditions. In addition, we report the mean correct and error RTs for old items in the standard CM (no repeat) condition. In all of the remaining CM conditions, missing data (due to zero error rates for some subjects) prevent the error-RT analysis. We should note that even in the cases presented in Table 3, the mean error RTs are often based on small sample sizes, so our focus is on the qualitative pattern of results rather than quantitative model fitting. In addition to presenting the observed error RTs, we also report in Table 3 the averaged predictions from the core model (bottom panel of Table 3).

Inspection of Table 3 reveals that in the standard (no-repeat) VM condition, mean error RTs were substantially longer than mean correct RTs for both old and new probes at all set sizes. The same pattern is observed for old items in the VM-repeat condition and the CM condition. Furthermore, this qualitative pattern of results is well captured by the model. The longer error RTs in these conditions occur for two main reasons. First, recall that within any given set size, items that have greater lags give rise to higher error probabilities and to longer RTs (see Figs. 3 and 4). Thus, within any set size, a
higher proportion of items with long lags contribute to error RTs than contribute to correct RTs and these “long-lag” errors are slow. Second, as is true of many evidence-accumulation models of choice RT, the present model makes allowance for variability across trials in both the drift rate and the start point of the evidence-accumulation process. As has been explained extensively in past work on these topics (e.g., Ratcliff, Van Zandt, & McKoon, 1999), drift-rate variability leads to longer error RTs than correct RTs, whereas start-point variability leads to the opposite pattern. Our quantitative fitting of the model made allowance for the free estimation of these across-trial variability parameters and the parameter-search routine apparently chose values of the free parameters that allowed the model to capture the main qualitative pattern of error-RT results.

Finally, an interesting exception to the pattern of results described above is that for new items in the VM-repeat condition, error RTs were slightly shorter than correct RTs (except at set-size 2). Furthermore, the model predicts little difference between mean error RTs and correct RTs in this condition. Recall that it was in this VM-repeat condition that new items had dramatically longer correct RTs. Because of the repeated new test probe from the previous trial, the drift rate towards the new (error) response threshold was considerably reduced in this condition. Apparently the accumulation of evidence to the old (error) boundary will not win the race under these conditions if it occurs even more slowly than the reduced-rate drift towards the new boundary.

2.3.3.4. Best-fitting parameters. Inspection of the best-fitting parameters from the model (Table 1) provides additional insights regarding the cognitive and memory processes underlying VM and CM performance. First, replicating the results from Nosofsky et al. (2014), estimated inter-item similarity (s) is near zero in the CM condition, suggesting that the memory representations for items from the positive vs. negative sets are highly distinct from one another in that condition. (We emphasize again that the data from the present paradigm did not allow us to obtain separate estimates of the extent to which items within the positive set were psychologically similar to one another. Selection attention to category-level features might lead to increases in within-category similarity at the same time as highly distinct between-category representations are formed.) Second, the magnitudes of the response
thresholds (OLD and NEW) were greater in VM than in CM. This result replicates findings from Strayer and Kramer (1994) for cases in which VM and CM are tested in blocked fashion. Because VM training yields noisier evidence-accumulation information regarding whether items are old or new, observers need to set stricter response thresholds in that condition in order to achieve reasonably accurate performance. Third, the rate of memory decay ($\beta$) for items on the current list was greater in CM than in VM. This result may imply that observers gave less processing capacity to the current memory set in CM than in VM (relying instead on long-term memory representations), but items on the current memory set with very short lags still automatically boosted performance.

2.4. Practice effects

Because we tested only four subjects, it is difficult to pinpoint the precise time course with which differences in VM and CM performance emerged. However, to gain some information regarding that issue, in Fig. 8 we plot the average of each subject’s median RT and error rates for each of Sessions 1–2, 3–4, 5–6, 7–8 and 9–10. The data are plotted as a function of condition (VM vs. CM), item type (old vs. new) and memory set size. (There is not nearly enough data to also investigate lag and repetition effects, so we collapse across those variables in this analysis.)

