Modeling short- and long-term memory contributions to recent event recognition

Robert M. Nosofsky[1], Rui Cao[2], Samuel M. Harding[1] and Richard M. Shiffrin[1]

1. Indiana University Bloomington
2. Boston University

Robert Nosofsky

Psychological and Brain Sciences

1101 E. Tenth Street

Indiana University

Bloomington, IN 47405

nosofsky@indiana.edu

Abstract

Participants gave recognition judgments for short lists of pictures of everyday objects. Pictures in a given list were an equal mixture of three types that varied according to the way they were used as targets and foils earlier in the same session. Under consistent-mapping (CM), targets and foils never switch roles; under varied-mapping (VM), targets and foils switch roles randomly across trials; whereas all-new (AN) items are novel on each trial of the experiment. Past research has shown that markedly enhanced performance occurs in CM conditions, leading to conclusions that item-response learning takes place in CM, perhaps automatically. However, almost all past research has compared CM, VM and AN performance in between-blocks designs in which participants may adopt different cognitive strategies across the conditions. The present mixed-list design holds constant the strategy that is used for CM, VM, and AN items, and produced patterns of performance dramatically different than those observed in pure-list control conditions. We develop an extended version of an exemplar-based random-walk model of probe recognition to account for the major qualitative effects in the data. The data and the model provide evidence for strong item-response learning for CM foils but weak item-response learning for CM targets. We consider possible explanations for these effects in our General Discussion.

Key Words: short-term memory, long-term memory, probe recognition, automaticity, mathematical modeling

Theoretical accounts of memory retrieval assume that performance in most tasks involves joint contributions from both short- and long-term memory (e.g. Atkinson & Shiffrin, 1968). This article explores these joint contributions using short-term probe-recognition memory, with both accuracy and response time (RT) measures of performance. On each trial, observers are presented with a short list of items (the "memory set") followed by a test probe. The task is to judge, as rapidly as possible while minimizing errors, whether or not the test probe was a member of the memory set. Test probes that are members of the memory set are termed "old" probes or "targets"; probes that are not members of the memory set are termed "new" probes or "foils". Short-term probe-recognition has been used extensively to study interactions between short-term memory (STM) and long-term memory (LTM) and the mechanisms that allow for the development of certain forms of automaticity (e.g. Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Our present research and modeling is aimed at further elucidating the processes at work, including the degree to which learning occurs automatically or is dependent on strategic choices.

In typical probe-recognition memory-search tasks, it is generally observed that RTs get longer and accuracy decreases as the size of the memory set grows, a pattern termed the *set-size effect*. The set-size effect was first reported in the classic studies reported by Sternberg (1966), and has been replicated in innumerable subsequent studies that investigated the underlying processes of memory-search tasks (e.g., McElree & Dosher, 1989; Monsell, 1978; Nosofsky, Little, Donkin, & Fific, 2011). Although the detailed processes that mediate short-term probe recognition might differ depending on the specific experimental conditions (for a comprehensive review and analysis, see Sternberg, 2016), essentially all theories assume that observers engage

the current memory-set items in STM to perform the task. Larger memory sets result in worse performance due to the capacity-limited nature of STM (Atkinson & Shiffrin, 1968).

In their now classic studies, Schneider and Shiffrin (1977; Shiffrin & Schneider, 1977) conducted hybrid memory/visual search tasks in which observers studied from one to four items and then searched for the presence of a studied item in a series of rapidly presented visual displays that had from one to four items per display. These researchers discovered that the set-size effect was greatly reduced or eliminated under "consistent mapping" (CM) conditions. In CM, the memory sets and target probes are chosen from a fixed set on every trial, and the foil probes are chosen from a different fixed set on every trial. Thus, the targets and the foils never switch roles across trials. As practice proceeded in Schneider and Shiffrin's studies, performance improved dramatically: participants were able to scan visual displays and make their old/new judgments with decreasing RT and few errors. Most importantly, the performance became largely invariant to set-size manipulations, suggesting reliance on a process other than the retrieval of the list held in STM (see also Logan & Stadler, 1991). In Schneider and Shiffrin's (1977) studies, participants were also tested in varied-mapping (VM) conditions, in which the items that served as old probes on some trials were new probes on other trials, and vice versa (a procedure typical of most STM probe-recognition studies). In contrast to CM, performance in the VM condition improved very little with practice, and the set-size effect persisted even after extensive practice. The researchers proposed that performance in the VM condition required an effortful, controlled process, regardless of the amount of practice; whereas practice in the CM condition allowed for the development of an extremely efficient form of information processing. They also showed that certain forms of automaticity developed, seen in visual search as an automatic attraction of attention to a target anywhere in a multiple item display, bypassing the

need for serial search of the display. There were also indications of automatic mechanisms at work in memory search, in both learning and task processing.

The dramatic contrast between CM and VM performance reported by Schneider and Shiffrin (1977) in their hybrid memory/visual multiple frame search tasks was shown after extensive practice, but the speed of learning was not reported. In paradigms involving pure memory-search tasks, ones involving only a single test probe on each trial rather than multiple displays, the dramatic contrast is typically observed after only minimal amounts of practice (e.g., Cao, Nosofsky, & Shiffrin, 2017; Cao, Shiffrin, & Nosofsky, 2018; Nosofsky, Cox, Cao, & Shiffrin, 2014). Our current investigations will focus on these early stages of performance in pure probe-recognition, memory-search tasks.

Despite the numerous studies aiming to explore the nature of the memory representations and cognitive processes that underlie CM versus VM performance, the bases for the dramatic performance differences remain to be clearly delineated (i.e. Cao et al., 2017; Cheng, 1985; LaBerge & Samuels, 1974; Logan & Stadler, 1991; Schneider & Fisk, 1982; Shiffrin and Schneider, 1977; for review, see Schneider & Chein, 2003). There are a number of fundamental issues to consider when accounting for these differences; as will be seen, teasing apart the controlling factors provides a major challenge. One factor that could account for much of the difference between CM and VM performance is that of learning responses: Participants could rely on information from LTM (the learned response to the test items) in performing CM tasks, but are forced to focus mainly on the current memory set to perform VM tasks. Such CM learning could be fostered by the fact that a present target would always have been a target on prior trials, and a present foil would always have been a foil on prior trials. As practice proceeds, the observer could in principle ignore the current memory set and rely solely on learning, i.e. on

information from LTM, to perform the task. By contrast, in typical VM tasks, a given test item on a current trial will have served randomly as an old target versus a new foil on numerous previous trials of the experiment. Thus, for VM, it behooves the observer to rely primarily on STM for the current memory set in order to perform the task.

Although observers can make reference to LTM under CM conditions, a challenging question concerns the detailed nature of the LTM information that is used. One contrast is between what we have referred to in previous work as *item-familiarity* models versus *item-response-learning* models (Cao et al., 2018; Nosofsky, Cao et al., 2014). In an item-familiarity model, the evidence that a test probe is "old" versus "new" is presumed to be based solely on the extent to which it activates items in the current memory set as well as items presented on previous trials of the experiment. If an item has occurred frequently in previous trials of the experiment, then its LTM familiarity will be high. By contrast, in item-response-learning models, the observer is assumed to store item-response *pairs* in memory. On trials in which an item has served as an old test probe, a memory trace is formed in which an "old" response is attached to the item; but on trials in which an item has served as a new test probe, a memory trace is formed with a "new" response attached to the item. Retrieval of memory traces with old response labels provides evidence that the test probe is old, but retrieval of traces with new response labels provides evidence that the test probe is new. As we review below, in past work, we have obtained evidence supporting the idea of item-response-learning under CM memory-search conditions (Cao et al., 2017, 2018; Nosofsky, Cao et al., 2014); however, the generality of that finding remains to be investigated and will be a central focus of the present research.

In addition to the differential reliance on LTM versus STM, another major reason for the dramatic difference between CM and VM performance likely involves strategic adjustments in

different types of criterion settings across the tasks (Strayer & Kramer, 1994a,b). Many modern formal models, including one that we will advance and test in this article, conceptualize memory-based decision making as involving an evidence-accumulation process (Ratcliff, 1978). In such models, evidence about whether a test item is old or new is extracted in step-by-step fashion, and a final decision is made once the accumulated evidence for one response versus the other reaches a criterion. The idea is illustrated schematically in the top panel of Figure 1 in terms of a simple random-walk process (Busemeyer, 1982; Feller, 1968). There is a counter with a starting point of zero, and the observer establishes an "OLD" response criterion and a "NEW" response criterion. On each step of the process, evidence is obtained that the test probe is either "old" or "new". If the evidence points to "old", then the random walk steps toward the OLD criterion, otherwise the random walk steps toward the NEW criterion. The evidence-accumulation process continues until one of the response criteria is reached, at which point the observer emits the corresponding response. The RT for making a decision is determined by the number of steps required to reach one or the other criterion. In cases in which the evidence is strong that the item, say, is "old", then most steps will go in a single direction toward the OLD criterion. Thus, there would be a high probability of responding "old" and RTs for making that decision would be short. By contrast, in cases in which the evidence for "old" versus "new" is equivocal, the random walk would tend to meander back and forth. Thus, RTs would tend to be long, and response probabilities associated with making "old" versus "new" responses would hover closer to chance.

In terms of the random-walk framework, one reason why RTs associated with CM versus VM performance might vary dramatically is that observers are likely to establish different settings of the response criteria (OLD and NEW in Figure 1) across the tasks. For example, we

will provide multiple reasons to expect that the evidence-accumulation process operates more efficiently under CM-training conditions than under VM-training conditions. Because the steps in the random walk are far more consistent in CM, the observer can place the response criteria much closer to the starting point without any appreciable loss in accuracy, resulting in even shorter RTs. This provides one illustration that differences in RT in CM versus VM may not result solely from differential reliance on LTM versus STM, but on changes in random-walk criterion placement as well.

Depending on the model, say one based on familiarity or one based on item-response knowledge, there is another type of criterion that could differ for CM and VM, a criterion determining the way that familiarity or item-response knowledge influences the drift rate of the evidence-accumulation process itself (Ratcliff, 1985). This can be illustrated with a simple item-familiarity model. Hypothetical distributions of "familiarity" for old and new test probes in CM versus VM conditions are illustrated schematically in Figure 2. Overall familiarity for old test probes would tend to be high for both CM and VM because these items have just appeared in the current memory set and have been seen often both in previous memory sets and tests. Familiarity would tend to be lower to some degree for new test probes, for both CM and VM, because they did not appear in the current memory set. However, as illustrated schematically in Figure 2, the familiarity of the new test probes would tend to be far greater in VM than in CM because the new test probes have appeared often in previous memory sets in VM, but have never appeared in previous memory sets in CM. It seems clear from the figure that the criterion best separating old and new would differ for CM and VM. Thus, even if there is little or no difference in LTM familiarity for old items across the CM versus VM conditions, the changed criterion setting would still impart a major benefit to the old items in the CM condition.

In a nutshell, it becomes difficult to determine the extent to which the dramatic CM advantage is the result of contributions from enhanced LTM representations, changed criterion settings, or some combination of both factors, particularly when CM and VM tasks are conducted in between-subjects or between-blocks designs, as is typical. Such paradigms allow different strategies, particularly different criteria, to be chosen for CM and VM.

In the present research we therefore investigate CM versus VM probe recognition by using a design with *mixed study lists*. Although mixing CM and VM trials within each block could reduce the use of differential strategies, it would still be possible for participants to change their strategy once seeing which items are presented on each trial. In contrast, if CM and VM items are present in the same study lists, such strategy change should be minimized. Furthermore, as explained in more detail below, to achieve still greater diagnosticity in our paradigm, each study list also includes "all-new" (AN) items never experienced in previous memory sets (see also Banks & Atkinson,1974; Nosofsky et al., 2014a,b). Thus, on each trial, a memory set composed of an equal mixture of CM, VM and AN items is presented to the observer. This is followed by a test probe equally likely to be one of the three types. The observer makes an old-versus-new judgment as to whether the test probe appeared in the current memory set. We suggest that the observer in this design will adopt the same criterion settings across the CM, VM and AN conditions. It then becomes informative to see whether there will still remain dramatic differences between CM and VM performance. We note here that Strayer and Kramer (1994a) conducted a related study aimed at similar issues, although they did not contrast pure CM versus VM memory-search paradigms. We consider the relation between our results and the earlier ones reported by Strayer and Kramer (1994a) in our General Discussion. To preview, our results will converge strongly with those reported by Strayer and Kramer

(1994a) by demonstrating that shifts of the OLD and NEW random-walk response criteria (see Figure 1) play a major role in influencing performance across the conditions.

