Journal of Experimental Psychology:
Learning, Memory, and Cognition
2014, Vol. 40, No. 6, 1524–1539

0278-7393/14/$12.00 http://dx.doi.org/10.1037/xlm0000015

An Exemplar-Familiarity Model Predicts Short-Term and Long-Term Probe Recognition Across Diverse Forms of Memory Search

Robert M. Nosofsky, Gregory E. Cox, Rui Cao, and Richard M. Shiffrin
Indiana University Bloomington

Experiments were conducted to test a modern exemplar-familiarity model on its ability to account for both short-term and long-term probe recognition within the same memory-search paradigm. Also, making connections to the literature on attention and visual search, the model was used to interpret differences in probe-recognition performance across diverse conditions that manipulated relations between targets and foils across trials. Subjects saw lists of from 1 to 16 items followed by a single item recognition probe. In a varied-mapping condition, targets and foils could switch roles across trials; in a consistent-mapping condition, targets and foils never switched roles; and in an all-new condition, on each trial a completely new set of items formed the memory set. In the varied-mapping and all-new conditions, mean correct response times (RTs) and error proportions were curvilinear increasing functions of memory set size, with the RT results closely resembling ones from hybrid visual-memory search experiments reported by Wolfe (2012). In the consistent-mapping condition, new-probe RTs were invariant with set size, whereas old-probe RTs increased slightly with increasing study–test lag. With appropriate choice of psychologically interpretable free parameters, the model accounted well for the complete set of results. The work provides support for the hypothesis that a common set of processes involving exemplar-based familiarity may govern long-term and short-term probe recognition across wide varieties of memory- search conditions.

Keywords: memory search, math modeling, response times, short-term memory, recognition

Supplemental materials: http://dx.doi.org/10.1037/xlm0000015.supp

A fundamental issue in cognitive science concerns the mental processes that underlie memory search and retrieval. These processes are often investigated by measuring both accuracies and response times (RTs) in tasks of probe recognition. In such tasks, observers are presented with a list of to-be-remembered items and then must classify a test probe as "old" or "new" as rapidly as possible while minimizing errors. In this article we present tests of a modern exemplar-familiarity model of memory search in tasks of probe recognition. As we will describe, the model builds upon and extends classic theories in the domains of categorization and memory and ties them together with evidence-accumulation models of decision making. We will show that the model provides a remarkably coherent account of a diverse set of results involving both short-term and long-term probes of memory across conditions that place different demands on the memory system. We will explore in particular the effects of varied versus consistent mappings, in which targets and foils either may switch roles across trials or instead receive fixed classification assignments. This manipulation will be shown to have dramatic effects upon memory-search performance, in ways analogous to those shown in studies of attention and visual search (e.g., Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). The proposed model will capture these effects with parameter choices that can be interpreted in ways aligning with those early studies. In addition, the model will be shown to also provide a viable process-level account of intriguing results reported recently by Wolfe (2012; Cunningham & Wolfe, 2014), who reported systematic functional relations between RT and memory set size in cases in which set size was varied across a wide range. In a nutshell, the work will advance the hypothesis that a common set of processes involving exemplar-based familiarity and retrieval may govern both long-term and short-term probe recognition across diverse forms of memory search.

This article was published Online First April 21, 2014.

Robert M. Nosofsky, Gregory E. Cox, Rui Cao, and Richard M. Shiffrin, Department of Psychological and Brain Sciences, Indiana University Bloomington.

This work was supported by a project development team within the Indiana Clinical and Translational Institute (National Institutes of Health/National Center for Research Resources Grant TR000006 to Robert M. Nosofsky) and by Grant FA9550-12-1-0255 from the Air Force Office of Scientific Research to Richard M. Shiffrin. We thank Michael Kahana and Jeremy Wolfe for their comments on an earlier draft of this article and for useful discussions.

Correspondence concerning this article should be addressed to Robert M. Nosofsky, Department of Psychological and Brain Sciences, 1101 East Tenth Street, Indiana University, Bloomington, IN 47405. E-mail: nosofsky@indiana.edu

Background

In the seminal "memory-scanning" paradigm introduced by Sternberg (1966, 1969), observers maintain short lists of items in memory and are then presented with a test probe. The observers' task is to classify the probe as "old" or "new" as rapidly as possible while minimizing errors. Under Sternberg's conditions of testing,

the result was that mean RTs for both old and new probes were linearly increasing functions of the size of the memory set. Furthermore, the RT functions for the old and new probes were parallel to one another. These results led Sternberg to formulate his classic serial-exhaustive model of memory search. Since that time, a wide variety of other information-processing models has been developed to account for performance in the task (for reviews and analysis, see Reed, 1973 and Townsend & Ashby, 1983).

One modern formal model of short-term probe recognition is the exemplar-based random walk (EBRW) model (Nosofsky, Little, Donkin, & Fific, 2011; Nosofsky & Palmeri, 1997). According to this model, short-term probe recognition is governed by the same principles of global-familiarity and exemplar-based similarity that are theorized to underlie long-term recognition and forms of categorization (Clark & Gronlund, 1996; Gillund & Shiffrin, 1984; Hintzman, 1988; Kahana & Sekuler, 2002; Medin & Schaffer, 1978; Murdock, 1985; Nosofsky, 1986, 1991; Shiffrin & Steyvers, 1997). The model assumes that each item of a memory set is stored as an individual exemplar in memory. When a test probe is presented, it causes the individual exemplars to be retrieved. The exemplars that are most readily retrieved are those that are highly similar to the test probe and that have the greatest memory strengths. Finally, the retrieved exemplars produce a combined result that gives rise to a familiarity-based evidence-accumulation process that determines the speed and the accuracy of old–new recognition decisions.

Nosofsky et al. (2011) and Donkin and Nosofsky (2012a, 2012b) showed that this exemplar-retrieval model provided excellent accounts of RTs and accuracies in a wide variety of short-term memory-search paradigms. The present research applies this modeling approach to two major empirical extensions. The first is to apply the model to a situation that involves both short-term and long-term probes of memory by including longer list lengths in the paradigm. This extension is aimed at bridging the gap between applications of exemplar-based familiarity models to short-term and long-term probe-recognition RTs and accuracies. The second extension is to investigate from a model-based perspective how relations between targets and foils across trials influence the process of probe recognition. Thus, we examine how relations between previously experienced memory sets and current sets impact performance. We expand upon the significance of these extensions below.

Bridging the Gap Using Short and Long Lists

The hypotheses that global familiarity and exemplar-based similarity govern long-term recognition and forms of categorization have been central ones in the field of cognitive psychology for decades (e.g., Gillund & Shiffrin, 1984; Hintzman, 1986, 1988; Medin & Schaffer, 1978; Nosofsky, 1986). The idea that those very same principles may underlie short-term probe recognition is less widely held, but evidence in favor of that hypothesis has been mounting in recent years (e.g., Donkin & Nosofsky, 2012a,2012b; Kahana & Sekuler, 2002; Nosofsky et al., 2011). More rigorous support for the idea would arise, however, if one could show that an exemplar-familiarity model accounted parsimoniously for probe recognition involving both short and long lists within the same experimental paradigm. We pursue that aim by further testing the EBRW model in the present work.

Furthermore, this aim of bridging short-term and long-term probe recognition with the EBRW model is a timely one, given intriguing results reported recently by Wolfe (2012). Following some of the early hybrid memory and visual search paradigms of Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977), Wolfe conducted experiments in which observers maintained lists of items in memory and then searched through visual arrays to locate whether a member of the memory set was present. Extending Shiffrin and Schneider's investigations, however, Wolfe tested not only memory sets that included a small number of items but ones that contained 8 or 16 items (and, in an extended paradigm, 100 items). Under his conditions of testing, he found that mean RTs were extremely well described as a logarithmic function of memory set size. In a related earlier investigation, Burrows and Okada (1975) examined memory search performance in cases involving memory sets composed of 2 through 20 items. Mean RT was well described as either a logarithmic or a bilinear function of memory set size. In this article, we explore the hypothesis that the principles of exemplar-based retrieval and global familiarity may provide an account of the curvilinear relation between mean RT and set size observed in probe-recognition paradigms that include longer list lengths.

Although the results from Wolfe (2012) and Burrows and Okada (1975) provide important targets for formal modeling, some aspects of the procedures used in these studies complicate the direct application of the exemplar-retrieval model. First, in both studies, the amount of time observers studied individual objects varied across the different memory-set sizes in an uncontrolled manner. From the perspective of the exemplar-retrieval model, the "memory strengths" associated with individual objects from the memory sets are therefore unknown. Furthermore, in both studies, observers were tested repeatedly on the same memory-set items for multiple trials. Because individual items from small memory-set sizes would serve as test probes more often than individual items from large memory-set sizes, effects of memory reinstatement at time of test could have exerted an impact on the patterns of results. Third, in the procedures used by Wolfe and by Burrows and Okada, the study–test lags (number of items intervening between a study item and a positive test probe) are unknown. It is often observed, however, that study–test lag exerts a major impact on performance in probe-recognition paradigms (e.g., McElree & Dosher, 1989; Monsell, 1978; Nosofsky et al. 2011; Ratcliff, 1978). Furthermore, as will be seen, study–test lag is assumed to be a fundamental controlling variable according to the exemplar-retrieval model. Finally, in the procedures used by Wolfe and by Burrows and Okada, performance was nearly error free (because observers studied long lists for greater periods of time than they studied short ones). Under such conditions, any speed–accuracy trade-offs that may vary across different memory-set sizes cannot be evaluated. Furthermore, the presence of errors would provide deeper and more challenging constraints for the evaluation of formal models of probe recognition.