Inspection of Fig. 8 reveals that there is already a big difference between VM and CM in overall RTs and error rates for both old and new probes by Sessions 3–4, and also evidence that these effects begin to emerge even in Sessions 1–2. For example, in Sessions 1–2, the error rates for the lures in the CM condition are essentially at zero and do not increase with memory set size. By Sessions 3–4, the RT function for the CM lures is also a flat function of memory set size. Thus, in general, it appears that the major differences between VM and CM emerged fairly quickly in the present paradigm. By comparison, the types of automaticity observed in studies such as those of Schneider and Shiffrin (1977) appear to have developed more slowly. However, the Shiffrin–Schneider paradigms used visual search and a crucial component of the practice effects they observed involved automatic visual attention drawn to the spatial position of consistently trained targets. By contrast, the present paradigm involved only memory search, and subjects would have received consistent training on all positive-set members multiple times by even the first block of the first session of training. Although future research is needed to unpack these effects, it appears that highly frequent presentations of all individual positive-set members, coupled with even infrequent presentations of individual negative-set members, allows highly distinct memory representations to be rapidly formed when subjects receive consistent training. Another interesting aspect of the practice-effect results is the convergence of the CM-old RTs to the CM-new RTs during the early sessions. It is possible that early in CM training attention to memory set items can slow responding, but that a gradual transition to reliance upon a long-term category strategy can reduce such costs, leaving only the residual and automatic effect of priming by the memory set items.

2.5. Discussion

In sum, the familiarity-plus-categorization EB-LBA model provides a good overall account of these diverse and intricate patterns of results involving varied- and consistent-mapping, old/new status of probes, set size of memory list, lag of positive probes, and repeat status of both old and new probes. Beyond accounting for the overall patterns of mean RTs and error rates as a function of conditions, the model captures reasonably well the detailed RT distributions for each subject. Perhaps the major qualitative result of note is that whereas repeating new probes from the previous list led to dramatic interference in the VM condition, there was no interference in the CM condition, and there was strong enhancement in the CM condition when old probes were repeated. Even when making allowance for major changes in psychological similarity across members of the positive and negative sets in the CM condition, the familiarity-only version of the model failed to capture these effects. Thus, the qualitative pattern of results and detailed model-based analyses support the hypothesis that longer-term category representations contributed in a major way to the CM performance of these practiced observers. In Experiments 2A and 2B we seek converging evidence for this role of categorization in CM memory search.
Fig. 8. Average of median response times (ms) and error probabilities plotted as a function of condition (CM vs VM), old–new status of probe, and set size during the early sessions of training. Session numbers are indicated at the top of each individual panel. CM = consistent mapping, VM = varied mapping.
3. Experiment 2A

The goal of Experiment 2A was to test for converging evidence for the contribution of categorization processes (above and beyond “familiarity” processes) to the CM memory-search results. The same four highly practiced observers from Experiment 1 participated in a transfer experiment. In addition to presenting memory sets and test probes from the positive and negative sets of the CM condition, we also included “all-new” (AN) items in the design. The AN items were sampled from the complete set of 2400 images available for use in the experiment (and that were not already used in the CM and VM conditions of Experiment 1). For each subject, any given AN item was used as a memory-set item or as a distractor at most once in each session of Experiment 2.

There were four main conditions defined both by the type of memory set presented on each trial and the type of distractor item that was presented (on negative-probe trials). The CM+/CM− condition was a continuation of the CM condition tested in Experiment 1. In the CM+/AN− condition, the memory set consisted of CM items, but the distractors were AN items. In the AN+/CM− condition, the memory set consisted of AN items, but the distractor was from the negative set of CM items. Finally, in the AN+/AN− condition, the memory set consisted of AN items, and the distractor was also an AN item.

We reasoned that to the extent that “familiarity” alone was the driving force behind observers’ recognition judgments, then observers should be fastest to correctly reject AN− items, because these items had never been experienced during the extensive training phase. By comparison, the CM− distractors had been experienced numerous times, so would be far more familiar than the AN− items, making it more difficult to correctly reject them. An advantage for the AN− lures might also be predicted by theories that posit an active search for negative evidence in the form of extra-list features that are not part of any members of current or recent memory sets or past probes (Mewhort & Johns, 2000). In contrast, we reasoned that to the extent that a “categorization” process was operating, then observers should be fastest to correctly reject the CM− items, because they belonged to the well-learned category of negative items.4

3.1. Method

3.1.1. Subjects

The subjects were the same as in Experiment 1 and we continued to pay them at the same rate.