A second related advantage of the mixed-study-list design is that it allowed us to explore the extent to which item-response learning under CM conditions may operate "automatically". Results from various previous research studies suggest that simple item-familiarity models fail to explain performance under pure CM-training conditions. For example, in one type of manipulation, a test item that is a foil is repeated across two consecutive trials. Under VM conditions, this manipulation leads to massive interference (increased false-alarm rates or longer correct-rejection RTs; Monsell, 1978; Nosofsky et al., 2014a); such interference is predicted by item-familiarity models because the repeated foil is highly familiar, providing misleading evidence that the foil is old rather than new. However, under CM conditions, if anything there is slight facilitation when a test foil is repeated across consecutive trials (Nosofsky et al., 2014a; for related findings, see Cao et al., 2018; Wolfe, Boettcher, et al., 2015). The facilitation under CM conditions is consistent with the idea that an item-response mapping between the foil and the new response is strengthened across the repeated trials. A question that arises, however, is the extent to which the item-response-learning process is a strategy-dependent one. Under pure CM conditions, the observer may learn that memorizing responses that are associated with individual items is an effective strategy for performing the task. Under mixed-study-list conditions, however, such a strategy will not generally be effective, because inconsistent old versus new responses have been assigned to VM items on the previous trials. If item-response learning operates automatically, independent of task-induced strategy, then we should still see evidence of the process for the CM-trained items. That evidence should disappear, however, if the process is a strategy-dependent one that operates only under pure CM-training conditions.

Although the main purpose of the present experiment was to investigate CM, VM and AN performance under the mixed-list conditions, as a source of comparison we also tested pure CM, VM and AN designs. As will be seen, we held fixed across the mixed and pure conditions a variety of experimental factors such as memory set sizes, individual-item frequencies, and frequencies with which individual items were assigned old versus new responses. These controls allowed us to compare in a meaningful way how CM, VM and AN performance varies across the mixed-list and pure-list designs and to thereby gain insights about the cognitive processes that govern performance. To preview just one example, to the extent that changed criterion settings play a major role in leading to enhanced CM performance (compared to VM performance) in pure-list designs, we should see greatly reduced differences in performance across these conditions in the mixed-list design (see also Strayer & Kramer, 1994a).

Finally, beyond making these kinds of qualitative empirical comparisons, our central goal involved the application and testing of a formal quantitative model for characterizing CM, VM and AN performance in both the mixed-list and pure-list conditions. The formal model is an extended version of the *exemplar-based random-wa*lk (EBRW) model that has been successfully applied to various forms of categorization (Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005) and old–new recognition memory (Donkin & Nosofsky, 2012; Nosofsky, 2016; Nosofsky, et al., 2011). In versions of the model applied to VM, CM and AN probe-recognition memory search (Nosofsky et al., 2014a,b; Cao et al., 2018), each item of the memory set is stored as an exemplar in short-term memory. Items and test probes from previous memory sets may also be stored in long-term memory. When the current test probe is presented, it activates exemplars to which it is similar (both short-term and long-term), and the activated exemplars race to be retrieved (see Formal Models section for details). The retrieved exemplars are evaluated and, as

was illustrated in Figure 1, lead an evidence accumulator to move toward either the OLD response criterion or the NEW response criterion. As described in detail in the Formal Models section, different versions of the model are based on different assumptions about the nature of the exemplar-retrieval process. The parameter estimates from the model help illuminate the extent to which item-response learning is taking place, whether there are differences in the nature of CM versus VM memory representations, and the presence or absence of criterion adjustments across the different tasks.

Experiment

The experiment was carried out twice. The results from the first iteration were surprising enough that an independent replication (with some additional conditions) was carried out after a long delay, with new programming of experiment and analysis and new participants. As described below, the results of the conditions in common were essentially identical, leading us to believe the findings are highly reliable.

The participants engaged in probe-recognition memory-search tasks that involved CM, VM or AN items. In the mixed-study-list condition (called mixed hereafter), the three item types were mixed in equal numbers in each study list; the test item was equally likely to be each of these types. In the pure conditions, separate participants were given CM, VM, or AN tasks. Across the mixed and pure conditions, individual items of each type occurred with the same frequency. Within both the mixed and pure conditions, we manipulated across trials whether the test probes were targets or foils, the size of the memory sets, and, in the case of target-probe trials, the serial position in which the target appeared in the memory set. In both the mixed and pure conditions, memory-set items were presented at relatively rapid rates and with only a brief

retention interval between presentation of the memory set and presentation of the test probe. Such conditions have been deemed to discourage idiosyncratic rehearsal strategies, so that the psychological recency of items on the study lists corresponds closely to their experimentally-manipulated recency (for extensive discussion, see, e.g., Donkin & Nosofsky, 2012a,b; Nosofsky & Donkin, 2016).

As mentioned above, we tested the mixed condition in two separate cohorts of participants recruited across separate academic years, with independent programming of experimentation and analysis. In addition, the 'matched' pure conditions were added when the replication was carried out. The patterns of aggregate data for the mixed condition across the first and second cohorts were extremely similar (see Figure S1 of supplementary materials) leading us to combine the first and second cohorts of participants in presenting the findings.

Method

Participants

Participants were members of the Indiana University community who either received credit towards an introductory psychology course requirement or were paid $12 for participating. There were 100 and 94 participants in cohorts 1 and 2 of the mixed condition, respectively. There were 35, 36 and 34 participants in the pure CM, VM and AN conditions, respectively. The same proportion (roughly half) of participants received course credit or were paid in cohort-2 of the mixed condition and in each of the pure conditions. All participants from cohort-1 of the mixed condition received course credit. All participants had normal or corrected-to-normal vision and all reported having normal color vision.

Stimuli and Apparatus

The stimuli were drawn from a pool of 2,400 unique everyday object images from the website of Talia Konkle and described in Brady, Konkle, Alvarez, and Olivia (2008). Each image subtended a visual angle of approximately 7 degrees and was displayed in the center of the computer screen on a gray background. The experiment was conducted with MATLAB Psychophysics Toolbox (Brainard, 1997) on personal computers. All participants were tested individually in private, sound-attenuated booths.

Procedure

For all conditions: Memory set size on each trial was 3, 6 or 9, chosen randomly on each trial. The order of presentation of the individual memory-set items was chosen randomly on each trial. Test-probe status (target or foil) was chosen randomly (50%) on each trial. On target trials, the test probe was randomly selected from among the items in the memory set. CM foils were randomly selected from the CM-foil set; VM foils were randomly selected from those members of the VM-set that were not memory-set items on that trial; and AN foils were randomly selected from the remaining items in the 2400-image set.

Mixed Condition. For each participant: In the mixed condition, 6 stimuli were randomly sampled from the 2400-image set to be the VM-set and another 6 stimuli were randomly sampled to be the CM-set (3 images for the CM-target set, 3 for the CM-foil set). For each trial, one-third of the memory set items were sampled randomly from the VM-set, one-third from CM-target set, and one-third of the memory set items were AN items that were never presented on previous trials. On each trial, the AN items were randomly selected from the remaining items in the 2400-image set. Foil test-probe trials were equally likely to be CM, VM or AN foils. Note, therefore,

that a CM-target-set item could only serve as a target; a CM-foil-set item could only serve as a foil; and VM-set items switched roles randomly from trial to trial. Because AN items were never presented on previous trials, they can be viewed as "unmapped" items. Because the memory set always consisted of a mixture of CM, VM and AN items in random order, participants could not anticipate whether a test probe would be CM, VM or AN.

Each participant completed a single session of testing that lasted about 45 minutes. The session consisted of 7 blocks with 25 trials for each block. Participants were instructed to memorize the memory-set items on each trial and indicate if the test probe was a member of the memory set (an old item, or target) or not a member of the memory set (a new item, or foil) by pressing a key on the computer keyboard (J=old, F=new). Participants were not informed of the CM-VM-AN manipulation before testing; however, instructions were provided that test probes that had appeared in previous trials of the experiment but not in the current memory set were defined as "new". On each trial, a fixation point ("*") appeared on the center of the screen for 0.5 seconds to indicate the start of that trial. Then each of the memory-set items was presented in the center of the screen for 1 sec with a 0.1 sec blank-screen inter-stimulus-interval. A blank screen then appeared for 1 sec, followed by another fixation point ("+") for 0.5 sec, followed by the test probe. The test probe stayed on screen until a key response was registered, after which there was another .5 sec blank screen and then feedback provided for 1 sec to indicate whether or not the response was correct. Participants were instructed to rest their index fingers on the response keys throughout the experiment and to respond as quickly as possible without making errors.

Pure Conditions. The procedure for the pure conditions was the same as for the mixed condition, except for the manner in which items were sampled to construct the memory sets. For

each participant: The VM-set was constructed by randomly sampling 18 items from the 2400-image set; the CM-target set by randomly sampling 9 items; and the CM-foil set by randomly sampling 9 items. In the pure-VM condition, the memory sets were constructed by randomly sampling from the VM-set; and in the pure-CM condition, the memory sets were constructed by randomly sampling items from the CM-target set. In the pure-AN condition, the memory sets were constructed by randomly sampling from all images in the 2400-image set that had not yet been used in the experiment. Target test probes and foil test probes were selected in the same manner as already described for the mixed condition.

These sampling procedures ensured that frequencies of individual item presentations representing the CM, VM, and AN conditions were equated across the mixed and pure conditions. In Table 1 we report these frequencies for the VM, CM-target, and CM-foil items. (By definition, each sampled AN item serves once in a memory set in both the mixed and pure conditions, and it may or may not then serve as a test probe.) Note that it is logically impossible to equate all potentially relevant CM and VM item-role frequencies in a single design. In the present experiment, we designed the sets of items such that the total frequency with which individual CM and VM items served as test probes was equated.

Results

We conducted the following pre-processing steps prior to full analysis of the data. First, we considered the first block to be a practice block and did not include it in our analyses. Second, we eliminated trials with RTs greater than 4000 ms or less than 180 ms (~1% trials). Third, we computed overall proportion correct and mean correct RT across all trials for the individual participants, and we eliminated participants who were severe outliers with respect to these measures in any of the conditions (9 of 100 participants in cohort 1 of the mixed condition,

4 of 94 participants in cohort 2 of the mixed condition and 8 of 105 participants in the pure conditions).[1] Finally, in a more fine-grained analysis, we examined set-size functions at the individual-participant level. In the pure-CM and pure-AN conditions, a few outlier participants had very long RTs or high error rates at the shortest set size, despite the ease of these conditions. Almost certainly, the poor performance at the shortest set size was due to attentional lapses and related factors that go outside the scope of the present investigation. We eliminated from analysis these participants as well (4 participants in the CM condition and 3 participants in the AN condition). We replaced the eliminated participants to achieve pre-planned sample sizes of 90 in cohort-2 of the mixed condition and 30 in each of the pure conditions.

The main results of the experiment are displayed in Figures 3-6. In each figure, the left panels display results from the pure conditions, and the right panels display corresponding results from the mixed condition. Figure 3 plots mean correct RTs as a function of conditions (VM, AN, CM), set size, and probe type (old vs. new); and Figure 4 plots proportion of errors as a function of these variables. Figures 5 and 6 provide a more detailed breakdown of the results for the old test-probe data by plotting mean correct RTs and proportion of errors as a joint function set size and *lag*, where 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 6, the item in the sixth serial position has lag 1, the item in the fifth serial position has lag 2, and so forth.

We present the results by describing the patterns that are evident in the figures. Results from extensive and detailed statistical tests to support these descriptions are provided in the appendix.

Pure Conditions. The results from the pure conditions broadly replicate patterns reported previously by Nosofsky, Cox et al. (2014b) for these conditions (although these researchers had

used different memory-set sizes and individual-item frequencies than in the present experiment). As shown in Figure 3, for both old and new test probes, the mean correct RTs are shortest in the CM condition, intermediate in the AN condition, and longest in the VM condition. In both the VM and AN conditions, RTs get longer as set size increases, whereas the set-size functions in the CM condition are flat. These same patterns are observed for the mean proportions of errors (Figure 4). Inspection of the joint set-size by lag functions in the left panels of Figures 5 and 6 reveals that, for the old test probes, the main controlling variable is lag, not set size: In the VM and AN conditions, RTs get systematically longer and error rates increase as lag increases. Once one conditions on lag, there is little if any remaining effect of set size. The main reason why one sees increasing set-size functions in Figures 3 and 4 for the old test probes is that larger-size memory sets contain targets with longer lags. In contrast to the VM and AN conditions, in the CM condition the lag functions are virtually flat.