Thus, the present research was designed to control the factors just mentioned. In our experiment, on each trial, subjects were sequentially presented a memory set consisting of 1, 2, 4, 8, or 16 items; each item was presented for the same fixed time. Following the memory set, observers were given a single test probe that they evaluated as old or new. Thus, (a) amount of study time is held roughly constant for each individual item across the different set

sizes; (b) effects due to repeated testing of individual items are greatly reduced; (c) study–test lag can be precisely measured; and (4) error data are produced along with the RT data to provide deeper constraints for the formal modeling. These procedures improve the ability to evaluate the predictions from the exemplar-familiarity model.

Relations Between Targets and Foils Across Trials

The second major extension in our studies involves the way targets and foils relate to each other across trials. We tested subjects in three conditions. Following the language from Shiffrin and Schneider (1977), in the varied-mapping (VM) condition, items that served as positive probes (old targets) on some trials might serve as negative probes (foils) on other trials and vice versa. In the consistent-mapping (CM) condition, one set of items always served as positive probes, and a second set always served as negative probes. Finally, in an all-new (AN) condition, on each trial, a completely new set of items formed the memory set (see also Banks & Atkinson, 1974). The VM condition places the greatest demands on the current list context by forcing the observer to discriminate whether a given item occurred on the current list rather than previous ones. The AN condition requires the observer to remember the current list, but it requires less contextual discrimination than VM because no target or foil had been presented on earlier lists (some contextual discrimination is presumably needed because some items on previous trials are similar to the test item on the current trial). The CM condition allows (but does not require) the observer to rely solely on long-term memory and to ignore the current-list context.

Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) demonstrated dramatic differences in patterns of performance across VM and CM conditions in their hybrid memory-visual search paradigms. VM conditions showed the usual pattern that performance depended on list length, and this pattern remained as practice continued. However, in CM conditions performance tended to become invariant with list length as practice continued (performance measured by RT in single-frame trials and accuracy in multiple-frame trials). We shall observe similar patterns in the present research (with AN performance intermediate between CM and VM).

The contrasting patterns of performance across VM and CM conditions in visual/memory search are among the most fundamental empirical results reported in the field of cognitive psychology, and they provide valuable information concerning how different forms of practice and experience influence controlled versus automatic human information processing. Yet, although Shiffrin and Schneider provided a conceptual theoretical account of the performance patterns in their VM and CM conditions, they did not develop a formal quantitative model. Our aim in the present work is to begin to make headway toward developing a unified formal-modeling account of memory-search performance across VM, CM, and AN conditions. We believe that the development of a successful, unified formal model will yield deeper insights into the cognitive processes that mediate varieties of memory search. To anticipate, we shall see that the present exemplar-retrieval model can account for the results from all three of these conditions, albeit with some parameters that differ widely across conditions. We will suggest that these parameter differences can be interpreted to align with the conceptual accounts of the role of VM and CM training provided by Schneider and Shiffrin. Finally, in the General Discussion we will elaborate the theme that the present modeling can bring together prior research and theory on attention and automatism, visual and memory search, short- and long-term memory retrieval, and categorization.

Experiment: Method

Subjects

The subjects were 150 undergraduates from Indiana University who participated in partial fulfillment of an introductory psychology class requirement. There were 50 subjects in each of the three conditions.

Stimuli

The stimuli were 2,400 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 degrees and was displayed on a gray background. The experiment was conducted on PCs running MATLAB and the Psychophysics Toolbox (Brainard, 1997).

Procedure

In the AN condition, a new set of stimuli was randomly sampled from the complete set of 2,400 images on each individual trial. No stimulus was used more than once in the experiment (unless it was an old test probe for the current list). In both the VM and CM conditions, for each individual subject, a set of 32 stimuli was randomly sampled from the 2,400 images and served as that subject's stimulus set for the entire experiment. In the VM condition, on each trial, the memory set was randomly sampled from those 32 stimuli. If the test probe was a foil, it was randomly sampled from the remaining members of the 32-stimulus set. In the CM condition, for each individual subject, 16 stimuli were randomly sampled and served as the positive set, and the remaining 16 stimuli served as the negative set. On each trial, the memory set was randomly sampled from the positive set. If the test probe was a foil, it was randomly sampled from the negative set.

The memory-set sizes were 1, 2, 4, 8, and 16. The size of the memory set was chosen randomly on each individual trial. The status of the test probe (old or new) was chosen randomly on each individual trial. If the test probe was old, its serial position on the study list was chosen randomly on each trial.

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 interstimulus 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.

Each subject participated for 5 blocks of 25 trials each. The computer reported to the subjects their overall percentage of correct responses at the end of each block.

Results

The first block was considered practice and was not included in the analyses. In addition, we deleted from analysis any trial in which the RT was less than 180 ms or greater than 5,000 ms (less than 1% of the data). Finally, on trials in which set size was equal to one and the probe was a foil, it was clear from our initial analyses that observers sometimes did not realize that they were being tested; most telling, a significant subset of these trials had very delayed RTs.[1] On these trials, the single study item was preceded by an asterisk and then the single test probe was preceded by a plus sign. This distinction was apparently not always sufficient to alert the subject that the test probe was being presented. Therefore, we delete from the modeling analyses trials in which memory-set size was equal to one.

The mean correct RTs are displayed as a function of conditions (VM, AN, CM), set size, and probe type (old vs. new) in Figure 1 (top panel). The mean proportions of errors are displayed as a function of these variables in Figure 2 (top panel). Mirroring the results from Wolfe (2012) and Burrows and Okada (1975), the mean RTs in the VM and AN conditions get substantially slower as set size increases, and this slowdown is curvilinear in form. In particular, the slowdown in RTs occurs at a decreasing rate as set size increases. This pattern is roughly the same for the old and new probes. The slowdown is much smaller in the CM condition and may be limited to the old probes. Unlike Wolfe's and Burrows and Okada's data, there are substantial proportions of errors in most of the conditions (see Figure 2). The overall pattern of error data is very similar to the mean RTs, the main exception being a pronounced increase in errors for new items in the VM condition at set-size 16.

In addition, across all set sizes and for both old and new probes, mean RTs are slower and error rates are higher in the VM condition than in the AN condition. (Banks & Atkinson, 1974, observed the same result in an experiment that involved only small set sizes. We consider their experiment in more depth in our General Discussion.) Clearly, mean RTs are much faster and error rates are much lower in the CM condition than in the other two conditions.

A more fine-grained breakdown of the data for the old probes is provided in Figures 3 and 4 (left panels), which plot mean correct RTs and error proportions as a joint function of set size and lag. "Lag" is defined as the number of items *back* in the study list with which the old probe was presented. For example, when set size is four, the item in the fourth serial position has lag 1, the item in the third serial position has lag 2, and so forth. To reduce noise, the data are averaged across lags 1–2, 3–4, 5–6, and 7–8 for set-size 8 and across lags 1–4, 5–8, 9–12, and 13–16 for set-size 16.

Inspection of the figures suggests that nearly all the effects of set size on the old items are due to the differential lags with the old items were tested (see also Nosofsky et al., 2011). That is, once one takes into account lag, there is little if any additional effect of set size per se. Across all conditions, old items with greater lags are responded to more slowly and with greater error probability. Thus, a major reason for the overall old-item set-size effects displayed in Figures 1 and 2 is that shorter lists tend to have items with smaller lags. To allow easier comparison of the lag effects across the different conditions, the results from the VM, AN and CM conditions are plotted together in Figures 5 and 6 (top panels). With few exceptions, at each lag, mean RTs are slowest and error probabilities are greatest in the VM condition, whereas mean RTs are fastest and error probabilities are lowest in the CM condition.

Formal Modeling Analyses

Outline of Formal Model

A schematic illustration of some of the main components of the EBRW model is provided in Figure 7. According to the model, each item of a study list is stored as an individual exemplar in memory. Under the present conditions, the "memory strength" of each individual exemplar is presumed to depend solely on the lag with which it was presented on the study list. On the basis of evidence reported by Donkin and Nosofsky (2012a; see also Wickelgren, 1974; Wixted & Carpenter, 2007), we assume more specifically that memory strength is a decreasing *power function* of lag *j*,

$$m_j = \alpha + j^{-\beta}, \tag{1}$$

where α is asymptotic strength and β reflects the rate at which memory strength decreases with lag.[2] The differential memory strengths are represented schematically in Figure 7 (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 et al., 2011; Nosofsky & Palmeri, 1997), exemplars are represented as points in a multidimensional space, and similarity is a decreasing function of distance between points in the space (see Figure 7A). 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 ($0 < s < 1$).[3]

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:

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

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

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 is residual "background" activation of exemplars from previous lists (and pre-experimental experience), given by free parameter *B*. As illustrated schematically in Figure 7B, when a test probe is presented, the exemplars stored in memory "race" to be retrieved, with rates that are proportional to their activations (cf. Logan, 1988).

To apply the EBRW model to the domain of old–new recognition, one assumes that the observer establishes "criterion

[1] Mean new-probe RTs were slower at set-size 1 than at set-size 2 in all three conditions (AN: 756 vs. 724; VM: 771 vs. 734; CM: 663 vs. 602).

[2] It is evident that other functions similar in shape to the power function would produce similar predictions. The power function is simple and works well over the ranges of lags in the experiments to which the model has been applied.

[3] In elaborated versions of the model (e.g., Nosofsky et al., 2011), sensitivity in ability to discriminate among distinct stimuli is also presumed to decrease with lag.