3.1.2. Stimuli

For each subject, the CM+ and CM− stimuli were identical to the ones used in Experiment 1. The AN+ and AN− stimuli were drawn from the remaining members of the 2400 object images that had not been used in the CM and VM conditions of Experiment 1.

3.1.3. Procedure

The basic procedure was the same as in Experiment 1, although we did not include the repeat manipulation. The different conditions of testing (CM+/CM−, CM+/AN−, AN+/AN−, AN+/CM−) were randomly mixed within blocks. The condition, old/new status of test probe, memory set size, and lag of positive probe were chosen randomly on each trial. Note that on trials in which a positive probe was presented, there is no logical distinction between the CM+/CM− and CM+/AN− conditions, and no logical distinction between the AN+/AN− and AN+/CM− conditions.

Again, there were four blocks of 25 trials in each session. Each subject completed 8 sessions, typically 2 consecutive sessions per day. Thus, each subject completed 800 trials, about 400 trials in which the probe was old and 400 trials in which the probe was new (and so approximately 100 trials of old and new in each main condition). Because there were again 5 different set sizes (2, 4, 6, 8, 16), each subject completed approximately 20 trials of old and 20 trials of new at each set size of each main condition.

4. Banks and Atkinson (1974) conducted related experiments that compared performance in VM conditions to AN conditions; however, in their paradigm, as well as in the experiments of Nosofsky et al. (2014), types of test probes were not mixed across different types of study lists. For example, if a study list consisted of AN items, then a lure would always be an AN item. By contrast, in our present experiment, across both CM and AN study lists, we vary whether the lure is a CM lure or an AN lure.
3.2. Results

For each subject, we computed the median correct RT and the proportion of errors separately for old probes and new probes at each set size in each of the four conditions. We display the average (across subjects) of these median RTs as well as the average error probabilities in Fig. 9.

Regarding the old probes (right panels of Fig. 9), RTs are longer and error rates are higher in the AN conditions than in the CM conditions. This pattern is as expected according to both the familiarity-only and familiarity-plus-categorization hypotheses (CM+ probes are more familiar than AN+ probes and also belong to the well-learned positive category). Also as expected, RTs for the AN+ probes get longer with increasing set size (for the same reasons as in the VM condition of Experiment 1). Interestingly, even after 32–40 sessions of practice, there is still a slight slowdown with set size for the CM+ items, suggesting a continued role of memory for the current list.

The more important results are the RTs and error rates for the new probes (left panels of Fig. 9). The most dramatic result is that RTs are much longer and error rates are much higher in the AN+/AN− condition than in the other three conditions. The comparison between the AN+/AN− and AN+/CM− conditions supports the predictions from the categorization hypothesis. The AN− items are far less familiar than are the CM− items, but the CM− items have been consistently assigned to the negative category. Subjects appear to be able to make use of these consistent category assignments to efficiently classify the CM− items as new.

On the other hand, there is little difference in performance between the CM+/CM− and CM+/AN− conditions. The hypothesis that subjects make use of the well-learned negative category assignment of the CM− items would seem to predict that subjects could respond more quickly in the CM+/CM− condition. It may be that in cases in which the memory set consists of CM+ items, subjects can efficiently evaluate whether a test probe is not a member of that extremely well-learned positive category, and there is little room for improved performance beyond use of that strategy.

The results of Experiment 2 tend to support the hypothesis that a categorization process plays a role above and beyond familiarity in CM memory search, but there are several interpretations of this conclusion. For example, one could argue that as a byproduct of the extensive practice, CM− items become less similar than AN− items to the AN+ items. In that case, the CM− items would give rise to lower summed activation than would the AN− items, allowing a “familiarity”-only model to account for the main qualitative pattern of results. However, one should ask why similarity might operate in this manner. It seems psychologically plausible for highly distinct representations to have developed between the CM+ and CM− sets, because subjects have been trained extensively to distinguish between those sets. However, the AN+ and CM− categories involve a whole new set of discriminations for which no training has been provided. Thus one would want to argue that training makes the category of CM− items more ‘distinctive’ generally. The present data are insufficient to distinguish this general distinctiveness hypothesis from the direct category-use hypothesis.