Mixed Conditions. In certain respects the results from the mixed condition are similar to those in the pure condition, whereas in other respects the results are dramatically different. We start by noting some of the similarities. First, averaged across the old and new items, mean correct RTs are still shortest in the CM condition, intermediate in the AN condition, and longest in the VM condition. For the error probabilities, there is little difference between the CM and AN items, but error probabilities are still greatest for the VM items. The VM items continue to show longer RTs and increased error probabilities with increases in set size; the AN target items also show these overall set-size effects. The lag x set-size functions for the VM and AN target items continue to show that mean RTs get longer and error probabilities increase as lag increases; and that once one conditions on lag, any additional effects of set size are rather small.

Of greater interest are the differences in the results across the pure and mixed conditions. One difference is that in the pure case, the old items showed big differences in overall RTs and error proportions across the CM, VM and AN conditions; but in the mixed case, the overall RTs and error proportions for the old items are similar in magnitude across the CM, VM, and AN conditions. Notably, in the mixed condition, the overall error proportions and RTs for the CM-old items are now nearly the same as for the VM-old items. Furthermore, whereas the lag functions for the CM-old items were nearly flat in the pure case, performance on the CM-old items now gets systematically worse with increases in lag, regardless of whether performance is measured in terms of errors or RTs. In a nutshell, there is a qualitative shift across the pure and mixed conditions in which performance on the CM targets gets dramatically worse in the mixed condition.

The overall RTs for the new items across the CM, VM, and AN conditions are also squeezed closer together in the mixed case compared to the pure case. Another interesting difference is that error proportions for the AN-new items are *reduced* in the mixed condition compared to the pure condition. At least on the surface, this type of result seems extremely surprising: Presumably, in the pure conditions, observers will have adopted a strategy that is more nearly optimal with respect to the types of items on which they are being tested. By contrast, in the mixed conditions, observers need to adopt a strategy that deals simultaneously with very different item types (CM, VM and AN). Yet error proportions on the AN-new items are significantly reduced in the mixed case compared to the pure case. Another notable feature of the mixed-conditions data is that error probabilities for the VM-new items sky-rocket relative to the other item types.

Interim Discussion

As observed in numerous previous studies, in the pure conditions, performance on the CM-old items is dramatically better than on the VM-old items, regardless of whether performance is measured in terms of RTs or error rates. Yet in the mixed conditions, RTs and error proportions associated with the CM-old items are nearly the same as for the VM-old items.

On the one hand, it is reasonable to expect reduced differences in CM-old versus VM-old performance across the pure and the mixed conditions due to adjustments in criterion settings. For example, as we noted in the introduction, the evidence distributions associated with old and new items are far more discriminable for CM than for VM; thus, in the pure case, which allows for separate setting of criteria for each condition, observers can set the drift-rate criterion in a far more effective location for CM than for VM (see Figure 2). But the mixed case requires that the criterion be set in a single location, so this dramatic advantage for CM-old disappears.

Nevertheless, even if a criterion change causes the performance differences to be reduced, one might still expect to see a substantial advantage in the mixed condition for the CM-old items compared to the VM-old ones. As reviewed in our introduction, past work has provided evidence that, under pure conditions, CM performance benefits from learning of consistent item-response mappings. If this type of item-response learning operates automatically, then a substantial advantage for CM-old compared to VM-old should still be observed in the mixed condition, because the CM targets are consistently mapped to old responses, whereas the VM items are randomly mapped to old versus new responses. Instead, the present empirical results provide initial evidence that this form of item-response learning may not proceed automatically, or is at least greatly weakened in the mixed-lists design.

Complicating the story, however, are the results for the CM-new items in the mixed condition. On the one hand, even if item-response learning does not operate, one still expects a substantial advantage for CM-new compared to VM-new in the mixed condition. The reason is that the CM-new items have low overall familiarity (they never appear in the memory sets, and are experienced only on the small proportion of trials in which they serve as foils). By contrast, the VM-new items have very high familiarity, so it is difficult for observers to correctly reject them. The more interesting result, however, is that the CM-new items are correctly rejected with significantly shorter RTs than are even the AN-new items (with the false-alarm rates for both item types being essentially at floor). Regardless of how low in overall familiarity they may be, the CM-new items are presumably at least as familiar as the AN-new items, given that the AN-new items have never been experienced prior to the current trial. In accord with previous work, these results are consistent with the idea that item-response learning *does* take place for the CM-new items, even in the mixed condition. We return to these issues in our modeling analyses section.

The general finding that the RT set-size functions tend to be squeezed toward one another in the mixed condition relative to the pure conditions is consistent with the idea that there is adjustment of random-walk (RW) response criteria (i.e., the values of OLD and NEW in Figure 1) across the conditions, leading to speed-accuracy tradeoffs. In the pure-VM condition, participants set the criteria far from the starting point to achieve reasonable accuracy; in the pure-AN condition they set the RW criteria at intermediate locations; but in the pure-CM condition they can set the criteria close to the starting point given the ease of the task. This form of condition-specific RW-criterion placement is not available in the mixed-lists case, so the criteria are likely set at an intermediate location. This change in the location of the RW criterion settings

is probably one of the major reasons why RTs for VM-new items get *shorter* in the mixed condition compared to the pure condition, but that the false-alarms associated with these items skyrocket (Figures 3 and 4). However, mere adjustment of the RW response criteria cannot explain all the changes across the pure and mixed conditions. For example, note that observers make virtually zero errors on the CM-old items in the pure condition, but make substantial errors on these same items in the mixed condition (Figure 4). If all that changes is that observers set the RW response criteria *farther* from the starting point in the mixed condition compared to the pure-CM condition, then the opposite pattern of error rates should be observed.

Still another hypothesis to consider is that there are generalized improvements in the evidence-accumulation process (improved drift rates) in the pure conditions relative to the mixed. Such improvements might arise for multiple reasons, including more effective placements of the drift-rate criteria in the pure conditions relative to mixed; or perhaps strategy-dependent mechanisms that operate more efficiently in the pure conditions (e.g., an item-response learning strategy in pure CM). Although improved drift rates are indeed likely to be an important part of the story, this hypothesis also fails to provide a complete explanation of the results. For example, as noted above, error rates for AN-new items are significantly *reduced* in the mixed condition compared to the pure; thus, apparently, something more than generalized improvement in drift rates is also at work.

In sum, our interim discussion suggests that multiple processes are involved in mediating the relation between pure-list and mixed-list performance in probe-recognition memory search. We now turn to the formal-modeling section of our article in an attempt to develop an account of the overall patterns of data and to elucidate the underlying mechanisms.

The Formal Model

As we previewed in our introduction, we use an extended version of the exemplar-based random-walk (EBRW; Nosofsky & Palmeri, 1997; Nosofsky et al., 2011) model as applied to probe recognition to develop an account of the data (for previous applications of simpler versions of the model to CM, VM and AN performance, see, e.g., Cao et al., 2018; Nosofsky, 2016; Nosofsky et al., 2014a,b). A schematic illustration of some of the main components of the EBRW model is presented in Figure 7. First, consider the study items from the memory set of the current trial. According to the model, each of the study items is stored in memory as an individual exemplar. The memory strength of each exemplar is presumed to decrease solely as a function of the lag with which it was presented on the study list (with the most recently presented items having the shortest lags).[2] Based on evidence reported by Donkin and Nosofsky (2012), we assume specifically that the memory strength decreases as a power function of lag *j*:

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

where β reflects the rate of decrease and α reflects asymptotic strength at long lags. The differential memory strengths are represented schematically in panel A of Figure 7, where the larger circles represent exemplars with greater memory strength.

As illustrated in panel B of Figure 7, when a test probe is presented, exemplars stored in memory are "activated" and "race" to be retrieved (cf. Logan, 1988). The exemplars race with rates that are proportional to their activations. The degree to which exemplar j (e_j) is activated by test-item i (t_i) is a joint function of exemplar j's memory strength and its similarity to test item i. For the present types of object-image stimuli, we assume for simplicity a binary match-mismatch similarity relation, with the similarity of an object to itself set equal to 1, and the

similarity between mismatching objects set equal to a free parameter s ($0 < s < 1$). The degree to which t_i activates e_j is then given by:

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

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

Thus, the memory-set exemplars that are most highly activated and most likely to be retrieved are those that match the test probe and that have short lags.

To apply the EBRW model to old-new recognition tasks, it is assumed that the observer establishes "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 Figure 7B) are activated and race to be retrieved. The degree of criterion-element activation is independent of the presented test probe, although it may depend on factors such as memory-set size. We presume that the level of criterion-element activation is at least partially under the control of the observer.

Finally, the retrieved exemplars and criterion elements drive a random-walk process that determines the "Old" vs. "New" decisions (Figure 7, Panel C). As discussed in our introduction, 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 process, if an old exemplar is retrieved, the random-walk counter takes a step toward the "OLD" response criterion; whereas if a criterion element wins the race, the random-walk counter takes a step toward the "NEW" response criterion. The retrieval process continues until one of the response criteria is reached, at which point the observer emits the appropriate response.

Given the assumptions described above (and some further technical assumptions described by Nosofsky and Palmeri, 1997, primarily that the distributions of each racing

exemplar are independent exponential distributions with parameters determined by the activation strength), it turns out that, on each individual step, the probability that the random-walk counter steps toward the +*OLD* response threshold is given by:

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

where A_i represent the summed activation of the test probe to all the memory-set items:

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

and c is the level of criterion-element activation. (The probability that the random walk steps toward the -*NEW* response boundary is given by $q_i = 1 - p_i$.)

Eq. 3 can be described as a 'familiarity' model, with familiarity represented by total activation A_i. Whether a step is taken toward the 'OLD' boundary is also influenced by c, which is best thought of as a criterion that will vary with conditions. This can be seen with reference to Figure 2: When the distributions (i.e. the distributions of A_i for new and old items) are far apart, as in pure CM conditions, then a low value of c will work well to produce good performance; when the distributions are close together then both old and new items will have high values of A_i and it will be necessary to have a high value of c to discriminate them. The observer is presumed to learn an appropriate setting of c through experience in the task, such that the summed activation (A_i) tends to exceed c when the test probe is old, but tends to be less than c when the test probe is new.

Whether c should vary with set size is not very clear. More traces are activated for larger set sizes, so one might think A_i would grow with set size. However, strength of traces falls off with lag, so the increase in A_i from non-matching traces is modest. Even more important, the primary contribution to A_i is a trace that matches the test item, and this trace is equally likely to

be in different serial positions, so on average has a larger lag for larger lists, and hence is weaker. Finally, a participant may find it difficult to change criteria on a trial by trial basis. Nonetheless we initially allowed for the possibility that *c* might vary with set size, but in all the models we examined the best estimates showed no evidence that it did so. In the models we apply and report in this article, we therefore held *c* fixed across the different set sizes.

The development thus far has considered the role of only the items in the current memory set. However, a key to providing a full explanation of CM, VM and AN performance requires formalizing the role of the study and test items presented on previous trials of the experiment.

Our current extended model implements the influence of past trials by assuming that exemplars from the past (before the current memory set exemplars) can also be activated when the test occurs, as depicted in Figure 7B. Different versions of the model arise by making alternative assumptions about the nature of these past exemplars. For example, according to one version of the model, when a LTM exemplar is retrieved, it acts to add familiarity, in the same way as retrieved exemplars from the current memory set, so the random-walk counter moves toward the +*OLD* threshold*,* regardless of whether the LTM exemplar originally served as a memory-set item, a target test probe or a foil test probe. We refer to this version of the model as an *item-familiarity* model.

Alternatively, according to a second version of the model, the observer stores the response labels associated with the test probes from previous trials. If a LTM exemplar with an "old" response label is retrieved, it will lead the random-walk counter to step toward +*OLD;* whereas if a LTM exemplar that served as a "New" test probe is retrieved, it will lead the random-walk counter to step toward -*NEW.* We refer to this second version of the model as an *item-response-learning* model.