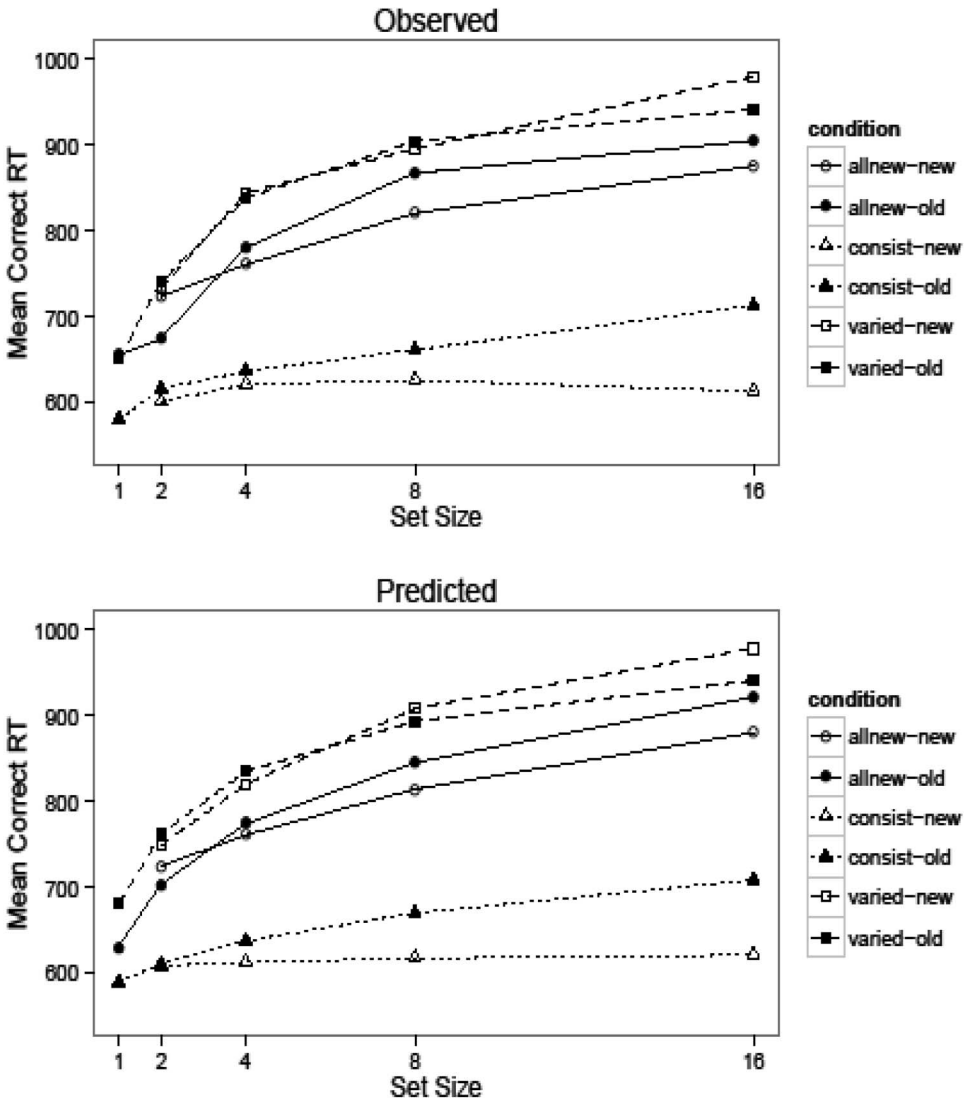

Figure 1. Mean correct response times (RTs) for old probes and new probes plotted as a function of set size in the varied-mapping, all-new, and consistent-mapping conditions. Top panel = observed, bottom panel = predicted. consist = consistent.

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 Figure 7B) race to be retrieved. However, 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 a random-walk process that governs old–new recognition decisions (see Figure 7C). The observer sets response thresholds $+OLD$ and $-NEW$ that establish the amount of evidence needed for making an "old" or a "new" response. On each step of the random walk, if an old exemplar wins the retrieval race, the random-walk counter takes a step in the direction of the $+OLD$ response threshold; whereas if a criterion element wins the race, then the counter takes a step in the direction of the $-NEW$ threshold. Thus, retrieval of criterion elements implements a mechanism for making "new" responses. The retrieval process continues until one of the response thresholds is reached.

Given more detailed processing assumptions described by Nosofsky and Palmeri (1997), it turns out that on each step of the random walk, the probability that the counter steps in the direction of the $+OLD$ threshold is given by

$$p_i = A_i/(A_i + k), \tag{3}$$

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

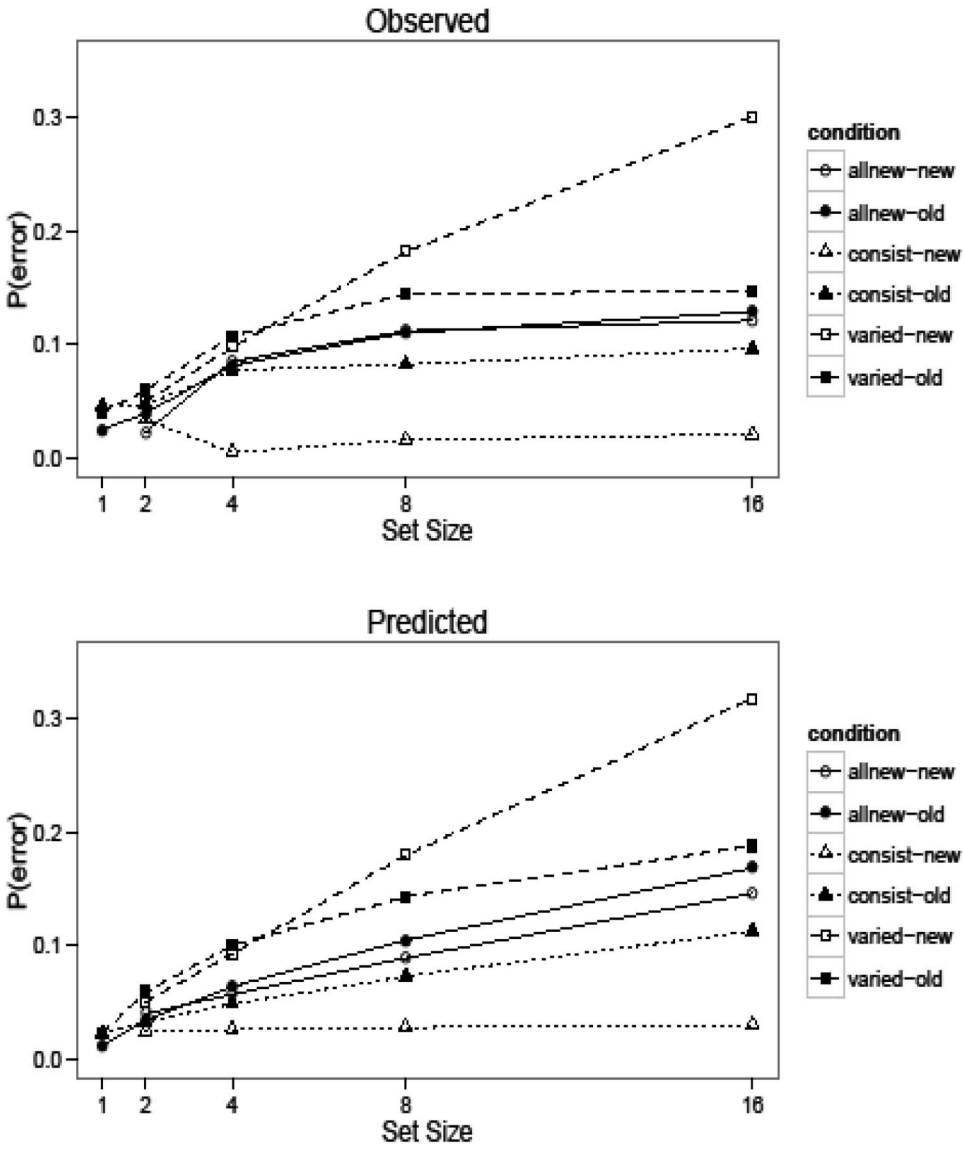

Figure 2. Mean error proportions for old probes and new probes plotted as a function of set size in the varied-mapping, all-new, and consistent-mapping conditions. Top panel = observed, bottom panel = predicted. consist = consistent.

$$A_i = \sum a_{ij} + B, \tag{4}$$

and k is the level of criterion-element activation. Note that test probes that match recently presented exemplars (with high memory strengths) will cause high summed activations (A_i), leading the random walk to march quickly to the $+OLD$ threshold and resulting in 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, the random walk will march quickly to 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 random walk will tend to step toward the appropriate response threshold on trials in which *old* versus *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:

$$k(M) = u + v \cdot M, \tag{5}$$

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 needs to set a stricter criterion for responding "old."