4. Experiment 2B

In a second follow-up, we tested the four subjects in another transfer phase that was closely related to the one from Experiment 2A. The main purpose was to seek generality for our finding that the efficiency of correct rejections is enhanced when the negative probes are drawn from the CM− set. We sought this generality by comparing performance on the different negative probes under a wider variety of testing conditions. In addition, we wished to explore how memories for AN+ and CM+ items might influence one another if such items were mixed together in the same memory sets.

We tested three main conditions. First, we continued to test the condition in which the memory set was composed purely of AN+ items (pure-AN); on trials in which the probe was a distractor, it could again come from either the AN− or CM− sets. Second, we tested a condition in which the first half of each list was composed of AN+ items and the second half was composed of CM+ items. We will refer to this condition as the sequential AN/CM condition (seq-AN/CM). One question was whether the AN+ items would continue to show the same strong effects of lag when the second half of each list was composed of CM+ items. For example, because the CM+ items were already so well learned, they might no longer interfere with the traces of other items that the subjects were trying to remember. Third, we...
tested a condition in which the memory sets were again composed of half AN+ items and half CM+ items; however, now the items within each list were mixed in a random order (mixed-AN/CM). Because the CM+ items were so well learned, we hypothesized that in the seq-AN/CM condition, some subjects might not even attend to the second half of each list once presentations of the CM items began. Indeed, subjects might even try to rehearse the earlier AN items. The present mixed-AN/CM design, however, would force the subject to attend to each memory-set item because he or she could
not anticipate the onset of a CM-only sequence of items. Thus, it provides an interesting alternative
test of the extent to which CM items will interfere with to-be-remembered AN items. Finally, we
decided to test the present conditions in a between-blocks design, with each condition tested in a sepa-
rate block, to allow subjects to develop any strategies that might be conducive to performance in
each condition.

4.1. Method

4.1.1. Subjects and stimuli

The subjects and stimuli were the same as in Experiment 2A.

4.1.2. Procedure

As described in the introduction to Experiment 2B, there were three main conditions defined by the
different types of memory sets: pure-AN, seq-AN/CM, and mixed-AN/CM. In the mixed-AN/CM con-
tion, the serial position of each item type was chosen randomly on each trial, subject to the constraint
that half the items were AN and half were CM. In each condition, negative test probes could be either
AN— or CM—. In the pure-AN condition, positive test probes could be only AN+; whereas in the other
two conditions, positive test probes could be either AN+ or CM+. The memory set sizes were again 2, 4,
6, 8 and 16.

The three conditions were tested in separate blocks of 25 trials each. Each condition was tested
once every three blocks in a random order. Subjects completed two sets of three blocks per session
of testing and each subject completed four sessions. On each trial, the memory set size, old/new status
of the probe, type of probe (AN or CM), and serial position of positive probes was chosen randomly,
subject to the constraints of the experimental design described above. All other aspects of the proce-
dure were the same as in Experiment 2A.

4.2. Results

The mean correct RTs and error rates across the three conditions are displayed in Figs. 10 and 11. To
allow for controlled comparisons across the three conditions, the results for positive probes are broken
down according to whether the probe occupied the first or second half of each memory set. (Recall
that in the seq-AN/CM condition, AN probes always occupied the first half of each list.)

One clear-cut result is that, in all three conditions, mean RTs are shorter and error rates tend to be
lower for the CM lures (CM—) than for the AN lures (AN—). This result lends considerable generality to
the main finding from Experiment 2A and again suggests a strong role for a learned “new” category in
influencing the subjects’ old–new recognition decisions.

Another clear-cut result is that in the mixed-AN/CM condition, mean RTs are shorter and error rates
are much lower for CM targets (CM+) than for AN targets (AN+), regardless of whether the items occu-
pied the first or second halves of each list. Although the same basic result was observed in Experiment
2A, note that in the present case it is observed in a within-list comparison, so cannot be attributed to
factors such as changed response-threshold settings across different list types. One interpretation is
that observers make efficient use of the long-term “old” category membership of the CM+ items in
making their recognition judgments. Another closely related interpretation is that CM+ items have
a greater degree of “self-match” to their memory traces than do AN+ items, yielding greater summed
activation for CM+ probes.