In the general form of the extended model, the probability that the random walk steps toward +OLD given test probe i is given by

$$p_i = (A_i + LTM_{Old}[i]) / [(A_i + LTM_{Old}[i]) + (c + LTM_{New}[i])], \quad (5)$$

where $LTM_{Old}[i]$ and $LTM_{New}[i]$ denote the long-term exemplar activations yielded by test probe i that drive the random walk toward the OLD and NEW response thresholds, respectively.[3] Different assumptions about the values of the LTM terms will yield models described above as 'item-familiarity' and 'item-response-learning'. There are a variety of special cases and constraints on parameters that will be covered when the model variants are presented. Note that the model as implemented with Eq. 5 combines the activation of current list items (A_i) and the long term contributions into a single process, even when the long-term component is item-response based.

Given the values of the random-walk step probabilities (p_i in Equation 5) and the settings of the random-walk criteria (OLD and NEW illustrated in Figure 7C), the computational formulas for choice probabilities and mean expected number of steps for correct and error responses have been derived in previous work (Busemeyer, 1982; Feller, 1968; for statements of these formulas in the context of the EBRW model, see Nosofsky & Palmeri, 1997, pp. 269-270). Predictions of mean RTs also require the estimation of a residual-time parameter t_0 reflecting processes such as encoding and motor-execution time; and a scaling parameter κ for converting individual steps in the random walk into time units.

Model Fitting Results

Our goal at this initial stage of research is to use the formal models to characterize the major qualitative patterns of results across the mixed and pure conditions. Accordingly, we limit consideration to fitting the mean correct RTs and error proportions at the group level. Once the formal-modeling directions that appear to be the most promising are identified, future research can be aimed at addressing more detailed issues, such as individual differences in performance, RT-distribution data, and distinctions between correct- versus error-RTs.

For simplicity, we used a weighted least-squares criterion of fit. In particular, we fitted the model to the mean correct-RT and error proportions data of: (a) the new items as a function of conditions (CM, VM, AN) and set size and (b) the old items as a joint function of conditions, 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 highly interpretable results were obtained when the mean RT data (measured in sec) and the error-proportion data were given equal weight, and when the individual data points for the new items were given 6 times the weight of the individual data points of the old items (justified because sample sizes for the individual new-item data points are on average 6 times greater than for the individual old-item data points because they are not broken down by lag.) Rather than allowing all parameters to vary freely, we imposed various constraints on the parameters with the aim of achieving theoretical parsimony; however, as described below, based on preliminary model-fitting results we also fitted more flexible versions when some of the limitations of the constrained models became evident. Summary fits of the models described below are reported in Table 2.

In general, the free parameters that need to be estimated in fitting the models to the data include: the memory-strength parameters (β and α in Equation 1); the similarity-mismatch

parameter *s* (Equation 2); the value of *c* (Equation 3); the random-walk criteria *OLD* and *NEW*; the residual-time parameter t_0 and time-scaling parameter κ; and the various *LTM* parameters that enter into Equation 5. Listings of the free parameters along with their best-fitting values for the models described below are provided in Tables 3-6.

Mixed Condition. Fitting the alternative models to the mixed-conditions data is more straightforward than fitting the models to the pure-conditions data, because certain parameters should be invariant across the conditions in the mixed case. In particular, across the CM, VM and AN conditions, there should certainly be fixed values of: t_0 and κ; fixed values of the criterion-element parameter *c*; and fixed values of the random-walk criteria *OLD* and *NEW*. In addition, for purposes of theoretical parsimony, our initial presumption was that the STM-related parameters β, α and *s* would be invariant across the CM, VM and AN conditions. The key parameters that are expected to vary across the conditions are the *LTM*-related parameters.

Our first step was to fit an item-familiarity model to the data. As discussed earlier, the familiarity-only model is defined by setting all LTM_{New} values in Equation 5 equal to zero. In addition, we constrained the settings of the LTM_{Old} values such that long-term familiarity for the CM-target and VM items was at least as great as for the CM-foils; with the AN items having the lowest long-term familiarity. (The familiarity for the AN items might still be non-zero, however, owing to similarity relations that AN test probes might have to items presented on previous trials.) Note as well that the LTM-familiarity values for VM targets and VM foils must logically be identical to one another (because whether each individual VM item serves as a target or foil on any given trial is random); and that the same holds true of the AN targets and AN foils.

The summary fits of the item-familiarity model are reported in Table 2, with best-fitting parameters reported in Table 3. The predictions from the model are presented in graphical form

in the supplementary materials (Figure S2). We do not discuss here the best-fitting parameters in depth because the fits from the model are substantially worse than are versions of the item-response-learning models (described below). In addition, inspection of the graphical predictions from the model revealed a number of salient qualitative shortcomings. First, the model was unable to account for the finding that RTs for the CM-new items were shorter than for the AN-new items: the reason is that CM foils are presumed to be at least as familiar as are AN foils, and greater levels of familiarity should interfere with an observer's ability to correctly reject the foils. Second, the item-familiarity model predicted substantial set-size effects for the CM-new and AN-new items (for both RTs and error proportions), whereas the observed set-size functions for these item types were essentially flat. The reason for this failed prediction is that the model estimated zero contribution of LTM familiarity for the CM-new and AN-new items, so the random walk was being driven solely by the STM components. Furthermore, in general, as set-size increases, summed activation increases for the new items, leading the random walk to step incorrectly towards the OLD response criterion, resulting in both increased probability of errors and longer correct-rejection RTs. Yet another difficulty for the item-familiarity model was that it predicted that AN targets would have substantially higher error rates and longer correct-old-response RTs than the CM and VM targets, but the observed data tend to go in the opposite direction (except for the largest set size). The reason for this failed prediction is that the CM and VM targets are presumed to have LTM familiarity values that are at least as great as the AN targets, and higher familiarity should facilitate old responses for the CM and VM targets compared to the AN targets.

Our next step was to fit a parameter-constrained item-response-learning model to the data. The model was the same as the item-familiarity described above, except it made allowance

for item learning of "new" responses (i.e., non-zero estimates of the LTM_{New} values in Equation 5). Although we allowed the LTM_{New} values to vary freely, our expectation was that the LTM_{New} estimates would be greatest in magnitude for the CM-foil items because of their consistent mappings to new responses. For simplicity, we also imposed the constraints that LTM *new*-response strengths for CM-targets were equal to zero and that LTM *old*-response strengths for CM-foils were equal to zero. In addition, we assumed that any LTM *old*-response strengths would be greatest in magnitude for CM targets, intermediate for VM items, and lowest in magnitude for AN items. (The assumption that LTM_{Old} should be at least as great for CM targets as for VM targets arises for two reasons: first, CM targets receive consistent mapping to old responses; second, in the current design, CM targets occur with greater frequency in the memory sets than do VM items.) Again, we imposed the logical constraint that these LTM response strengths were symmetric for VM targets and VM foils, and symmetric for AN targets and AN foils, because assignment of VM and AN items as targets versus foils was randomized.

The summary fits of the baseline IR model are reported in Table 2, with best-fitting parameters reported in Table 4. Again, we provide the predictions from the model in graphical form in the supplementary materials (Figure S3). In short, the baseline IR model yielded some significant improvements compared to the familiarity-only model, but still suffered from important shortcomings. Among the main improvements was that, because it allows estimate of a LTM_{New} response-strength for the CM foils (see Table 4), it was able to account for the very short RTs associated with these items, and did a much better job of predicting the nearly flat set-size functions associated with those items. In addition, it predicted the slight advantages in responding for the CM targets compared to the VM targets. (There are multiple parameter settings that yield this slight CM advantage. These include setting the LTM_{Old} values slightly

higher for the CM targets than for the VM targets; or assigning a small-magnitude LTM_{New} value to the VM targets, which causes interference with the random walk's movement toward the OLD response criterion. The latter is a reasonable possibility because the VM items have been mapped to both old *and* new responses on previous trials.)

Despite these improvements, the baseline IR model was still unable to capture the detailed results involving the AN items: It still predicted a significantly increasing set-size function for the AN foils, and still predicted longer RTs and higher error rates for the AN targets compared to the CM and VM targets. The model also failed to capture the manner in which the set-size functions for the AN targets exhibit a "cross-over" with respect to the set-size functions for the CM and VM targets (see Figures 3 and 4). These limitations of the baseline IR model suggested strongly that it was mischaracterizing the nature of the memory representations involving the AN items in the mixed condition.

In considering the reason for these limitations, we took note of the fact that whereas the CM and VM items are viewed repeatedly throughout the experiment, each time an AN item is presented in a memory set it is being experienced for the very first time by the observer. Much as in a *von Restorff* effect, the presentation of the highly novel item is likely to be attention-attracting and to lead to a highly distinctive memory representation for the item. However, this attention attraction and distinctive coding may be short-lived, as the observer continues to experience subsequent items from the memory sets (including *new* AN items) that need to be stored in STM.

To potentially account for these effects, we decided to fit an extended version of the IR model to the data that allowed separate STM-coding parameters for the AN items. Although the modeling is post hoc in character, we believe it needs to be taken seriously, given that it allows

dramatically improved fits to the complete set of data, and given that the parameter adjustments that are introduced have some psychological plausibility. We modeled the attention-attracting properties of the AN items by assuming that upon initial presentation, the memory strength was adjusted by a *boost* parameter (where *boost* > 1); but that this enhanced memory strength may fade away and be short-lived by allowing separate β and α values in the AN memory-strength function. Finally, we made allowance for the possibility that the AN items are more distinctive than the CM and VM items by allowing a separate similarity-mismatch parameter (*s*) for the AN test probes. In all other respects, this extended IR model was the same as the baseline version described above.[4]

The summary fits of the extended IR model are reported in Table 2, with best-fitting parameters reported in Table 5. (The modeling analyses revealed that the LTM_{New} values associated with the VM and AN items could be set equal to zero with no loss in fit.) The parameter-described predictions from the model are presented in graphical form in the right panels of Figures 8-11. To aid comparisons, we replot the observed data from the mixed condition in the corresponding left panels of the figures. As can be seen from comparing the summary fit measures in Table 2, the fit of this extended IR model is substantially better than those of the alternative models that we discussed above. Furthermore, inspection of Figures 8-11 reveals that the model provides an outstanding account of the overall patterns in the data. For the probability-of-error data, these successful accounts include: a) the extremely accurate responding for the CM and AN foils, and the extremely inaccurate responding for the VM foils; b) the nearly flat set-size functions for the CM and AN foils, and the steeply rising set-size function for the VM foils; c) the intermediate and more tightly packed accuracy levels for the CM, VM, and AN targets; d) the intermediate slopes associated with the set-size functions for the target items; e)

the slight advantage in accuracy associated with the CM targets compared to the VM targets; and f) the tendency for the AN-target set-size function to cross-over with respect to the CM-target and VM-target set-size functions. Moreover, the model captures the more nuanced set-size by lag functions shown in the panels of Figure 10. First, it captures the result that the main controlling variable is lag rather than set size. Second, it accounts well for the curvilinear shape of the lag functions. Third, although the functions are somewhat noisy, it provides a good overall quantitative account of the accuracy levels for each of the item types at each of the lags. Finally, although the error data associated with the different set sizes are nearly overlapping once one conditions on lag, for the CM and VM items there does seem to be a slight tendency for the set-size-3 function to lie above the set-size-6 function; and for the set-size-6 function to lie above the set-size-9 function. This tendency is also captured by the model.

Our discussion above was with respect to the error-probability data only. However, almost all of the patterns noted above were mirrored in the observed and predicted correct-RT data as well – see Figures 9 and 11. In short, while making use of a reasonably small number of free parameters, the model is providing an outstanding account of an extremely rich set of short-term probe-recognition data in the present mixed-list CM-VM-AN paradigm.