This version of the EBRW model makes use of 10 free parameters: the parameters α and β in the memory-strength power function; the similarity parameter s; the background-activation B;

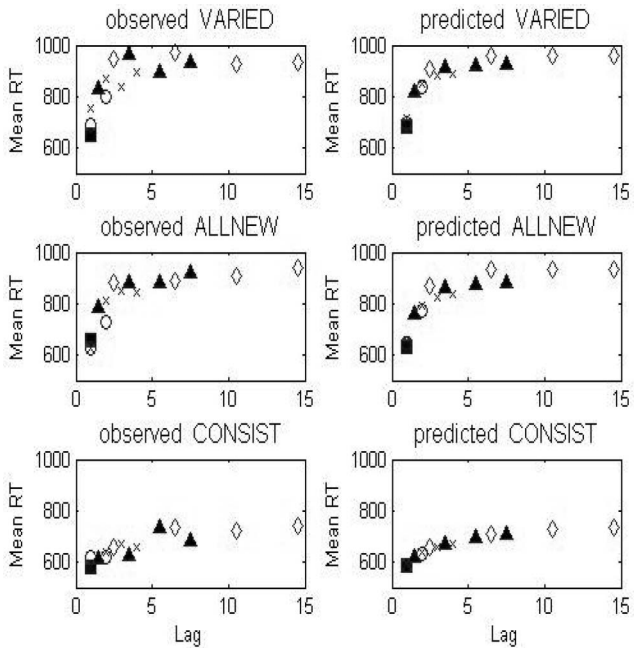

Figure 3. Mean correct response time (RT) for old probes plotted as a joint function of set size and lag in the varied-mapping, all-new, and consistent-mapping conditions. Left panels = observed, right panels = predicted. Solid squares = set-size 1, open circles = set-size 2, crosses = set-size 4, solid triangles = set-size 8, open diamonds = set-size 16. CONSIST = consistent.

criterion-setting parameters *u* and *v*; response-thresholds *OLD* and *NEW*; a scaling constant κ for transforming the number of steps in the random walk into units of time; and a mean residual-time parameter T_r corresponding to factors not associated with recognition decision making (e.g., encoding and motor-execution times). The equations for predicting mean RTs and choice probabilities from the model were reported by Nosofsky and Palmeri (1997, pp. 269–270). The application of the equations involves use of simple analytic formulas rather than requiring simulation or numerical integration.

Applying the Model

We had three interrelated goals in applying the formal model to the data. The first was to assess the ability of the model to account in parsimonious fashion for the major trends in performance across the three conditions. The second was to assess in more rigorous fashion the manner in which the model parameters varied across conditions. The third was to consider what those parameter changes might imply about memory storage and retrieval.

To pursue the first goal of testing whether the model could capture the major trends in performance, we fitted different versions of the model to the averaged data. For simplicity, we used a weighted least-squares criterion of fit. In particular, we fitted the model to the mean RT and error proportions data of the (a) new items as a function of set size and (b) the old items as a joint function of set size and lag. When fitting group RT and error-proportion data one needs to decide how to weight each component of the data in determining overall fit. We found that reasonable results were obtained when the RT data (measured in seconds) were given 5 times the weight of the error-proportion data and the individual data points for the new items were given 4 times the weight of the individual data points of the old items. (Sample sizes for the new-item data points are much greater than for the old-item data points because they are not broken down by lag.) We use more rigorous methods of model evaluation based on hierarchical Bayesian model fitting in our subsequent analyses but note that the two methods paint a similar picture.

Rather than allowing all parameters to vary freely, we were interested to discover whether some parameters could be held fixed across conditions without a significant decline in overall fit. In this initial stage of model evaluation, we addressed this question informally by assessing relative changes in the weighted sum-of-squared deviations (WSSD) and relying on visual inspection of the model-fitting results. Note that the CM condition differs in principle from the VM and AN conditions in the sense that subjects can perform the task without relying on memory for any individual list. In particular, subjects can form long-term categories corresponding to the positive and negative sets and classify each test probe based on its category membership (cf. Shiffrin & Schneider, 1977). Although efficient use of this categorization strategy may require more extended practice than is available in a single session of testing, it seems likely that it plays at least some role even under the present conditions. Therefore, in evaluating the ability of the EBRW model to fit the memory-search data with some parameters fixed across conditions, we treated the CM condition separately from the VM and AN conditions.

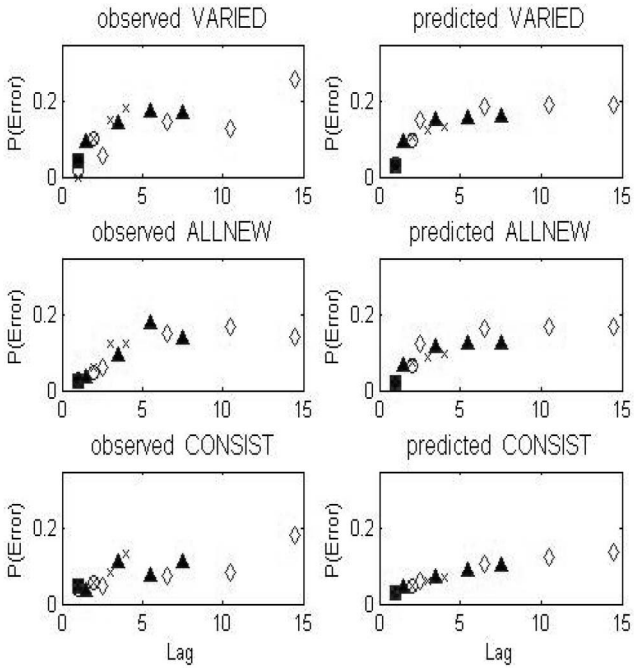

Figure 4. Mean error proportions for old probes plotted as a joint function of set size and lag in the varied-mapping, all-new, and consistent-mapping conditions. Left panels = observed, right panels = predicted. Solid squares = set-size 1, open circles = set-size 2, crosses = set-size 4, solid triangles = set-size 8, open diamonds = set-size 16. CONSIST = consistent.

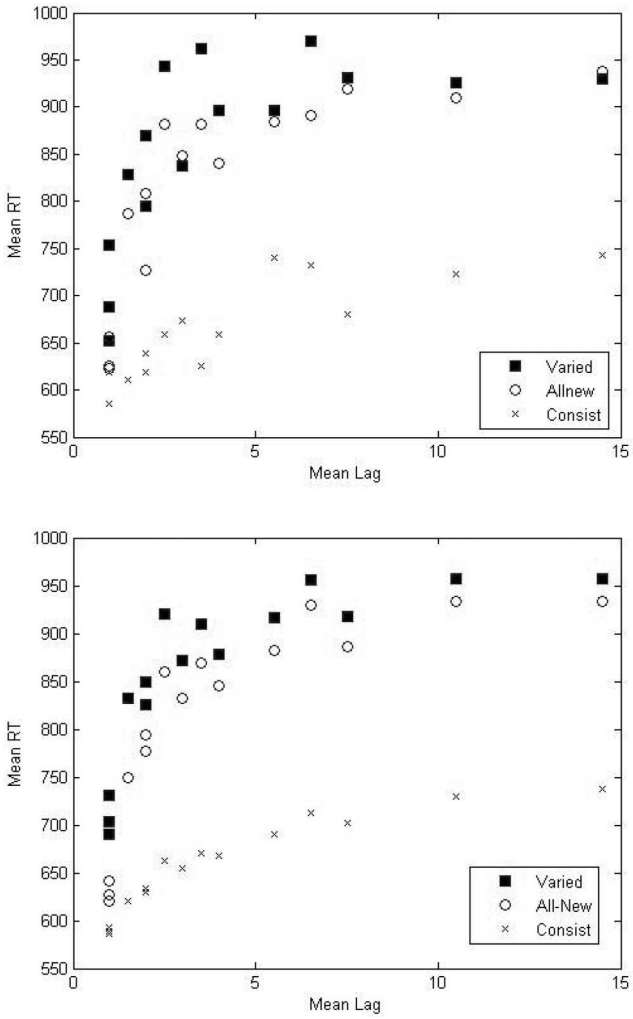

Figure 5. Mean correct response time (RT) for old probes plotted as a function of lag in the varied-mapping, all-new, and consistent-mapping conditions. Top panel = observed, lower panel = predicted. consist = consistent.

Fits to Group Data

The WSSD between predicted and observed mean RTs and error probabilities are listed for a series of different versions of the EBRW model in Table 1. The full version of the model (Version 1) allows all 10 parameters to vary freely across each of the three main conditions (VM, CM, AN) and provides a baseline for comparison of more constrained versions of the model. In Version 2, we assume that the scaling parameter κ and residual-time parameter T_r are invariant across the three conditions. As can be seen in Table 1, this constrained version yields a WSSD fit that is essentially identical to the full version. In Version 3 (the "core model"), we impose the further constraints that the β, α, *OLD*, *NEW*, and *B* parameters are identical across the VM and AN conditions. (We imposed these constraints because results from fits of the full version of the model suggested near equality of these parameters across the VM and AN conditions.) Again, the increase in WSSD is relatively small.

The predictions from this core version of the EBRW model are shown alongside the observed data in each of Figures 1–6. The best fitting parameter values are reported in Table 2. In brief, the model appears to capture the major trends in performance extremely well: the curvilinear increase in mean RTs as a function of set size that is observed for old and new probes in both the VM and AN conditions (Figure 1); the increase in error proportions for old and new probes that is observed as a function of set size in these conditions (Figure 2); the finding that mean RTs are slower and error proportions are greater in the VM condition than in the AN condition; the finding that RTs are much faster in the CM condition than in the other conditions, and that error rates are lower, particularly for the new probes (Figures 1 and 2); and the joint lag by set-size functions observed for the mean RTs and error rates across all three conditions (Figures 3–6).