An additional result concerns the comparison of performance on the AN+ items and the CM+ items
across the three list types. Recall that we were interested in exploring how the inclusion of CM+ items
on a list might influence observers’ memories of the AN+ items. On the one hand, error rates are
greater and mean RTs tend to be longer for the AN+ items when they appear together with CM+ items
(i.e., in the seq-AN/CM and mixed-AN/CM conditions) than when they appear alone (in the pure-AN
condition). Indeed, the magnitude of the error rates on the AN+ items is rather dramatic in the
mixed-list conditions, especially at the larger set sizes. This result is the opposite of what we had
hypothesized, namely that the highly distinct traces of the CM items might cause less interference
with to-be-remembered AN items. Note as well that the patterns of performance tend to go in the
opposite direction for the new items: error rates tend to be lower for the AN\(-\) items in the seq-AN/CM and mixed-AN/CM conditions than in the pure-AN condition. Thus, one possibility is that the results are reflecting changed biases for responding “old” vs. “new” across the conditions. Another possibility

![Fig. 10. Transfer Phase 2: Mean correct response times (ms) plotted as a function of condition (pure-AN, seq-AN/CM, mixed-AN/CM), set size, type of test probe (AN+, AN\(-\), CM+, CM\(-\)), and whether the test probe was a member of the first or second half of each study list. AN = all-new, CM = consistent mapping. “+” = positive test probe (old), “-” = negative test probe (new).](image1)

![Fig. 11. Transfer Phase 2: Mean probability of error plotted as a function of condition (pure-AN, seq-AN/CM, mixed-AN/CM), set size, type of test probe (AN+, AN\(-\), CM+, CM\(-\)), and whether the test probe was a member of the first or second half of each study list. AN = all-new, CM = consistent mapping. “+” = positive test probe (old), “-” = negative test probe (new).](image2)
is that in cases in which CM items are members of the memory sets, observers place more emphasis on use of a long-term positive-category membership strategy and de-emphasize short-term memory for the current list members: If the test probe is a CM+ item, respond “old”, otherwise respond “new”. Such a strategy would lead to worse performance on AN+ items in mixed lists than in pure lists, but to the opposite pattern for AN– items. This hypothesis involving a relative emphasis on whether a test probe is or is not a CM+ item (in cases in which CM+ items were members of the study list) is also consonant with the pattern of results that we observed in Transfer Phase 1.

5. General discussion

5.1. Summary

In their studies that examined varieties of controlled and automatic human information processing, Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) observed dramatic differences in the patterns of visual/memory search performance exhibited by subjects trained under varied-mapping (VM) vs. consistent-mapping (CM) conditions. Although they provided a conceptual framework for understanding these performance differences, they did not develop or test a formal quantitative model. One aim of the present work was to fill that gap and move towards the development of a cognitive-based formal quantitative model of memory-search performance under both VM and CM conditions.

Nosofsky et al. (2014) made some preliminary progress along these lines by demonstrating that, with appropriate choice of parameter settings, a modern exemplar-retrieval model of probe recognition provided a good qualitative account of VM and CM memory-search performance. However, the present research goes well beyond that recent work in several important respects. First, in Nosofsky et al. (2014), the demonstrations were limited to showing that a version of the exemplar model could capture the main qualitative pattern of results seen in averaged subject data. By contrast, in the present work, our aim was to develop and test versions of the model that could quantitatively capture detailed RT distributions and choice probabilities observed at the individual-subject level. Second, Nosofsky et al. (2014) examined the behavior of subjects who participated in the tasks for only a single session. By contrast, a central theme in the studies of Shiffrin and Schneider was to demonstrate how forms of automatic processing developed in subjects who received extensive CM practice. Thus, in the present work, our goal was to model the performance of highly practiced subjects who participated in the tasks for numerous sessions. Third, and perhaps most important, our goal in the present work was to conduct manipulations to test for a central role of “categorization-based” processing in CM memory search that operates along with “familiarity-based” processing. As we clarify in our General Discussion, these demonstrations of categorization-based processing, and their implementation in the extended formal model, go significantly beyond important related demonstrations contributed in some past research.