The best-fitting parameters from the model (Table 5) provide suggestions about the underlying memory mechanisms that governed performance in the task. Perhaps of greatest interest are the estimates of the LTM-related parameters: The most salient result is that the model estimates a large-magnitude value of LTM_{New} for the CM-NEW test probes, while the LTM_{New} estimates for the VM and AN items can be set at zero with no loss in predictive accuracy. (The fit of the model with LTM_{New} set at zero for the CM-NEW items was markedly worse than the one presented here.) This result suggests that an item-response-learning process for the

consistently-mapped new items did indeed occur under the present mixed-list conditions. At the same time, the LTM_{Old} estimate for the CM-OLD items is only slightly higher than the one associated with the VM and AN items, suggesting only weak item-response learning for the CM-OLD items. At present, the reason for this asymmetry in old-item versus new-item response learning remains something of a mystery to us, although we speculate about its basis in our General Discussion. Our tentative conclusion, however, is that not all item-response learning takes place automatically at the same rate in the context of the present kinds of mixed-list, short-term probe-recognition tasks.

Finally, we note as well that the STM-parameter estimates for the AN items are in accord with the intuitions that we developed above: the *boost* parameter is slightly greater than unity, capturing the initial attention-attracting properties of the AN items; the decay parameter β for the AN items is greater in magnitude than for the CM and VM items, reflecting that this enhanced memory strength may be short-lived; and the similarity-mismatch parameter *s* is smaller in magnitude for the AN test probes than for the CM and VM test probes, reflecting that the AN test probes are more distinctive than are the CM and VM test probes.

Pure Conditions. Although our main goal involved formal modeling of the mixed-conditions data, for completeness we also report example fits for the pure-conditions data. Unfortunately, as noted earlier, there is no longer any expectation that participants will hold criterion settings constant across the CM, VM and AN conditions when those conditions are tested in between-subjects fashion. As a result, certain parameter estimates become non-identifiable. For example, for the present pure-lists design, the parameters *c* and LTM_{New} cannot be estimated separately in the VM and AN conditions, one can only estimate their sum (see

Equation 5). Here, for purposes of illustration, we report fits from an item-response-learning version of the model. We arbitrarily set LTM_{New} equal to zero in the VM and AN conditions while allowing *c* to vary freely. To draw some potentially meaningful comparisons with the results from the mixed-condition, we hold fixed the time-scaling constant κ at 38 msec, the value estimated from the fits to the mixed-condition data.[5]

The summary fits of the item-response learning model to the pure-conditions data are reported in the lower panel of Table 2, with best-fitting parameters listed in Table 6. The parameter-described predictions from the model are presented in graphical form in the right panels of Figures 12-15. Again, to aid comparisons, we replot the observed data from the pure conditions in the corresponding left panels of the figures. In brief, the model yields reasonably good accounts of the pure-conditions data, a result that provides a general replication of previous tests of the EBRW probe-recognition model in similar pure-list paradigms (e.g., Nosofsky et al., 2014a,b). We emphasize, however, that in the present experimental design, the psychological interpretation of some of the model-fitting results is uncertain (because we cannot estimate with any precision the values of the *c* and *LTM* parameters across the conditions).

One parameter-estimation result that is of central interest, however, concerns the RW-criterion parameters (*OLD* and *NEW*). As can be seen in Table 6, the criteria are placed closest to the starting point in the CM condition; at intermediate locations in the AN condition; and farthest from the starting point in the VM condition. Recall that we had anticipated such a pattern in introducing the research issues in our article. To review, because the evidence-accumulation process operates highly efficiently in the CM condition, participants can set the RW-criteria close to the starting point without any appreciable loss in accuracy. And because the evidence-accumulation process operates with low efficiency in the VM condition, participants

need to set the RW-criteria far from the starting point to achieve reasonable accuracy. Within the present modeling framework, making allowance for these shifts in RW-criterion placements across the conditions is crucial: As reported in the bottom panel of Table 2, a version of the model in which the RW-criteria are constrained to be the same across the CM, VM and AN conditions yields substantially worse accounts of the data.

Comparing the criterion placements in the pure conditions to those in the mixed condition provides additional insights. In the mixed condition, observers needed to set the RW-criteria at single locations that could not vary when making old-new judgments for CM, VM and AN items. Inspection of Table 5 reveals that in the mixed condition, observers chose intermediate "compromise" locations of roughly the same magnitude as occurred for the pure-AN condition (compare to Table 6). In the pure-CM condition, the RW-criteria are pushed "inward" compared to the mixed condition; whereas in the pure-VM condition, the RW-criteria are pushed "outward" compared to the mixed condition. The changed RW-criterion placements across the pure and mixed conditions go a long way towards explaining some of the major changes in performance patterns across these conditions. For example, according to this interpretation, observers did not accumulate as much evidence in the mixed-VM condition as they did in the pure-VM condition. This explains why RTs for the VM items were markedly shorter in the mixed condition than in the pure condition, but why the error rates (averaged across old and new test probes) for the VM items tended to skyrocket in the mixed condition (see Figures 3 and 4).

General Discussion

Summary

A classic finding in probe-recognition memory search is the dramatic contrast in performance across consistent-mapping (CM) and varied-mapping (VM) conditions. Even with relatively little practice, observers perform dramatically better under CM-training conditions than under VM-training condition. Furthermore, unlike in VM, under CM-training conditions performance is relatively unaffected by increases in memory set size or by increases in the lag with which targets are presented on the study lists.

One of the major reasons for the performance contrast appears to be that participants can rely on different forms of LTM under CM conditions, but must rely on capacity-limited STM under VM conditions. Although a variety of LTM strategies are available (e.g.,. see Logan & Stadler, 1991), recent research reported by Nosofsky et al. (2014a) and Cao et al. (2017, 2018) suggested that item-response learning plays a major role in enhancing CM memory-search performance during the early stages of practice. Such forms of item-response learning have been theorized to be among the key mechanisms that lead to "automatic" skilled performance in varieties of cognitive tasks (Logan, 1988).

However, developing a clear picture of the underlying mechanisms that distinguish CM and VM performance is difficult, because the tasks are generally conducted in between-blocks or between-subjects fashion. As a result, rather than reflecting the operation of hard-wired mental mechanisms, at least some of the differences may reflect changes in strategy as well. One type of strategy change involves the use of changed criterion settings across tasks. These include changes in the total amount of accumulated evidence required before an observer is willing to

initiate a response (i.e., changes in the RW-criterion settings); as well as criterion changes that drive the step-by-step evidence-accumulation process itself (i.e., changes in the "drift-rate" criterion [Ratcliff, 1985], modeled here in terms of "criterion-element activation"). A second type of strategy change may reflect the types of LTM representations that observers rely upon across different conditions. For example, rather than being an automatic, hard-wired mechanism, observers may "choose" to rely on item-response learning under CM conditions because it is a generally useful strategy.

The key idea that we pursued in the present work to address these issues was to test CM versus VM probe-recognition memory search in a mixed-list design, and to compare performance in the mixed-list design to performance observed under pure-list conditions (for closely related work, which we discuss in more depth below, see Strayer & Kramer, 1994a). In the mixed-list design, items that receive CM and VM training appear within the same lists, and observers cannot anticipate whether they will be tested on a CM or a VM item. To achieve even greater diagnosticity, we included "unmapped" all-new (AN) items in the paradigm as well. The critical point is that the mixed-list design forces observers to use a common set of criterion settings across the different item types. By holding this strategic factor constant across conditions, one can better evaluate the nature of hard-wired LTM mechanisms associated with CM versus VM training.

In general agreement with earlier results reported by Strayer and Kramer (1994a), we observed some dramatically different patterns of CM, VM, and AN performance across the mixed-list and pure-list conditions. Many of these changes were indeed consistent with the hypothesis that observers adopted changed criterion settings across the pure and mixed conditions. Overall, RT differences across the CM, VM and AN conditions were vastly reduced

in the mixed condition compared to the pure conditions; but error rates for VM new items skyrocketed in the VM-mixed condition compared to the VM-pure condition. Our formal modeling suggested that a major factor driving these changes was that, in the pure conditions, participants set the RW-criteria close to the starting point in CM, at intermediate locations in AN, but far from the starting point in VM. In addition, participants can set the drift-rate criterion for what constitutes "old" evidence at a lax location in the pure CM and AN conditions, but must set it at a strict location in the pure VM condition (see Figure 2). By contrast, in the mixed condition, the criteria are held fixed across the different item types, causing many of the performance differences across conditions to be vastly reduced (e.g., RT differences across tasks), but others to be magnified (e.g., error rates for VM-new items).

Under the fixed-criterion situation that operates in the mixed-list condition, the key question that then arises is whether fundamental performance differences continue to be observed across the CM, VM and AN conditions. Among the main results was that performance on CM-new items continued to be vastly superior to performance on VM-new items; in addition, RTs for CM-new items were significantly shorter than for AN-new items (with error rates on both types of items being near floor). By comparison, there was relatively little difference in performance across CM-old, VM-old, and AN-old items, regardless of whether performance was measured in terms of RTs or error rates. In addition, all three types of items showed significant lag x set-size functions in the mixed-list situation, whereas the lag x set-size functions for CM-old items in the pure condition were essentially flat. Thus, there was a dramatic change in the observed patterns of performance for the CM-old items across the pure and mixed conditions – the types of "automaticity" attributed to processing of CM-old items under pure-list conditions were no longer observed under the mixed-list conditions. Formal modeling indicated that item-

response learning continued to operate for the CM-new items, even under the present mixed-list conditions. By comparison, any differences in old-item-response learning that may have existed across the CM-old and VM-old items were small in magnitude.

We should emphasize that the patterns of results we observed in the mixed-list experiment for the CM and VM items are extremely robust. Cao (2018) conducted a second mixed-list experiment nearly identical to the present one, except that the AN items were not included in the design. As in the current study, there was a dramatic performance advantage for the CM foils compared to the VM foils, regardless of whether performance was measured in terms of error rates or RTs. By contrast, there was only a small and non-significant advantage for the CM targets compared to the VM targets. Formal modeling again indicated that new-item response learning had occurred for the CM foils but not the VM items; however, there was little difference in the magnitude of old-item response learning for the CM targets versus the VM items.

We did not anticipate this apparent asymmetry in old- versus new-item response learning for the CM items in the mixed-list situation. Determining its basis remains a topic for future research. One possibility involves an asymmetry between old and new test probes and whether or not they are present in the memory sets themselves. Along with VM items, the CM targets appear often in the study lists, but each individual CM or VM target rarely appears as a test probe. Possibly, the cognitive system may interpret the high frequency of appearance on study lists along with the low frequency of appearance as test probes as a form of inconsistent mapping, so item-response learning of the CM targets is slowed. By contrast, CM foils never appear on the study lists, and when they appear as test probes they are always assigned new responses, so item-response learning takes place more efficiently for the CM foils. Future tests

of this idea might systematically manipulate the conditional probability with which individual CM targets appear as test probes when they appear on study lists.[6] It is also possible that the presence of VM items on the same study list causes confusion because some study list items are not consistently-mapped targets. Such inconsistency might prevent learning about all study list items, and even inhibit learning about such items at test.

A complication that arose in the modeling of the mixed-lists data involved the need to extend the item-response-learning model with a special set of STM parameters for the AN items. For reasons of theoretical parsimony, we had hoped to characterize all differences between CM, VM and AN mixed-list performance in terms of changes in the LTM parameters, but this version of the baseline model had significant shortcomings. Instead, the results from the best-fitting model suggested that AN items had attention-attracting properties when they appeared on the current study lists, and also that AN test probes were coded in a more distinctive fashion than were VM and CM test probes. Indeed, according to the extended-IR model, the main reason why AN foils had lower false-alarm rates than did VM foils is that the AN foils activated items on the current memory set far less than did the VM foils (see parameter estimates in Table 5). The present model-fitting results suggest the possibility of strong interactions between LTM and STM, with the nature of the way that items are coded in STM being influenced by LTM representations. Future research is needed to provide converging evidence for the current proposal involving the STM coding of the AN items.

Relations to Previous Work

As noted throughout our article, Strayer and Kramer (1994a) conducted studies related to our own. These researchers also contrasted CM-like and VM-like conditions across mixed-list

and pure-list designs; and they also used formal evidence-accumulation modeling to help characterize the results. They did not include AN items in their designs.

Most of Strayer and Kramer's (1994a) experiments involved hybrid memory-search designs involving the simultaneous presentation of multiple test probes; however, their Experiment 2 involved a single test-probe design. However, their paradigm blurred the distinction between CM and VM. On each trial, a single memory set would be presented, but it was then followed by 24 consecutive test-probe presentations, rather than a single test probe (as in the standard task). Thus, even their VM condition can be considered to be a type of "within-trial" CM condition, because items receive consistent mappings to old and new responses across 24 consecutive tests. In addition, in Strayer and Kramer's paradigm, subjects were allowed to study the memory set for as long as they wished. This procedure would tend to reduce set-size differences and also tend to erase any systematic effects of lag on performance (because participants could study and rehearse the memory sets in multiple idiosyncratic ways).