Next we provide some intuition about the basis for these predictions. Because stimuli with shorter lags have greater memory strengths, the summed activation (A_i) is greatest for old test probes with short lags, resulting in fast mean RTs and low error rates for these stimuli, and also causing the dependence of old-probe-item

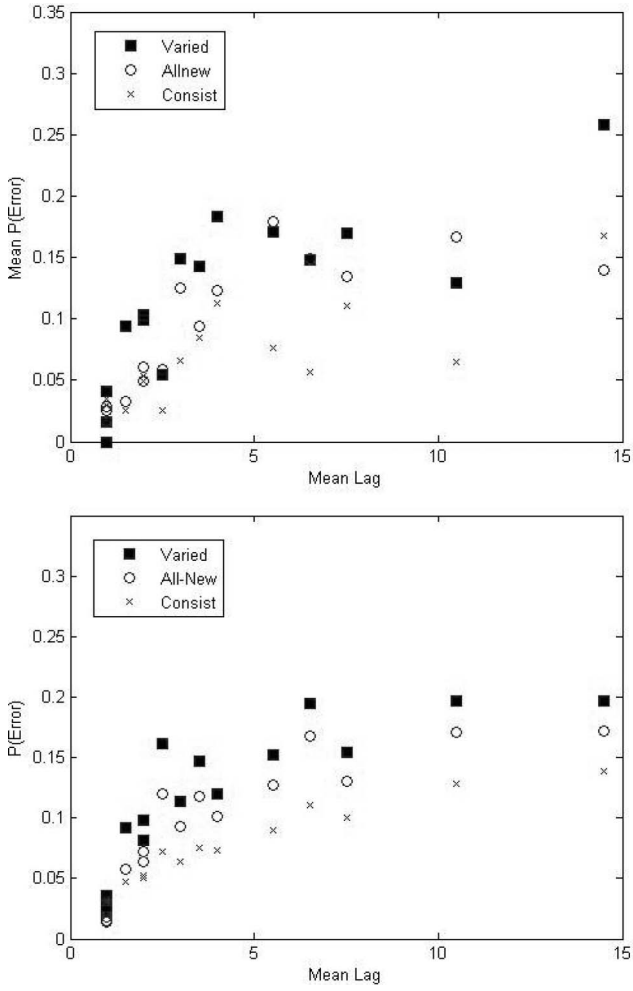

Figure 6. Mean error proportions for old probes plotted as a function of lag in the varied-mapping, all-new, and consistent-mapping conditions. Top panel = observed, lower panel = predicted. consist = consistent.

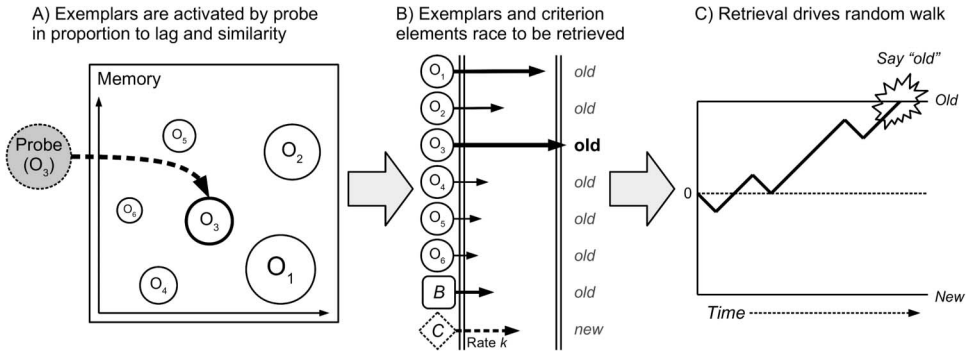

Figure 7. Schematic illustration of the workings of the exemplar-based random-walk model as applied to the probe-recognition paradigm. Panel A: Old exemplars (O) are activated in proportion to their memory strength (which is a function solely of lag) and their similarity to the test probe. Panel B: The old exemplars (O), background elements (B), and criterion elements (c) race to be retrieved with rates that depend on their activations. Panel C: The retrieved exemplars, background elements, and criterion elements drive a random-walk process for making old–new recognition decisions. Each time that an old exemplar or background element is retrieved, the random walk steps toward the OLD criterion; each time that a criterion element is retrieved, the random walk steps toward the NEW criterion.

RT upon memory set size because longer lists tend to include stimuli with greater lags. Furthermore, across a broad range of parameter settings, as lag increases, the old-item step-probabilities in the random walk decrease toward .5, first rapidly and then more gradually. This property lies at the core of the model's predictions that old-item mean RTs and error probabilities increase in curvilinear fashion with increases in lag and set size, at least for the range of different set sizes tested in the present paradigm. Turning to new probes, we note that summed activation increases as set size increases. As a result the probability that the random walk takes correct steps toward the *–NEW* threshold *decreases*, so mean RTs for the new probes get slower. Again, the changes in magnitude of these new-item step probabilities tend to be curvilinear with set size, a core property of the model.

A key parameter change that allows the model to account for the differences in performance across the VM, AN, and CM conditions is the change in the value of the similarity parameter *s*. (In Version 4 of the model we constrained the similarity parameter *s* to be equal across the VM, AN, and CM conditions. As reported in Table 1, this constraint led to a steep increase in the WSSD compared to the full version of the model, suggesting that the changes in similarity across the three conditions are highly significant. We also observed a large increase in WSSD even if the parameter *s* was constrained to be equal across only the VM and AN conditions.) As reported in Table 2, the psychological similarity between distinct objects is greatest in the VM condition, intermediate in the AN condition, and near-zero in the CM condition. Note that these differences in similarity are accompanied by

Table 1
Weighted Sum of Squared Deviation (WSSD) Fits of Different Versions of the EBRW Model to the Mean Correct RTs and Error-Probability Data of Experiment 1

Model version	AN	VM	CM	Total
1. Saturated	.0131	.0193	.0105	.0430
2. Fixed scale and residual	.0131	.0195	.0112	.0438
3. Core model	.0143	.0209	.0112	.0463
4. Fixed similarity	.0151	.0486	.0423	.1060
5. Fixed power decay	.0720	.0498	.0265	.1482
6. Zero asymptote	.0314	.0360	.0112	.0789
Artificial data	.0738	.1798	.0338	.2874

Note. Model versions: 1. Saturated model in which all parameters are free to vary. 2. All parameters free except κ and T_R, which are held fixed across conditions. 3. Version 2 with β, α, *OLD*, *NEW*, and *B* held fixed across the AN and VM conditions. 4. Version 2 with *s* held fixed across conditions. 5. Version 2 with β held fixed across conditions. 6. Version 2 with $\alpha=0$. Artificial data = fit of Version 2 of the model to an artificial data set with linearly increasing RT functions. In the WSSD fits, response times are measured in seconds and errors are measured in proportions. AN = all-new; VM = varied mapping; CM = consistent mapping; EBRW = exemplar-based random walk; RTs = response times.

Table 2
Best Fitting Parameters From the Core Version (Version 3) of the EBRW Model

Parameter	AN	VM	CM
s	0.056	0.107	0.003
α	1.980		1.596
β	2.069		0.324
B	1.716		1.352
u	2.596	2.723	2.943
v	0.108	0.187	0.009
Old	8.742		4.622
New	7.219		11.361
T_r	211.190		
κ	13.389		

Note. Cells without entries had parameter values constrained to be equal to parameter values from conditions listed to their left. EBRW = exemplar-based random walk; AN = all-new; VM = varied mapping; CM = consistent mapping; *s* = similarity; α = memory-strength asymptote; β = memory-strength decay rate; *B* = background activation; *u* = criterion-activation intercept; *v* = criterion-activation slope; *OLD* = old response threshold; *NEW* = new response threshold; T_r = residual time (ms); κ = time-scale parameter (ms).

corresponding changes in the magnitude of the *v* parameter, which governs how the observer adjusts the criterion *k* with changes in set size: As summed activation grows with increases in set size, the observer compensates by setting stricter values of *k*. Greater compensation is needed when interexemplar similarity is high than when it is low.

The processes that might produce changes in similarity are an important issue, and will be considered in the General Discussion. Here we explain how the value of the similarity parameter affects the predictions. First, because psychological similarity is near zero in the CM condition, summed activation for new probes is near zero, regardless of set size. Thus, the random walk marches in the same efficient fashion toward the *NEW* response threshold regardless of set size, resulting in the nearly flat mean RT function. For old probes, however, lag continues to play a role in the activation function, and memory strength of the old probes decreases with increasing lag. Thus, even in the CM condition, mean RTs for old probes get somewhat slower, on average, with increasing lag. It seems likely that this effect of lag would eventually disappear with continued practice in the CM condition (i.e., if observers form long-term categories corresponding to the positive and negative sets; cf. Shiffrin & Schneider, 1977). Such a process would require an extension of the present version of the model. Finally, the slowdown and increased errors in the VM condition compared to the AN condition arise because of the greater similarity among items in the VM condition. As *s* (and *v*) increase, the random-walk step probabilities for both old and new probes tend toward .5, resulting in a noisier and slower random-walk process.

It is instructive to consider the failings of other constrained versions of the EBRW model (see Table 1). In Version 5, we held fixed the power-function decay parameter at $\beta = 0$; unsurprisingly, this model fitted considerably worse than did the full version, being unable to predict the observed strong effects of lag. In Version 6, we held fixed the memory-strength asymptote at $\alpha = 0$. The increase in WSSD compared to the full version of the model is quite large, showing that a power function descending slowly enough to 0 to capture the performance at long lags would not fit the data at early lags. It is well known that people have excellent recognition memory for the present types of stimuli even at very long lags (e.g., Brady et al., 2008). However, it is less clear whether the lag function asymptotes at a constant value, as assumed in the model we have fit to a limited range of lags, or continues dropping but at a slower and slower rate.

It seems clear that the model can do a good job capturing the observed data. However, the model has a fair number of parameters, and good fit would not be meaningful if any pattern could be fit. This is not generally the case, as we illustrate in Appendix A: When linear RT data are substituted for the curvilinear observed data, the model cannot fit the results.