In brief, according to our proposed model, the items from a memory set are stored as separate exemplars in memory, with strengths that vary as a function of the lag with which they were presented on the list. Traces from previous lists are stored as well. In the case of CM practice, because the traces from previous lists are consistently assigned to either the “old” or the “new” category, observers can make use of their memories for these long-term category assignments to help make their old–new recognition judgments. When a test probe is presented, the stored exemplars and/or remembered category assignments are retrieved, and these retrieved sources of information drive an evidence-accumulation process for making old–new recognition judgments. The retrieval of old exemplars from the memory set drives the evidence-accumulation process towards an “old” response threshold, whereas retrieval of “criterion elements” (a familiarity-based mechanism for responding “new”) drives the evidence-accumulation process towards a “new” response threshold. The direct retrieval of any long-term category-based information occurs along with the retrieval of exemplars from the memory set, and will also drive the evidence-accumulation process towards the appropriate threshold. We refer to the retrieval of exemplars from the memory set (or from previous lists) as “familiarity-based” processing, whereas the retrieval of long-term category assignments constitutes “category-based” processing.
The model outlined above provided an excellent account of the wide range of empirical phenomena reported in this article. Under VM conditions, mean RTs and error rates for both old and new probes were curvilinear increasing functions of memory set size (at least for the range of set sizes tested here). In the case of old probes, however, the set-size effects appeared to be derivative of a more fundamental effect of study-test lag. In particular, mean RTs and error rates for old probes were curvilinear increasing functions of the lag with which the test probe was presented, and the various set-size functions were largely overlapping within this lag function. Similar results have been reported by Monsell (1978), McElree and Dosher (1989) and Nosofsky et al. (2011) in the case of a smaller range of set sizes, and more recently by Nosofsky et al. (2014) for the large range of set sizes tested here. The exemplar-retrieval model predicts these effects because memory strength of stored exemplars is a decreasing function of their lag. Thus, the efficiency with which the old exemplars are retrieved decreases as their lag increases. As explained in more depth by Nosofsky et al. (2014), the drift rates that emerge from the exemplar-retrieval process decrease in curvilinear fashion with increases in lag, and this factor is among the key properties underlying the exemplar model’s predictions. The lengthening in old-item mean RTs that is seen with increases in set size arises because, as set size increases, there is an increased proportion of cases in which the test probe has a long lag. Regarding the new probes, as set size increases, the summed activation yielded by a new probe also increases (see Eq. (4)), and this increased activation of old exemplars takes the evidence-accumulation process in the wrong direction.

The model also provided an excellent account of the pattern of results in the standard CM condition. As expected, for both old and new probes, performance in the CM condition was far better than in the VM condition, including shorter mean RTs and much lower error rates. In addition, the set-size function for the new probes was flat, although we continued to observe a curvilinear increase in mean RTs and error rates for the old probes as their lag increased. The slope of this lag function was much lower in the CM condition than in the VM condition. The exemplar model accounts for this overall pattern of results due to several factors. First, estimated similarity of new probes to the consistently-mapped old exemplars is nearly zero under the current conditions of testing. Thus, regardless of set size, when a new probe is presented, it yields essentially zero summed activation of the old exemplars, so evidence accumulates towards the new response threshold in extremely efficient fashion. For old probes, however, just as is the case in VM, the memory strength of a positive probe decreases with its lag of presentation on the current list, so the summed activation to which it gives rise decreases with lag. Note that the lag functions observed for the positive probes suggest strongly a continued role of “familiarity-based” processing even in the CM condition, i.e., a role of memory for the exemplars on the current list. Finally, although the present design did not allow us to obtain reliable estimates of the long-term category activations associated with old and new probes in the CM condition, it seems likely to us that this process plays a major role as well.