Another procedural difference is that Strayer and Kramer (1994a) used categorized word lists instead of arbitrary images of objects. CM items were drawn from one set of pre-existing semantic categories and VM items from other sets of pre-existing semantic categories. Subjects likely adopted category-coding strategies under such conditions. By contrast, research reported by Cao et al. (2017) suggested strongly that participants do not adopt category-coding strategies during the early stages of practice with the types of arbitrary object images we used in the present experiments. Still another procedural difference is that Strayer and Kramer (1994) studied the performance of highly practiced subjects, whereas our studies were focused on performance during the early stages of learning.

Given the numerous differences across our paradigms, it is difficult to directly compare our results to the ones reported by Strayer and Kramer (1994a). Broadly speaking, they observed bigger differences in CM vs. VM RT in blocked conditions than in mixed, which is one of our central findings as well. However, they did not distinguish between old vs. new test probes in reporting the RT data (see their Figure 2), so the comparisons here are limited. Strayer and Kramer (1994a, Table 4) reported little difference in VM accuracy for either targets or foils across the mixed and blocked conditions; this result contrasts dramatically with our own, where false alarms for VM foils skyrocketed in the mixed-list case. Unfortunately, so many factors differ across our experiments that it is difficult to know which factor is the key one that is responsible for this difference in results.

In their formal modeling, Strayer and Kramer (1994a) fitted Ratcliff's (1978) diffusion model to RT-distribution data. The modeling was intended to be descriptive in nature, and Strayer and Kramer (1994a) did not develop process-oriented accounts of the mechanisms that drive the evidence-accumulation process in that model. Instead, the purpose of the modeling was to estimate overall evidence-accumulation drift rates across conditions as well as the RW-criterion settings. (Apparently, they did not estimate separate drift rates for old versus new items and they did not make reference to the "drift-rate" criterion, which is critical to our own modeling.) Strayer and Kramer (1994a) considered changes in RW-criterion settings across mixed-list and pure-list conditions to be "strategic" effects; and changes in drift rates to be "data-driven" effects. Our own results converge strongly with theirs in finding systematic changes in RW-criterion settings across the mixed and pure conditions, with the pattern of changes in CM versus VM the same in the two studies. Thus, despite the numerous differences across our

experimental paradigms, this evidence for the role of shifting RW-criterion settings across mixed-list and pure-list conditions seems quite robust.

With regard to the evidence-accumulation process itself, Strayer and Kramer (1994a, pp. 337) summarize their modeling results by writing: "However, one surprising result of the modeling was the finding that the drift rate decreased from CM blocked to CM mixed conditions. This suggests some factor, in addition to the setting of response criteria, was responsible for the differences in performance between blocked and mixed CM conditions. We interpret this as evidence that strategic factors also affect the build up of perceptual-cognitive evidence, and that data-driven or bottom-up processing is not strategy independent. Thus, automatic components of performance can be modified by strategic components of performance." Our finding of dramatic changes in CM drift rates across the pure and mixed conditions provides converging evidence for this aspect of Strayer and Kramer's (1994a) findings as well. In addition, we have initiated the formal development of mechanistic accounts of the basis for these changes in CM drift rate. They appear to arise as a result of changes in drift-rate criterion settings across pure and mixed conditions, and perhaps to changes in the extent to which observers rely on item-response-learning strategies as well.

Conclusions

A major contributing factor to the dramatic contrast in performance in CM and VM probe-recognition memory search in pure-list conditions involves changed criterion settings across the tasks. The changed criterion settings pertain both to the total amount of accumulated evidence required before an observer is willing to initiate a response; as well as criterion changes that drive the step-by-step evidence-accumulation process itself. Once one controls for these

changes in criterion settings by testing a mixed-list paradigm, the CM-VM performance differences are dramatically reduced for old target items, but remain for new foils. The empirical pattern of results as well as formal modeling suggest that an automatic form of new-item-response learning continues to operate for new CM foils under the mixed-list conditions, but this form of item-response learning may be greatly weakened for the old target items. In addition, the manner in which items are encoded in STM during probe-recognition memory search may interact strongly with the nature of the LTM representations that develop as observers are trained in the tasks. Fleshing out the detailed basis for these latter findings is a central direction for future research.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes1. In *Psychology of learning and motivation* (Vol. 2, pp. 89-195). Academic Press.

Banks, W. P., & Atkinson, R. C. (1974). Accuracy and speed strategies in scanning active memory. *Memory & Cognition*, *2*(4), 629-636.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

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.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Busemeyer, J. R. (1982). Choice behavior in a sequential decision-making task. *Organizational Behavior and Human Performance*, *29*(2), 175-207.

Cao, R. (2018). The role of long-term memory in automaticity development. Unpublished Ph.D. dissertation, Indiana University Bloomington.

Cao, R., Nosofsky, R. M., & Shiffrin, R. M. (2017). The development of automaticity in short-term memory search: Item-response learning and category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(5),* 669.

Cao, R., Shiffrin, R. M., & Nosofsky, R. M. (2018). Item frequency in probe-recognition memory search: Converging evidence for a role of item-response learning. *Memory & cognition*, *46*(3), 450-463.

Cheng, P. W. (1985). Restructuring versus automaticity: Alternative accounts of skill acquisition. *Psychological review, 92,* 414-423.

Duncan, M., & Murdock, B. (2000). Recognition and recall with precuing and postcuing. *Journal of Memory and Language*, *42*(3), 301-313.

Feller, W. (1968). An introduction to probability theory and its applications, 3rd ed., Vol. 1. New York: Wiley.

Kuznetsova A, Brockhoff P., Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293-323.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms?. *Cognitive Psychology*, *22*(1), 1-35.

Logan, G. D., & Stadler, M. A. (1991). Mechanisms of performance improvement in consistent mapping memory search: Automaticity or strategy shift?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 478.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18*, 346–373.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465–501.

Nosofsky, R. M., Cao, R., Cox, G. E., & Shiffrin R. M. (2014a). Familiarity and categorization processes in memory search. *Cognitive Psychology, 75*, 97-129.

Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014b). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1524.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 188*, 280–315

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31(3)*, 608-629.

Osth, A. F., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review*, *126*,

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological review*, *92*(2), 212.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive science*, *27*(3), 525-559.

Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception & Psychophysics, 31(2)*, 160–168.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1–66

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

Singmann, H., & Kellen, D. (in press). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Neuroscience and Cognitive Psychology*. Psychology Press.

Strayer, D. L., & Kramer, A. F. (1994a). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(2)*, 318–341.

Strayer, D. L., & Kramer, A. F. (1994b). Strategies and automaticity: II. Dynamic aspects of strategy adjustment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 342.

Sternberg, S. (1966, August 5). High-speed scanning in human memory. *Science, 153*, 652– 654.

Sternberg, S. (2016). In defence of high-speed memory scanning. *The Quarterly Journal of Experimental Psychology*, *69*(10), 2020-2075.

Wolfe, J. M., Boettcher, S. E., Josephs, E. L., Cunningham, C. A., & Drew, T. (2015). You look familiar, but I don't care: Lure rejection in hybrid visual and memory search is not based on familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1576.

.

Appendix:

Statistical Analysis

We collected two dependent measures – the proportion of errors and the reaction time for correct responses. In each of the experimental conditions, corresponding to the factorial combination of Condition (CM, VM, AN), Set Size (3, 6, 9), and Probe (Old, New), we computed the mean of these two dependent measures for each subject, and applied this 3x3x2 design to the data from both experiments (Pure, Mixed).

In the Mixed experiment, a small number of subjects failed to provide at least one correct response in some of the more-challenging experimental conditions (e.g. VM, Set Size = 9), resulting in missing cells in the data. Rather than exclude these subjects from the analysis, or to adopt a post-hoc approach to impute these missing cells, we chose to perform statistical analysis using Linear Mixed Effects Models, which are well-suited to situations with missing data, and to unbalanced, hierarchical and/or repeated-measures designs (Singmann & Kellen, in press).

Analysis was performed using the "lme4" package (Bates, Maechler, Bolker, Walker, 2015) within the R programming language (R Core Team, 2019). Experimental manipulations were treated as Fixed Effects, with Condition (CM, VM, AN) and Probe (Old, New) coded as categorical factors, and Set Size (3, 6, 9) treated as a continuous variable. Defining the appropriate Random Effects structure can be difficult with some authors suggesting the full experimental structure (Barr, Levy, Scheepers, & Tily, 2013). In our case, however, because we have reduced the data to a single observation per observation cell, these hierarchical effects cannot be estimated; instead we adopt the standard convention of allowing for individuals to differ in their overall performance, irrespective of condition, by estimating a Random intercept for each participant. While these Mixed Models can be applied directly to the raw data, this

approach is sufficiently powerful to detect the theoretically relevant results with minimal additional complexity.

General Procedures:

In the sections below, we describe the results for each experiment separately, focusing first on correct response times, followed by the error proportions. First, we describe the procedure used to construct and evaluate different statistical models which was employed in all of the subsequent analyses unless otherwise noted.

Technical Notes:

1. By default, most statistical software utilizes Type III sum-of-squares within analysis of variance. However, this has the consequence of testing for Main Effects *after* accounting for the influence of interactions. We instead chose to utilize Type I, or *sequential*, sum-of-squares which better matches standard conceptual practice within psychology, starting with the simplest models and adding complexity only as needed. In cases with unbalanced designs, this approach can be problematic, as the order in which the terms are entered into the model can influence the results; however, we have a minimally-unbalanced design. For completeness, we checked all orders and found no difference in the results.

2. Estimation of the appropriate denominator degrees of freedom in Mixed Models can be challenging, especially when the Random Effects structure is complicated. We utilized Satterthwaite's (Satterthwaite, 1946) method to compute the appropriate df, as recommended by the authors of the lme4 package.

3. Hypothesis tests were assessed using the *lmerTest* (Kuznetsova, Brockhoff, Christensen, 2017) package in R, which allows users to define and test coefficient contrasts. We utilized the *contest1D* function and defined contrast vectors to conduct pairwise comparisons in marginal

means, as well as testing hypotheses such as “the slope of the set size function in the CM condition was flat.”

Model Selection:

We first began by constructing the full model within each of the two experiments. We then iteratively removed non-significant terms from the model specification until reaching the minimal model, the summary ANOVA of which is shown below.

Table A0.1

Pure - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	231.857	2	510	< .001
Set Size	-	7.934	1	510	< .001
Probe	Old, New	0.316	1	510	= .574
Condition:Set Size	-	3.251	2	510	< .05
Condition:Probe	-	3.814	2	510	< .05

Table A0.2

Pure - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	105.744	2	510	< .001
Set Size	-	98.714	1	510	< .001
Probe	Old, New	0.444	1	510	= .506
Condition:Set Size	-	32.803	2	510	< .001
Condition:Probe	-	3.643	2	510	< .05

Table A0.3

Mixed - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	230.417	2	3069	< .001
Set Size	-	88.390	1	3069	< .001
Probe	Old, New	6.649	1	3069	< .01
Condition:Set Size	-	9.321	2	3069	< .001
Condition:Probe	-	206.645	2	3069	< .001
Condition:Set Size:Probe	-	10.0225	3	3069	< .001

Table A0.4

Mixed - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	377.006	2	3074.2	< .001
Set Size	-	153.990	1	3074.2	< .001
Probe	Old, New	0.091	1	3074.2	= .763
Condition:Set Size	-	71.610	2	3074.2	< .001
Condition:Probe	-	254.303	2	3074.2	< .001
Set Size:Probe	-	35.2952	1	3074.2	< .001
Condition:Set Size:Probe	-	59.9135	2	3074.2	< .001

Hypothesis Tests:

In the following section, we report the supporting statistics to corroborate the claims made in the main text. Each is formatted by first quoting the original text (in *italics*), followed by the description of the specific statistical model used to test each claim. Supplemental tables are used to organize the results.