Hierarchical Bayesian Modeling of Individual-Subject Data

In a second approach to analyzing the data, we implemented the EBRW as a continuous diffusion model (Ratcliff, 1978; Ratcliff, Van Zandt, & McKoon, 1999) and used a software package developed by Wabersich and Vandekerckhove (2014) to compute the predictions from the model (see the supplemental materials for details). The advantage of this approach is that it allows computation of the joint likelihood of responses and their RTs at the level of individual trials. Furthermore, we used the method of hierarchical Bayesian modeling to estimate the posterior distribution of each individual parameter across the three conditions (Plummer, 2011; for general reviews, see Kruschke, 2011; Lee & Wagenmakers, 2014). In this approach, we assume that each subject has his or her own value of each of the model parameters. However, these values are presumed to be sampled from group-level distributions, with the parameters describing each group-level distribution allowed to vary among the three main conditions (AN, VM, CM). An advantage of Bayesian hierarchical modeling is that properties at the group level help constrain estimates of the individual-subject parameters, thereby reducing the noise of the parameter estimates at the individual-subject level. The details of this hierarchical Bayesian analysis are reported in the supplemental materials.

The predictions of the model at the group level were essentially the same as described in the previous section, albeit at the expense of a very large number of individual-subject parameter estimates. (The method assumes a group level distribution and assigns each subject a value from that distribution. Although the total number of parameters is large, we show in the supplement that the model captures not only the mean RTs but also the RT distributions.) The most important new information that is provided by the analysis are the posterior means and 95% highest density intervals (HDIs) for each group-level parameter, which are reported in Table 3. One can determine whether two conditions differ in their parameter values by computing the 95% HDI of their difference. If this interval excludes zero, the parameters are "credibly" different from one another (Kruschke, 2011). These credible differences are also shown in Table 3.

Consistent with the results from our fits of the model to the averaged data, the Bayesian hierarchical analysis suggests that measured similarity (*s*) differs across conditions, with similarity greatest in the VM condition and least in the CM condition. Likewise, the adjustments in the criterion setting with changes in set size (*v*) are greatest in the VM condition and least in the CM condition. The Bayesian analysis of the individual-subject data points to other parameter differences across conditions as well. Perhaps of greatest interest is that the decay in memory strength (β) is greatest in the VM condition and least in the CM condition. We consider potential explanations of this effect in our General Discussion.

General Discussion

The Big Picture

The present work adds significant new support to a line of research showing that a set of common processes can account for results from diverse paradigms aimed at memory storage, category judgments, and memory retrieval. The processes envisioned in our present model build upon those assumed in very successful models of category judgment based on global activation of category exemplars (Medin & Schaffer, 1978; Nosofsky, 1986). Likewise, the model is similar to many models of long-term recognition memory that were also based on global familiarity (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Shiffrin & Steyvers, 1997). Extending those earlier approaches, the present model posits a dynamic

Table 3
Posterior Means and 95% HDI (in Parentheses) for the Group-Level Mean of Each Parameter From the Hierarchical Bayesian Analysis

Parameter	Varied condition		All-new condition		Consistent condition
S	0.304 (0.274–0.333)	>	0.093 (0.046–0.139)	>	0.001 (0–0.002)
α	1.35 (1.19–1.51)	=	0.963 (0.648–1.25)	=	0.989 (0.660–1.29)
β	1.90 (1.42–2.44)	>	0.972 (0.640–1.36)	>	0.343 (0.079–0.657)
B	2.89 (2.56–3.16)	<	4.29 (4.08–4.62)	=	4.51 (4.08–4.89)
u	3.94 (3.58–4.21)	<	5.09 (4.85–5.48)	=	5.59 (5.12–6.10)
v	0.394 (0.348–0.433)	>	0.105 (0.037–0.182)	>	0.002 (0–0.005)
A	54.9 (51.6–58.2)	=	57.5 (53.4–61.1)	=	57.3 (52.9–61.7)
c	0.483 (0.466–0.500)	=	0.490 (0.472–0.507)	<	0.553 (0.532–0.575)
T_r	381 (376–386)	>	336 (331–341)	>	288 (281–294)

Note. The results of pairwise comparisons are indicated between the columns. Two distributions are considered credibly different (greater than or less than one another, as indicated) if the 95% HDI of their difference excludes zero. The diffusion-model implementation involves a different parameterization of the response-threshold parameters (A and c) than does the random-walk implementation; see the supplemental materials for details. HDI = highest density interval; s = similarity; α = memory-strength asymptote; β = memory-strength decay rate; B = background activation; u = criterion-activation intercept; v = criterion-activation slope, T_r = residual time; A = old-response threshold, c = starting-point proportion.

exemplar-retrieval mechanism that results in the emergence of a familiarity-based evidence-accumulation process, allowing it to account for the time course of categorization and recognition decision making and to predict categorization and recognition RTs (Nosofsky et al., 2011; Nosofsky & Palmeri, 1997). A major new contribution is the present demonstration that such models can predict data from recognition paradigms involving both short and long lists by incorporating a form of short-term memory loss. The modeling thus predicts the findings that performance drops sharply with lag and that the lag functions for different list lengths lie atop each other. (Of interest, this result seems reminiscent of a long line of research on free and cued recall; e.g., Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1981. For example, free-recall serial position functions show recency effects that are largely independent of list length. The deeper formal connections between the results in these related memory paradigms remain to be investigated.) The present research goes even further by exploring the effects of varied versus consistent stimulus–response mappings across trials. This manipulation was shown to have dramatic effects upon memory-search performance, in ways analogous to those shown in studies of attention and visual search (e.g., Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Furthermore, the model seems to be a viable candidate for accounting for the effect of these mapping manipulations on memory search. Finally, the present research raises the strong likelihood that the same processes can account for recent studies by Wolfe and colleagues that combined visual search with memory search over a wide range of memory set sizes (Cunningham & Wolfe, 2014; Wolfe, 2012). In sum, the present research provides a remarkably coherent account of results from quite different paradigms, using a model with a common set of underlying processes.

The Present Experiment

The above discussion provided a broad-based characterization of the import of the work, but it is useful to be more specific about the new contributions from the present experiment. We have shown that a representative from the class of exemplar-familiarity models (the EBRW model) accounts successfully for probe recognition involving both short and long lists within the same experimental paradigm and accounts for differences in performance due to manipulations in the relations between targets and foils across trials. Our studies presented lists of lengths ranging from 1 to 16 followed by single-item recognition tests; accuracy and RT were measured. Previous studies that tested both short and long lists showed that mean RTs increased at a decreasing rate as memory set size increased (Burrows & Okada, 1975; Wolfe, 2012), but study time of individual items, memory reinstatement due to repeated testing, and study–test lag varied in an uncontrolled manner across the different memory set sizes. The present studies controlled these factors or allowed them to be measured, and under our conditions of testing subjects produced sufficient errors to provide strong constraints for model evaluation.

Going beyond the Burrows and Okada (1975) and Wolfe (2012) studies, we also manipulated memory requirements by using three main conditions: varied-mapping (VM), all-new (AN), and consistent-mapping (CM). Previous work has provided a general conceptual account of how these differing stimulus–response mappings impact visual and memory search. The present application of the EBRW model provides a quantitative account that captures the results from all three conditions. In addition the EBRW model gives parameter estimates that can be used to interpret the differences between the conditions, an issue that we consider in some depth later in our discussion.

The overall memory set-size effects in our experiments were similar in form to those reported by Burrows and Okada (1975) and Wolfe (2012). However, for old targets the present data showed that the effect of memory set size is primarily due to changes in the lag between study and test (i.e., number of intervening items). After differential lags were taken into account, there was little additional effect of set size per se. The lag effect was therefore primarily responsible for the observed curvilinear dependence of old-item RT on memory set size. Similar results involving lag have been reported previously by Monsell (1978); McElree and Dosher (1989); and Nosofsky et al. (2011) for cases involving only

small memory-set sizes. In the VM and AN conditions, mean correct RTs for new probes were also curvilinear increasing functions of memory set size; whereas in the CM condition, the RT function for new probes was flat. The functions relating error proportions to memory set size tended to be very similar in form to the mean RT functions across all conditions. Finally, mean RTs were slowest in the VM condition, intermediate in the AN condition, and fastest in the CM condition. The error proportions across these conditions mirrored the mean RTs.

The EBRW model accounted well for all of these major trends in performance and provided good quantitative fits to the averaged group data. Its prediction that mean RTs and error probabilities will increase in curvilinear fashion with set size seems to arise because the familiarity-based step probabilities of the random walk are themselves curvilinear functions of lag (in the case of old items) and set size (in the case of new items)—at least for the range of lags and set sizes tested here.

With appropriate choice of parameter settings, the model also accounted extremely well for the patterns of performance across the VM, AN, and CM conditions. Parameter estimates were obtained both by fitting the model to the averaged group data as well as by applying hierarchical Bayesian modeling to the individual-trials data of the individual subjects. Both approaches indicated that a major form of parameter change across the VM, AN, and CM conditions was in the estimate of interitem similarity (s): Similarity was greatest in the VM condition, intermediate in the AN condition, and near-zero in the CM condition. The implication is that in the CM condition, positive versus negative items are highly distinct from each other in memory, whereas they are most confusable in the VM condition.