Perhaps the most important new qualitative effect from the present study involves the pattern of results from our “repeat” manipulation. When the probe from the previous list was repeated as a test probe on the current list, it yielded a dramatic dissociation in the pattern of results for new probes across the VM and CM conditions. In particular, in cases in which the test probe was new, there was a considerable lengthening in mean correct RTs and a considerable increase in error rates in the VM condition, replicating well known results of the effect of recent negatives on short-term memory performance (e.g., Berman et al., 2009; Monsell, 1978). Critically, however, there were no such adverse effects in the CM condition. According to the exemplar model, repeating a new probe from the previous list increases the summed activation to which it gives rise (i.e., increases its “familiarity”). Because only the familiarity component operates in the VM condition, this increase in summed activation takes the evidence-accumulation process in the wrong direction, i.e., more strongly towards the old response threshold and less strongly towards the new response threshold. The result is a dramatic increase in false alarms and a lengthening of correct-rejection RTs. By contrast, in the CM condition, a categorization process operates along with the familiarity process. When a new probe is repeated from the previous list, the observer’s memory for its assignment to the “new” category is enhanced. This enhanced memory for membership in the new category offsets the adverse effects of boosts in familiarity, and allows the exemplar model to account for the effects of the repeat manipulation in the CM condition.
5.2. Familiarity vs. categorization

Previous researchers have also provided important evidence for a role of categorization strategies in CM memory search. In our view, however, the present evidence points to an additional form of categorization than shown in these previous studies.

For example, Logan and Stadler (1991) contrasted “process-improvement” accounts of CM search with “process-switching” mechanisms. According to process-improvement accounts, a single mechanism operates across VM and CM conditions and the mechanism simply gets more efficient with CM practice. By contrast, in process-switching accounts, new strategies, such as categorization, arise in CM situations. To provide evidence for the role of some form of categorization, Logan and Stadler introduced catch trials in their CM designs. Following practice in CM search, observers were presented with a single catch trial, in which the memory set consisted of some subset of the complete positive set of CM+ items, but in which the negative probe was some member of the positive set not included in the current memory set. Logan and Stadler observed very high false alarm rates or else very slow correct rejections on the catch trial. Thus, it appeared that the memory search process had not simply become more efficient under CM training; instead, the results were consistent with the view that some form of long-term categorization process was an important part of the picture. Logan and Stadler concluded that “These results rule out the possibility that practiced CM performance depends (entirely) on comparing probes with an explicit representation of the memory set in short-term memory...” (p. 493).

Although we strongly agree with Logan and Stadler that some form of categorization process is operating, our view is that a flexible-similarity version of the “familiarity-only” exemplar model can account for their catch-trial findings. Recall that we hypothesize that, as a result of extensive CM training (or due to pre-experimental learning), observers may learn to selectively attend to category-level features that are useful for discriminating between the classes of old and new items (e.g., Schneider & Fisk, 1984). Because of these selective-attention processes, there may be increases in within-category similarity among items, and decreases in between-category similarity (Nosofsky, 1984, 1986). Thus, although the negative probe on the catch trial is not an exact match to any single memory-set item, it still may have high psychological similarity to numerous of those items in the memory set, which could cause high false-alarm rates or lengthened correct-rejection RTs.

Our present findings involving the repeat trials go beyond the previous evidence reported by Logan and Stadler because they suggest that matches to members of the negative category can promote recognition decision-making under CM conditions. Even if there are changes in psychological similarity among items, a familiarity-only model predicts that repeat presentations of the same negative probe should interfere with the ability to correctly reject that probe. The lack of interference observed in our CM condition, coupled with the strong interference in the VM condition, suggests strongly the use of the long-term “negative-set” categorization strategy in the CM condition.