1. (Pure) *For both old and new test probes, the mean correct RTs are shortest in the CM condition, intermediate in the AN condition, and longest in the VM condition.*

We constructed two GLME models, separately for Old and New items; within each analysis, we examined the effects of Condition (CM, VM, AN) and Set Size (3, 6, 9) on the mean correct response time.

For Old test probes, there was a significant main effect of Condition, $F(2, 240) = 78.915$, $p < .001$ (see Table A1.1). Pairwise tests revealed a significant difference between CM (687 $\pm$ 38ms) and VM (1184 $\pm$ 38ms), $t(240) = -12.480$, $p < .001$, between CM and AN (886 $\pm$ 38ms), $t(240) = -4.988$, $p < .001$, and VM and AN, $t(240) = 7.492$, $p < .001$.

Table A1.1

Pure (Old) - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	78.915	2	240	< .001
Set Size	-	1.763	1	240	= .186
Condition:Set Size	-	1.317	2	240	< .05

For New test probes, there was also a significant main effect of Condition, $F(2, 240) = 155.276$, $p < .001$ (see Table A1.2). Pairwise tests again showed a significant difference between CM (596 $\pm$ 37ms) and VM (1235 $\pm$ 37ms), $t(240) = -17.605$, $p < .001$, CM and AN (891 $\pm$ 37ms), $t(240) = -8.123$, $p < .001$, and VM and AN, $t(240) = 9.482$, $p < .001$.

Table A1.2

Pure (New) - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	155.276	2	240	< .001
Set Size	-	7.013	1	240	< .01
Condition:Set Size	-	1.983	2	240	< .05

2. (Pure) *In both the VM and AN conditions, RTs get longer as set size increases, whereas the set-size functions in the CM condition are flat.*

We next constructed a GLME model including data from both Old and New test probes. We examined the effects of Condition (CM, VM, AN) and Set Size (3, 6, 9) on mean correct response time.

We found a significant interaction between Condition and Set Size, $F(2, 510) = 3.201$, $p < .05$ (see Table A1.2). Post-hoc t-tests comparing the estimated Set Size slope for each condition against 0.0 revealed a non-significant difference in the CM condition (-2.7 $\pm$ 8 ms/item), $t(510) = -0.347$, $p = .729$; by contrast, the slopes for the VM (24.3 $\pm$ 8 ms/item), $t(510)$

= 3.157, $p < .01$, and AN (15.6 ± 8 ms/item), $t(510) = 2.031$, $p < .05$, conditions were significantly different from 0.0.

Table A2.1

Pure - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	228.301	2	510	< .001
Set Size	-	7.812	1	510	< .01
Condition:Set Size	-	3.201	2	510	< .001

3. (Pure) *These same patterns are observed for the mean proportions of errors.*

- *For both old and new test probes, the proportion of errors is smallest in the CM condition, intermediate in the AN condition, and largest in the VM condition.*

For Old test probes, there was a significant main effect of Condition, $F(2, 240) = 42.963$, $p < .001$ (see Table A3.1). Pairwise tests revealed a significant difference between CM (2.3 ± .7%) and VM (2.3 ± .7%), $t(240) = -9.242$, $p < .001$, between CM and AN (6.8 ± .7%), $t(240) = -5.240$, $p < .001$, and VM and AN, $t(240) = 4.002$, $p < .001$.

Table A3.1

Pure (Old) - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	42.963	2	240	< .001
Set Size	-	60.540	1	240	< .001
Condition:Set Size	-	14.256	2	240	< .001

For New test probes, there was a significant main effect of Condition, $F(2, 240) = 65.756$, $p < .001$ (see Table A3.2). Pairwise tests revealed a significant difference between CM (0.8 ± .8%) and VM (12.1 ± .8%), $t(240) = -11.416$, $p <$

.001, between CM and AN (7.3 ± .8%), $t(240) = -6.648$, $p < .001$, and VM and AN, $t(240) = 4.768$, $p < .001$.

Table A3.2

Pure (New) - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	65.756	2	240	< .001
Set Size	-	43.607	1	240	< .001
Condition:Set Size	-	20.401	2	240	< .05

- *In both the VM and AN conditions, proportion of errors gets larger as set size increases, whereas the set-size functions in the CM condition are flat.*

We next constructed a GLME model including data from both Old and New test probes. We examined the effects of Condition (CM, VM, AN) and Set Size (3, 6, 9) on error proportions. We found a significant interaction between Condition and Set Size, $F(2, 510) = 32.313$, $p < .01$ (see Table A4.1). Post-hoc t-tests comparing the estimated Set Size slope for each condition against 0.0 revealed a non-significant difference in the CM (0.1 ± .1% per item) condition, $t(510) = 0.419$, $p = .675$; by contrast, the slopes for the VM (2.2 ± .1% per item), $t(510) = 11.714$, $p < .001$, and AN (1.0 ± .1% per item), $t(510) = 4.946$, $p < .001$, conditions were significantly different from 0.0.

Table A3.3

Pure - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	104.165	2	510	< .001
Set Size	-	97.240	1	510	< .001
Condition:Set Size	-	32.313	2	510	< .001

4. (Mixed) *Averaged across the old and new items, mean correct RTs are still shortest in the CM condition, intermediate in the AN condition, and longest in the VM condition.*

We ran a GLME model on mean correct RTs in the Mixed condition, with Condition (CM, VM, AN) and Set Size (3, 6, 9) as predictors. There was a main effect of Condition, $F(2, 3064) = 200.582$, $p < .001$ (see Table A4.1). Pairwise comparisons showed that responses in the CM condition (855 $\pm$ 21ms) were shorter than in VM (1008 $\pm$ 21ms), $t(3064) = -19.105$, $p < .001$, and AN (889 $\pm$ 21ms) conditions, $t(3064) = -4.284$, $p < .001$; the average response time in the VM condition was longer than in the AN condition, $t(3064) = 14.801$, $p < .001$.

Table A4.1

Mixed - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	200.582	2	3064	< .001
Set Size	-	76.923	1	3064	< .001
Condition:Set Size	-	8.118	2	3064	< .001

5. (Mixed) *For the error probabilities, there is little difference between the CM and AN items, but error probabilities are still greatest for the VM items.*

We ran a GLME model on the proportion of errors in the Mixed condition, with Condition (CM, VM, AN) and Set Size (3, 6, 9) as predictors. There was a significant main effect of Condition, $F(2, 3069) = 309.528$, $p < .001$ (see Table A5.1). Pairwise comparisons showed that there were fewer errors in the CM (5.8 $\pm$ .1%) than VM (18.0 $\pm$ .1%) conditions, $t(3069) = -21.636$, $p < .001$, while the CM and AN (5.9 $\pm$ .1%) conditions did not differ, $t(3069) = -0.168$, $p = .867$. Finally, the VM and AN conditions were significantly different from one another, $t(3069) = 21.452$, $p < .001$.

Table A5.1

Mixed - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	309.528	2	3069	< .001
Set Size	-	126.445	1	3069	< .001
Condition:Set Size	-	58.793	2	3069	< .001

6. (Mixed) *The VM items continue to show longer RTs and increased error probabilities with increases in set size; the AN target items also show these overall set-size effects.*

Using the same GLME as in (#4), we estimated the Set Size slope for each Condition on the mean correct response time. This revealed a significant interaction between Condition and Set Size, $F(2, 3064) = 8.118$, $p < .001$. Pairwise comparisons of the Set Size slope within each Condition revealed a significantly non-zero slope for the VM (17.7 $\pm$ 2.3 ms/item), $t(3064) = 7.631$, $p < .001$, and AN (12.9 $\pm$ 2.3 ms/item), $t(3064) = 5.557$, $p < .001$ conditions. There was a significant, albeit smaller slope in the CM condition (4.6 $\pm$ 2.3 ms/item), $t(3064) = 2.009$, $p < .05$.

Using the same GLME as in (#5), we estimated the Set Size slope for each Condition on the proportion of errors. This revealed a significant interaction between Condition and Set Size, $F(2, 3069) = 58.793$, $p < .001$. Pairwise comparisons of the Set Size slope within each Condition revealed a significantly non-zero slope for the VM (2.5 $\pm$ .2% per item), $t(3069) = 15.284$, $p < .001$, and AN (0.5 $\pm$ .2% per item), $t(3069) = 3.007$, $p < .001$ conditions. There was not a significant slope in the CM condition (0.2 $\pm$ .2% per item), $t(3069) = 1.181$, $p = .238$.

7. (Pure v. Mixed) *In the pure case, the old items showed big differences in overall RTs and error proportions across the CM, VM and AN conditions; but in the mixed case, the overall*

RTs and error proportions for the old items are similar in magnitude across the CM, VM, and AN conditions.

We performed an analysis of the Old items from the two experiments, considered together. We included Condition (CM, VM, AN) and Experiment (Pure, Mixed) as predictors, and conducted separate analyses for the mean correct response times and proportion of errors. There was a significant interaction between Experiment and Condition, $F(2, 1687) = 155.389$, $p < .001$ for both mean response times (see Table A7.1) and error proportions (see Table A7.2)

Table A7.1

Both Experiments (Old Items) - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	32.437	2	1687	< .001
Experiment	Mixed, Pure	0.013	1	211	< .001
Condition:Experiment	-	155.389	2	1687	< .001

Table A7.2

Both Experiments (Old Items) - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	12.063	2	1688	< .001
Experiment	Mixed, Pure	7.184	1	211	< .01
Condition:Experiment	-	5.985	2	1688	< .001

In the Pure conditions, pairwise comparisons showed a significant differences in mean response times between all comparisons of CM (687 $\pm$ 53ms), VM (1183 $\pm$ 53ms), and AN conditions (886 $\pm$ 53ms) (see Table A7.3). The same pattern of results held for the mean proportion of errors; CM (2.3 $\pm$ 2%), VM (10.1 $\pm$ 2%), AN (6.8 $\pm$ 2%) (see Table A7.4).

Table A7.3

Both Experiments (Old Items) - Correct Response Times - Pairwise Comparisons

Comparison	Mean Difference (SE)	t-Statistic	DF	p-value
Pure CM - Pure VM	-496.1 (25.8)	15.543	1687	< .001
Pure CM - Pure AN	-198.3 (25.8)	-17.771	1687	< .001
Pure VM - Pure AN	297.9 (25.8)	-19.234	1687	< .001

Table A7.4

Both Experiments (Old Items) - Error Proportions - Pairwise Comparisons

Comparison	Mean Difference (SE)	t-Statistic	DF	p-value
Pure CM - Pure VM	-0.078 (0.017)	-4.658	1688	< .001
Pure CM - Pure AN	-0.044 (0.017)	-2.641	1688	< .01
Pure VM - Pure AN	0.034 (0.017)	2.017	1688	< .001

In the Mixed conditions, pairwise comparisons showed no significant differences in mean response times between any of the CM (922 ± 22ms), VM (930 ± 22ms), or AN (922 ± 22ms) conditions (see Table A7.5); error rates revealed a significant difference between the CM (9.5 ± .1%) and VM (11.4 ± .1%) condition, as well as a difference between VM and the and AN (9.0 ± .1%) conditions; CM and AN did not differ from each other(see Table A7.6).

Table A7.5

Both Experiments (Old Items) - Correct Response Times - Pairwise Comparisons

Comparison	Mean Difference (SE)	t-Statistic	DF	p-value
Mixed CM - Mixed VM	2.8 (6.1)	0.457	1687	= .647
Mixed CM - Mixed AN	-5.26 (6.1)	-0.868	1687	= .385
Mixed VM - Mixed AN	-8.04 (10.5)	-0.765	1687	= .444

Table A7.6

Both Experiments (Old Items) - Error Proportions - Pairwise Comparisons

Comparison	Mean Difference (SE)	t-Statistic	DF	p-value
Mixed CM - Mixed VM	-0.02 (0.007)	-2.888	1688	< .01
Mixed CM - Mixed AN	0.005 (0.007)	0.665	1688	= .506
Mixed VM - Mixed AN	0.024 (0.007)	3.552	1688	< .001

8. (Mixed) *Notably, in the mixed condition, the overall error proportions and RTs for the CM-old items are now nearly the same as for the VM-old items.*

We constructed a GLME to examine the patterns for the Old test probes within the Mixed Experiment; we entered Condition (CM, VM, AN) into two separate analyses, one for the mean correct response time and one for the proportion of errors.