Interpretation of Parameter Estimates

The near-zero estimate of similarity in the CM condition seems to have a natural interpretation. As discussed previously in our article, performing well in the CM condition requires only that the subject form two long-term categories, one corresponding to the positive set and the second corresponding to the negative set. In principle, the subject can then classify a test probe as old or new without even memorizing the individual memory sets presented on each trial. All stimuli used in our experiment can be easily discriminated on a pairwise basis. Furthermore, subjects also have the opportunity to learn to attend to any category-level features that may useful for separating the two classes of items (e.g., Nosofsky, 1986). Thus, it seems reasonable that highly distinct memory representations can be maintained for members of the two fixed categories, so estimated similarity between items is extremely low.[4]

The finding that similarity is greater in the VM condition than in the AN condition can be explained by task differences and demands: First, in AN the test item likely has some unique features not overlapping with other items either in the present list or prior lists (cf. Mewhort & Johns, 2000). Unique features would reduce similarity. Second, in VM the observer's probe of memory would likely give more emphasis to list context cues in order to access the items on the recent list and not the items on prior lists (e.g., Raaijmakers & Shiffrin, 1981). A likely consequence would be a relative deemphasis of content features associated with individual items. Because the content features determine the similarity of a test probe to the trace of a different item, similarity would be greater in VM than AN conditions. These interpretations would be consistent with a variety of feature-based models such as retrieving effectively from memory (REM; Shiffrin & Steyvers, 1997) and storing and retrieving knowledge and events (SARKAE; Nelson & Shiffrin, 2013). An important direction for future research is to develop and implement precise mechanisms of feature-based similarity change to allow one to predict in greater detail how the history of previous lists impacts performance on current lists.

These interpretations involving the similarity parameter seem plausible, but we had also expected that there would be a difference between VM and AN in the background-noise parameter (B), which was intended to reflect the extent to which items from previous lists are activated by the test probe. Of course, B would not represent the presence of unique features in the test probe. In addition, B also represents the contribution of pre-experimental familiarity, which is invariant across the VM and AN conditions. Still, to gain more sensitive measures of the influence of previous-list activation, an interesting direction for future research would be to explicitly manipulate the recency with which test probes on current lists are presented on previous lists. Assuming that there is residual activation of items from previous lists, the EBRW predicts the general qualitative result that correct rejections will be slower for lures that had been presented on recent lists than for lures presented in the distant past (cf. Monsell, 1978; for a discussion, see Nosofsky et al., 2011, pp. 291–292). However, it is an open question whether the power function that relates memory strength to lag operates in the same manner across different lists as it does within lists.

As noted earlier in our article, Banks and Atkinson (1974) reported a probe-recognition experiment that, like ours, involved a comparison of performance across VM and AN conditions. Although their paradigm was restricted to cases involving only small memory-set sizes (and they did not report their data conditionalized on study–test lag), Banks and Atkinson found, as did we, that mean RTs were slower in their VM condition than in their AN condition. In Appendix B we report a modeling analysis in which the EBRW is fitted to the Banks and Atkinson data. Beyond providing an excellent account of their data, the EBRW modeling analyses reveal interesting effects of their experimental manipulations on the parameters of interest. Most relevant to the present discussion, similarity in the VM condition was again estimated as being higher than in the AN condition. Thus, the similarity effect appears to be robust.

Finally, another effect revealed by the Bayesian hierarchical modeling analysis was that the magnitude of the memory-strength decay parameter (β) was greatest in the VM condition and least in the CM condition. This effect may be closely related to the similarity effect: High-similarity items may be more interfering than low-similarity ones, leading to more rapid declines in memory strength for items on higher similarity lists (e.g., Nairne, 1990; Oberauer & Kliegel, 2006). In addition, an emphasis on list context

[4] A more refined version of the model might allow separate estimates of the extent to which positive-set items are similar to one another versus the extent to which negative-set items are similar to positive-set items. This added complexity, however, was not necessary for modeling the present data.

in VM might produce more rapid decay of individual-item memory strengths than in the AN condition.

Logarithmic RT Functions

In his 2012 article and research Wolfe showed that the relation between memory set size and RT was well fit by a logarithmic function. Although there were various procedural differences between our studies, we too obtained curvilinear RT functions that at least very roughly probably could be described by logarithmic functions. We have not tried to fit any particular descriptive function to our data, preferring to let the curvilinear results flow from the EBRW model that has been used successfully in similar paradigms. It is important to note that the form of an RT function will depend on trade-offs of errors and speed, and we use our model to predict both. In fact the model can explain how the form of the RT functions varies across conditions that stress accuracy versus speed (see Appendix B).

As noted earlier in our discussion, in the case of old probes, the set-size functions appear to be largely derivative of a more fundamental effect of study–test lag. A similar mechanism may contribute to the logarithmic functions observed by Wolfe (2012) in his hybrid memory/visual search paradigm. For example, under Wolfe's conditions, items from shorter memory set sizes will serve more frequently as targets of visual search. Assuming that finding a target on trial N reinstates the memory for that target (boosts its memory strength), search for items from short lists will proceed more efficiently than search for items from long lists.

Error RTs, Effects of Practice, and Model Extensions

An important limitation of the present work is that we made no attempt to account for error RTs. Under the present conditions, which involve the analysis of averaged data, the basis for relations between correct RTs and error RTs is extremely complex. For example, subjects with poor memories may be more likely to produce slow RTs and high error rates; subjects with high response caution would produce slow RTs and low error rates; and subjects with low response caution would produce fast RTs and high error rates. A more fruitful approach to testing the ability of the model to account jointly for correct and error RTs in the present paradigms is to collect extensive data from individual subjects and model performance at the individual-subject level. Extending the model to account for both correct and error old–new recognition RTs of individual subjects has been accomplished in other paradigms by making allowance for variability in response-threshold settings and variation in exemplar-based activations across trials (Nosofsky & Stanton, 2006). We are pursuing this route of modeling individual-subject performance in VM and CM memory search in ongoing work.

Another limitation of the present work is that subjects participated for only a single session. A natural question is how performance patterns may change with more extended practice in the tasks. This question is particularly relevant for the CM condition, which makes allowance for the use of long-term categorization as a basis for performance (Shiffrin & Schneider, 1977). Under the present conditions of testing, the lag with which old items were presented on the individual memory sets continued to exert an impact on CM performance, suggesting that memory for the current list influenced responding. With more extended practice, the lag effects on the old items might eventually disappear. Alternatively, performance might involve a mix between long-term categorization supplemented with recent memory.

The EBRW model was originally formulated as a model of categorization, so extending the model to account for the CM performance of highly practiced subjects should be straightforward. In the categorization version of the model, members of the negative set would be stored in memory along with members of the positive set. Any time a positive-set member is retrieved, the random walk steps in the direction of the OLD response threshold, whereas any time a negative-set item is retrieved, the random walk steps in the direction of the NEW response threshold. This categorization-based decision process might be supplemented by the retrieval of criterion elements as well, so that performance is a mix of "categorization-based" and "recognition-based" responding. It is an open question whether this type of hybrid model could account for the detailed performance of highly practiced subjects in CM versions of the probe-recognition task.

Conclusions

In conclusion, the present results support the hypothesis that common processes of global exemplar-based familiarity may underlie probe-recognition performance involving both short and long lists. Furthermore, the curvilinear relations between mean RT and set size observed in memory-search paradigms that use short and long lists seem to emerge naturally from the predictions of the exemplar-familiarity model. The model also seems to be a promising candidate for understanding the varying patterns of memory search that are observed across conditions in which mapping relations between targets and foils are manipulated across trials. In short, the present modeling has brought together and extended prior research and theory on attention and automaticity, categorization, short- and long-term memory, and evidence-accumulation models of choice RT to move the field closer to achieving a unified account of diverse forms of memory search.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2). New York, NY:Academic Press.

Banks, W. P., & Atkinson, R. C. (1974). Accuracy and speed strategies in scanning active memory. *Memory & Cognition, 2,* 629–636. doi:10.3758/BF03198131

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA, 105,* 14325–14329. doi:10.1073/pnas.0803390105

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Burrows, D., & Okada, R. (1975, June 6). Memory retrieval from long and short lists. *Science, 188,* 1031–1033. doi:10.1126/science.188.4192.1031

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60. doi:10.3758/BF03210740

Cunningham, C. A., & Wolfe, J. M. (2014). The role of object categories in hybrid visual and memory search. *Journal of Experimental Psychology: General*. Advance online publication. doi:10.1037/a0036313

Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science, 23,* 625–634. doi:10.1177/0956797611430961

Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: An investigation using response-time distribution models. *Psychonomic Bulletin & Review, 19,* 363–394. doi:10.3758/s13423-012-0236-8

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. doi:10.1037/0033-295X.91.1.1

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. doi:10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. doi:10.1037/0033-295X.95.4.528

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research, 42,* 2177–2192. doi:10.1016/S0042-6989(02)00118-9

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. New York, NY: Cambridge University Press.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527. doi:10.1037/0033-295X.95.4.492

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 118,* 346–373. doi:10.1037/0096-3445.118.4.346

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238. doi:10.1037/0033-295X.85.3.207

Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: A test of item matching in short-term recognition memory. *Journal of Experimental Psychology: General, 129,* 262–284. doi:10.1037/0096-3445.129.2.262

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10,* 465–501. doi:10.1016/0010-0285(78)90008-7

Murdock, B. B., Jr. (1985). An analysis of the strength–latency relationship. *Memory & Cognition, 13,* 511–521. doi:10.3758/BF03198322

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18,* 251–269. doi:10.3758/BF03213879

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120,* 356–394. doi:10.1037/a0032020

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27. doi:10.1037/0096-1523.17.1.3

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 118,* 280–315. doi:10.1037/a0022494

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300. doi:10.1037/0033-295X.104.2.266

Nosofsky, R. M., & Stanton, R. D. (2006). Speeded old–new recognition of multidimensional perceptual stimuli: Modeling performance at the individual-participant and individual-item levels. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 314–334. doi:10.1037/0096-1523.32.2.314

Oberauer, K., & Kliegel, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language, 55,* 601–626. doi:10.1016/j.jml.2006.08.009

Plummer, M. (2011). *JAGS: Just another Gibbs sampler*. Retrieved from http://mcmc-jags.sourceforge.net/

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134. doi:10.1037/0033-295X.88.2.93

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300. doi:10.1037/0033-295X.106.2.261

Reed, A. V. (1973, August 10). Speed–accuracy trade-off in recognition memory. *Science, 181,* 574–576. doi:10.1126/science.181.4099.574

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66. doi:10.1037/0033-295X.84.1.1

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190. doi:10.1037/0033-295X.84.2.127

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

Sternberg, S. (1966, August 5). High-speed scanning in human memory. *Science, 153,* 652–654. doi:10.1126/science.153.3736.652

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 57,* 421–457.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. New York, NY: Cambridge University Press.