5.3. On the nature of the categorization process in CM

A question left unanswered in our present research concerns the detailed nature of the categorization process that operates in CM memory search. Logan and Stadler (1991), for example, distinguished between several different types of categorization-related process-switching mechanisms. According to item-based learning accounts, observers store category associations with each of the items that receive consistent mappings during training. When presented with a test probe, memories for those long-term individual-item associations are used to enable old–new recognition decisions. Indeed, such accounts are consonant with exemplar models of categorization (Medin & Schaffer, 1978; Nosofsky, 1986), which posit that categories are represented in terms of individual stored items along with their associated category feedback. According to this view, training on individual items is fundamental to CM performance. An alternative view posits a category-comparison strategy in which the observer learns general categories that distinguish positive-set members from negative-set members, but learning is not necessarily specific to each item. Thus, there may be some higher-order category nodes stored in memory and observers may learn that objects assigned to node A should receive “old” responses whereas objects assigned to node B should receive “new” responses. Old–new recognition decisions for individual items are not based directly on the items but rather on the category node to which they are assigned.
Following the completion of Experiments 1 and 2 in this article, we conducted an additional transfer experiment designed to distinguish between these hypotheses. Unfortunately, the subjects were not tested for a sufficiently long period to yield clear conclusions. The key idea in this final transfer study was to conduct a reversal paradigm, in which formerly negative-set items now composed the memory sets, and formerly positive-set items now became the distractors. Furthermore, in the initial part of this reversal phase, only half the items from each of the positive and negative sets were retrained. Then, in the second phase of the reversal paradigm, the remaining items were introduced, again with reversed category assignments. We expected to observe a good deal of interference in the first phase of the reversal paradigm, which would eventually be overcome as the reversal training proceeded. The key question was what would happen in the second phase, when the remaining items were reintroduced. To the extent that observers learn to assign higher-order category nodes to old vs. new responses, we expected little remaining interference in the second phase, because reassignment of the category nodes would already have taken place during the first phase of reversal training. By contrast, to the extent that it is individual item-learning that is involved, we expected renewed interference at the start of the second phase, because observers still need to unlearn the item-category associations that were formed during CM training. Unfortunately, we apparently did not conduct the first reversal phase for a sufficiently long period of time, because subjects continued to show interference at the end of that phase (relative to their performance in the standard CM condition). Therefore, the results from the second reversal phase were not informative.5

In addition to testing these types of reversal tasks at time of transfer, other related paradigms could also provide deeper insights into the nature of the memory and categorization processes that underlie VM and CM performance. As one example, one might vary the degree of consistent training across individual items, analogous to studies that used such manipulations in the context of visual detection paradigms (e.g., Schneider & Fisk, 1982). Such paradigms could provide evidence of the role of long-term categorization processes that do not rely solely on memory for the test probe from the previous trial. As a second example, we hope to test categorized VM tasks in future research (e.g., Schneider & Shiffrin, 1977). In such tasks, there are again two long-term categories of items, just as in CM. However, unlike CM, the categories switch roles across trials in terms of whether they serve as targets or distractors. One possibility is that the sole benefit of CM training compared to VM training in memory search lies in the presence of the long-term categories. The presence of those categories allows both for selective attention to relevant category-level features as well as use of a positive–negative categorization strategy to supplement or bypass familiarity-based memory search. Another possibility, however, is that CM training confers something above and beyond categorization. For example, in the context of their hybrid visual/memory search paradigms, Shiffrin and Schneider provided evidence for the existence of “automatic attention attraction” responses attached to consistently trained members of the positive set. Possibly, one might find evidence of analogous processes in pure memory-search paradigms. For example, the very high frequency with which members of the positive set are experienced during training may yield high long-term memory strengths and forms of “automatic memory attraction” that are fundamental to performance.

Along with conducting these deeper empirical inquiries, future research should also investigate alternative formal models for how CM memory search operates. Our present proposed modeling approach combines a categorization-based strength with an individual-exemplar-based strength into a single drift rate. Of many alternatives, perhaps the simplest is a ‘race’ model in which the two processes operate separately and in parallel, the response being governed by the winner of the race (e.g., Strayer & Kramer, 1990). We plan further studies with conditions that have the potential to allow these model classes to be distinguished. In sum, future research is needed to further unravel the precise nature of the categorization process whose operation we have highlighted in this article, and,

---

5 The distinction between item-based and category-based performance is not as clear cut as this discussion seems to imply. If as in most recent models item encoding is allowed to converge on features of greatest utility to solve the task (say those that best distinguish the two CM categories from each other), then a category code can be viewed as just another feature. Possibly the models could be distinguished if the item features are defined to be those pre-existing prior to training (and re-weighted during training) and category features are new ones formed as a result of training. These are subtle distinctions and a resolution must await further research.
more generally, to understand the interactions of short-term and long-term memory and categorization in probe recognition.

Acknowledgments

This work was supported by a Project Development Team within the ICTSI NIH/NCRR Grant Number TR000006 to Robert Nosofsky and by Grant FA9550-12-1-0255 from the Air Force Office of Scientific Research to Richard Shiffrin.

References