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods, 46,* 15–28. doi:10.3758/s13428-013-0369-3

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition, 2,* 775–780. doi:10.3758/BF03198154

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science, 18,* 133–134. doi:10.1111/j.1467-9280.2007.01862.x

Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychological Science, 23,* 698–703. doi:10.1177/0956797612443968

(*Appendices follow*)

Appendix A

Fits to Artificial Linear Response-Time Data

A question that arises is whether the prediction of a negatively accelerated, curvilinear increase in response times (RTs) is an a priori prediction from the exemplar-based random walk (EBRW) model or whether the model could fit other functions that relate RT to set size. In a preliminary attempt to address this question, we constructed an artificial data set in which the mean RTs for negative probes were a linearly increasing function of set size, and the mean RTs for positive probes were a linearly increasing function of their lag. We used the present observed RT data from the small set sizes and lags to determine the slopes of these functions and then extrapolated linearly to construct the artificial data set. (Because linear extrapolation would produce error rates that exceeded .5 at the larger set sizes and lags, we did not use linear extrapolation to modify the error proportions.) The fit of the highly parameterized Version 2 of the EBRW model to this artificial data set is reported in Table 1. As can be seen, the weighted sum-of-squared deviations is extremely large. This result suggests that the most natural prediction from the EBRW model is that the mean RTs will increase curvilinearly with increases in set size and lag, as was observed in our data and in the related paradigms of Wolfe (2012) and Burrows and Okada (1975). Although parameters are available that allow the model to fit linear-RT functions in isolation, those parameter settings then yield poor predictions of other aspects of the complete set of data.

Appendix B

Application of the EBRW Model to the Probe-Recognition Data of Banks and Atkinson (1974)

In Banks and Atkinson's (1974) experiment, subjects engaged in short-term memory-search tasks involving lists of words. In their *familiar* condition, which corresponds to our varied-mapping condition, memory sets were drawn from a small, well-learned set of words, and targets and foils were chosen randomly from the set on each trial. In their *infinite* condition, which corresponds to our all-new condition, memory sets were sampled without replacement from a very large pool of words on each trial. In addition, Banks and Atkinson tested subjects in payoff conditions that emphasized either *accuracy* or *speed*. Thus, there were four main conditions: familiar–accuracy, familiar–speed, infinite–accuracy, and infinite–speed. In all conditions, memory set size on each individual trial was 2, 3, 4, 5, or 6. Of the trials, half tested positive probes and half tested negative probes. The lag of positive probe items was randomly chosen on each trial, but the effect of lag was not analyzed in their article. Banks and Atkinson described the presentation rates of study items as occurring at normal reading speed, whereas the retention interval between study and test probe was subject controlled.

The observed mean RTs and error probabilities are plotted as a function of conditions in Figures B1 and B2 (symbols). (We estimated these data by eye from the figures plotted in the original article. In the accuracy conditions, Banks and Atkinson reported the error rates only averaged across set sizes and stated that there was little effect of set size. Thus, for simplicity, we assume the observed error data are flat functions of set size in the accuracy conditions.) As can be seen, within this range of small set sizes, mean RTs for both old and new probes were roughly linear functions of set size in all conditions. However, the slopes of the set-size functions were clearly less in the speed conditions than in the accuracy conditions. In the speed conditions, error rates were increasing functions of the size of the memory set (see Figure B2). False alarms (responding "old" to "new" items) predominated in the familiar condition, whereas misses (responding "new" to "old" items) predominated in the infinite condition. (Error rates were low in the accuracy condition, and only the rates averaged across set sizes were reported.).

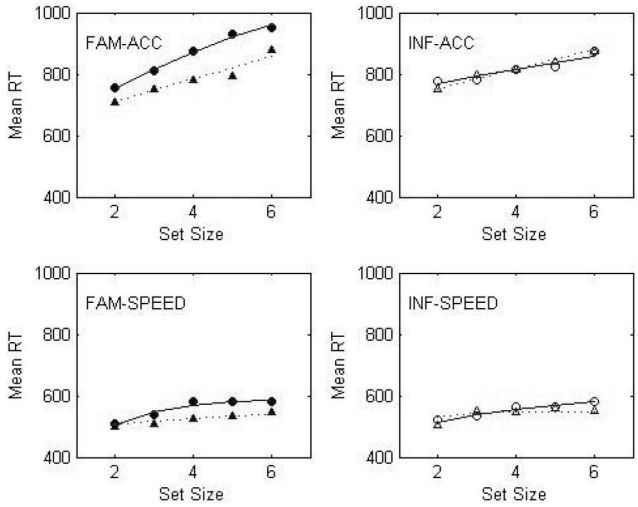

Figure B1. Mean response time (RT) as a function of conditions in the Banks and Atkinson (1974) experiment. Triangles = old items, circles = new items. Dotted lines = old-item predictions from EBRW model. Solid lines = new-item predictions from EBRW model. EBRW = exemplar-based random walk; FAM = familiar; INF = infinite; ACC = accuracy.

(*Appendices continue*)

We fitted the EBRW to these group-averaged data using the methods described in the main text. However, because Banks and Atkinson did not report lag functions, we could fit the model to the set-size functions only. The best fitting parameters from a moderately constrained version of the model are reported in Table B1, with the model's predictions plotted as different line types in Figures B1 and B2.

As is clear from inspection of Figures B1 and B2, the model provides an excellent fit to the data and accounts for all of the major performance trends. Furthermore, the parameter estimates are generally easily interpretable. For example, the magnitudes of the response thresholds are greater in the accuracy conditions than in the speed conditions. (In random-walk and diffusion models, adjustment of the response thresholds is the most direct approach to meeting the demands of speed stress versus accuracy stress. Increasing the magnitude of the thresholds leads to increased accumulation of evidence and so greater accuracy but at the expense of slower responding.) In addition, the mean residual time was faster in the speed than in the accuracy conditions, which seems a reasonable result as well. Because lag functions were not reported, we cannot reliably estimate the β memory-decay parameter. Similar fits are obtained across a wide range of values of β. An interesting result is that measured similarity was greater in the speed than in the accuracy conditions. Under speed stress, subjects may not have sufficient time to carefully compare all of the features of the probe to the members of the memory set, resulting in greater confusability. Finally, replicating the model-based result from our own experiment, in the conditions closest to our own, estimated similarity in the familiar–accuracy (VM) condition was greater than in the infinite–accuracy (AN) condition.

One parameter change for which we do not have a ready explanation involves the magnitude of *v* in the infinite–speed condition. The large magnitude of *v* implies that subjects set an increasingly strict criterion for responding "old" as set size increases in that condition. The result is that there is a high probability of "misses" in the infinite–speed condition.

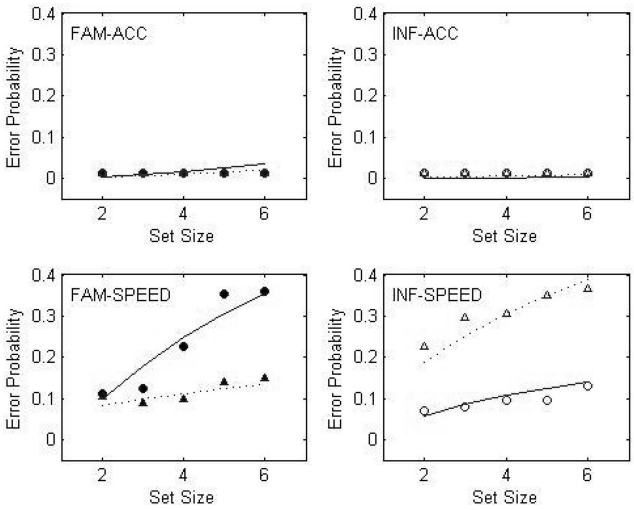

Figure B2. Error probabilities as a function of conditions in the Banks and Atkinson (1974) experiment. Triangles = old items, circles = new items. Dotted lines = old-item predictions from EBRW model. Solid lines = new-item predictions from EBRW model. EBRW = exemplar-based random walk; FAM = familiar; INF = infinite; ACC = accuracy.

Table B1
Best Fitting Parameters From the EBRW Model Applied to the Banks and Atkinson (1974) Data

Parameter	Fam-Acc	Fam-Speed	Inf-Acc	Inf-Speed
s	.040	.102	.016	.103
α	.183			
β	.001			
B	.001			
u	.266	.282	.406	.452
v	.056	.075	.034	.193
Old	3.713	2.947	3.188	2.041
New	5.631	2.610	7.799	3.804
T_r	285.8	195.4		
κ	53.809			

Note. Cells without entries had parameter values constrained to be equal to parameter values from other conditions. EBRW = exemplar-based random walk; Fam = familiar; Inf = infinite; Acc = accuracy; *s* = similarity; α memory-strength asymptote; β = memory-strength decay rate; *B* = background activation; *u* = criterion-activation intercept; *v* = criterion-activation slope; *Old* = old response threshold; *New* = new response threshold; T_r = residual time (ms); κ = time-scale parameter (ms).

Received November 14, 2013
Revision received February 11, 2014
Accepted March 8, 2014 ■