There was not a significant main effect of Condition on the mean correct response time, $F(2, 1447) = 0.492$, $p = .611$; pairwise comparisons demonstrated that response times in the CM condition (923 $\pm$ 22ms) were not significantly different from those in the VM condition (931 $\pm$ 22ms). However, there was a significant main effect of Condition on the proportion of errors, $F(2, 1448) = 6.438$, $p < .01$; pairwise comparisons revealed a significant difference between the CM (9.5 $\pm$ .1%) and VM (11.4 $\pm$ .1%) conditions.

9. (Comparison) *Another interesting difference is that error proportions for the AN-new items are reduced in the mixed condition compared to the pure condition.*

We constructed a GLME to compare the New items between the Mixed and Pure experiments; we entered Condition (CM, VM, AN) and Experiment (Pure, Mixed) as predictors. We found a significant interaction between Condition and Experiment, $F(2, 1685) = 38.441$, $p < .001$; pairwise comparisons revealed fewer errors in the Mixed (2.7 $\pm$ .1%) than in the Pure (7.4 $\pm$ 1.5%) experiment, $t(761) = -2.874$, $p < .01$.

10. (Mixed, Basic IR Model) *The set-size functions for the AN targets exhibit a "cross-over" with respect to the set-size functions for the CM and VM targets.*

We constructed a GLME for mean correct response times for the Old test probes in the Mixed experiment, entering Condition (CM, VM, AN) and Set Size as predictors. There was a significant interaction between Condition and Set Size, $F(2, 1447) = 10.703$, $p < .001$ (see Table A10.1). Examination of the Set Size slope within each condition revealed a slope which did not differ from zero in the CM condition (3.6 ms $\pm$ 2.6ms/item), $t(1447) = 1.382$, $p = .167$; by

contrast, the slopes were significantly different in both in the VM (8.47 ± 2.6ms/item), $t(1447) = 3.273$, $p < .01$, and AN (20.07 ± 2.6ms/item), $t(1447) = 7.746$, $p < .001$.

Table A10.1

Mixed (Old) - Correct Response Times - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	0.516	2	1447	= .597
Set Size	-	51.233	1	1447	< .001
Condition:Set Size	-	10.703	2	1447	< .001

We utilized this same approach to analyze the proportion of errors as well. Again, there was a significant interaction between Condition and Set Size, $F(2, 1448) = 3.170$ $p < .05$ (see Table A10.2). Examining the Set Size slopes within each condition revealed a flat slope in the CM condition (0.1 ± .1% per item), $t(1448) = 0.682$, $p = .495$, but significant slopes in the VM (0.7 ± .1% per item), $t(1448) = 3.343$, $p < .001$, and AN conditions (0.8 ± .1% per item), $t(1448) = 4.062$, $p < .001$.

Table A10.2

Mixed (Old) - Error Proportions - Summary ANOVA

Term	Levels	F-Statistic	DF1	DF2	p-value
Condition	CM, VM, AN	6.563	2	1448	< .01
Set Size	-	21.803	1	1448	< .001
Condition:Set Size	-	3.170	2	1448	< .05

Footnotes

1. We computed overall proportion correct and mean-correct RT for each participant on the CM, VM, and AN items, averaged across all memory set sizes and whether a test probe was a target or foil. Participants were defined as outliers if their overall proportion correct was more than three standard deviations below the mean in any of the CM, VM, or AN conditions; or if their mean correct RT was greater than three standard deviations above the mean in any of the CM, VM, or AN conditions. These computations were performed separately across the pure and mixed experiments.

2. We emphasize that the form of the memory-strength function likely varies across different cognitive tasks. For example, participants may rely on different strategies in free-recall and serial-recall tasks than in probe-recognition tasks, altering the form of the memory-strength gradient as well the mechanisms that underlie performance (e.g., Duncan & Murdock, 2000; Osth & Farrell, 2019).

3. Although not made explicit in Equation 5, the LTM activations should be conceptualized in terms of two components: a hard-wired LTM strength, and a cognitive weighting of that strength. For example, in VM, a test item might have high LTM strength due to its frequent occurrence in memory sets from previous trials; however, if the observer attempts to focus attention on the current list, then the weight given to LTM would be low.

4. Summarizing, the degree to which AN test probe t_i activates an exemplar with lag j in the memory set (e_j) is given by $a_{ij} = m_j$, if $t_i = e_j$; $a_{ij} = m_j s$, if $t_i \neq e_j$; where $m_j = boost^*(j^{-\beta} + \alpha)$; and where the parameters β, α, and *s* are estimated separately for the AN items than from the CM and VM items.

5. An additional technical detail is that, following Nosofsky et al. (2011), we made allowance for a *primacy* parameter in fitting the pure-conditions data. In this extension, the memory strength for the item in the first serial position of the memory set is multiplied by *primacy*, allowing the model to account for systematic decreases in RT and increases in accuracy for this item. These residual primacy effects are often observed in pure-list designs; for unexplained reasons, the residual primacy effect did not seem to arise in our mixed-list design.

6. Although we have raised the issue of the extent to which old-item-response learning occurs in the present mixed-list design, essentially the same issue arises for pure-list designs. Much of the past evidence for item-response learning under pure CM conditions pertains to the CM foils, not the targets. The extremely good performance associated with CM targets in pure-list designs could arise, for example, simply because the targets are far more familiar than are the foils. This would allow participants to set their drift-rate criterion in a location that is highly effective for responding to both target and foil test probes (see Figure 2).

Table 1. Relative frequency of trials in which individual items from the CM and VM classes are presented in study and test roles in both the mixed and pure conditions.

	Memory-Set Item	Old Test Probe	New Test Probe	Total Test Probe
CM-target	2/3	1/18	0	1/18
CM-foil	0	0	1/18	1/18
VM	1/3	1/36	1/36	1/18

CM=Consistent Mapping; VM=Varied Mapping.

Table 2. Weighted Sum of Squared Deviation (WSSD) Fits of the Models to the Mixed-Conditions and Pure-Conditions Data.

Mixed Conditions				
Model	**CM**	**VM**	**AN**	**Total**
Familiarity-Only	0.086	0.173	0.115	0.374
Baseline IR	0.047	0.134	0.103	0.283
Extended IR	0.030	0.026	0.046	0.101

Pure Conditions				
Model	**CM**	**VM**	**AN**	**Total**
Baseline IR	0.041	0.112	0.057	0.210
Criterion-Constrained IR	0.058	0.158	0.173	0.389

Table 3. Best-fitting parameters for the familiarity-only model as applied to the mixed-condition data.

	Item Type		
Parameter	**CM**	**VM**	**AN**
β	1.357	--	--
α	0.437	--	--
s	0.020	--	--
c	0.353	--	--
OLD-crit	2.636	--	--
NEW-crit	2.709	--	--
t_0	356.0	--	--
κ	108.0	--	--
AN-OLD			0.000
VM-OLD		0.137	
CM-OLD(old)	0.104		
CM-OLD(new)	0.000		

Note. Cells with dashes denote cases in which parameter values were held fixed at those values that appear in columns to the left. Cells enclosed by brackets denotes cases in which a parameter was set at a default value. t_0 and κ are measured in msec.

The LTM_{Old} parameters in each condition are denoted *AN-OLD, VM-OLD, CM-OLD(old) and CM-OLD(new).* In this notation, for example, *VM-OLD* is the LTM_{Old} value associated with VM test probes. Because old and new VM and AN test probes are logically equivalent in terms of their LTM familiarity, a single parameter estimate applies to these types of probes. Separate parameter estimates are applied to CM-old versus CM-new test probes, however, because those item types are logically distinct in terms of their LTM familiarity. By definition, all LTM_{New} values are held fixed at zero for the familiarity-only model.

Table 4. Best-fitting parameters for the baseline item-response learning model as applied to the mixed-condition data.

	Item Type		
Parameter	**CM**	**VM**	**AN**
β	1.399	--	--
α	0.408	--	--
s	0.021	--	--
c	0.346	--	--
OLD-crit	2.546	--	--
NEW-crit	2.774	--	--
t_0	391	--	--
κ	104	--	--
AN-OLD			0.000
AN-NEW			0.000
VM-OLD		0.142	
VM-NEW		0.039	
CM-OLD	0.142		
CM-NEW	0.242		

Note. Cells with dashes denote cases in which parameter values were held fixed at those values that appear in columns to the left. Cells enclosed by brackets denotes cases in which a parameter was set at a default value. t_0 and κ are measured in msec.

The LTM_{Old} and LTM_{New} parameters in each condition are denoted *AN-OLD, AN-NEW, VM-OLD, VM-NEW, CM-OLD and CM-NEW.* In this notation, for example, *VM-OLD* is the LTM_{Old} value associated with VM test probes.

Table 5. Best-fitting parameters for the extended item-response learning model as applied to the mixed-condition data.

	Item Type		
Parameter	**CM**	**VM**	**AN**
Boost			1.158
β	1.228	--	1.511
α	0.276	--	0.478
s	0.070	--	0.010
c	0.587	--	--
OLD-crit	3.485	--	--
NEW-crit	4.479	--	--
t_0	519.0	--	--
κ	38.0	--	--
AN-OLD			0.143
AN-NEW			[0.000]
VM-OLD		0.143	
VM-NEW		[0.000]	
CM-OLD	0.203		
CM-NEW	0.623		

Note. Cells with dashes denote cases in which parameter values were held fixed at those values that appear in columns to the left. Cells enclosed by brackets denotes cases in which a parameter was set at a default value. t_0 and κ are measured in msec.

The LTM_{Old} and LTM_{New} parameters in each condition are denoted *AN-OLD, AN-NEW, VM-OLD, VM-NEW, CM-OLD and CM-NEW.* In this notation, for example, *VM-OLD* is the LTM_{Old} value associated with VM test probes.

Table 6. Best-fitting parameters for the item-response learning model applied to the pure-conditions data.

	Item Type		
Parameter	**CM**	**VM**	**AN**
primacy	1.189		
β	0.854	--	--
α	0.299	--	--
s	0.036	--	--
c	0.477	0.845	0.569
OLD-crit	2.856	4.968	3.774
NEW-crit	3.768	5.345	4.693
t_0	450.0	--	--
κ	[38.0]	--	--
AN-OLD			0.105
AN-NEW			[0.000]
VM-OLD		0.405	
VM-NEW		[0.000]	
CM-OLD	0.405		
CM-NEW	18.240		

Note. Cells with dashes denote cases in which parameter values were held fixed at those values that appear in columns to the left. Cells enclosed by brackets denotes cases in which a parameter was set at a default value. t_0 and κ are measured in msec.

The LTM_{Old} and LTM_{New} parameters in each condition are denoted *AN-OLD, AN-NEW, VM-OLD, VM-NEW, CM-OLD and CM-NEW.* In this notation, for example, *VM-OLD* is the LTM_{Old} value associated with VM test probes.

Figure Captions

1. Schematic sketch of a random-walk evidence-accumulation process.

2. Schematic sketch of familiarity distributions for VM versus CM, along with alternative effective placements of the criterion c for portioning the old and new item distributions.

3. Mean correct response times (msec) in the pure and mixed experiments, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9).

4. Mean probability of errors in the pure and mixed experiments, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9).

5. Mean correct response times (msec) for the old test probes in the pure and mixed experiments, plotted as a joint function of condition (CM, VM, AN), lag, and set size.

6. Mean probability of errors for the old test probes in the pure and mixed experiments, plotted as a joint function of condition (CM, VM, AN), lag, and set size.

7. Schematic illustration of the application of the exemplar-based-random-walk model to the short-term probe-recognition task. Note: O_k is the old item on the current study list that is presented in serial-position k.

8. Right panel: Predictions from the extended item-response learning model of the mean probability of errors in the mixed-list experiment, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9). Left panel: the observed data are re-plotted for ease of comparison.

9. Right panel: Predictions from the extended item-response learning model of the mean correct response times (msec) in the mixed-list experiment, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9). Left panel: the observed data are re-plotted for ease of comparison.

10. Right panel: Predictions from the extended item-response learning model of the mean probability of errors for old test probes in the mixed-list experiment, plotted as a joint function of condition (CM, VM, AN), lag, and set size. Left panel: the observed data are re-plotted for ease of comparison.

11. Right panel: Predictions from the extended item-response learning model of the mean correct response times (msec) for old test probes in the mixed-list experiment, plotted as a joint function of condition (CM, VM, AN), lag, and set size. Left panel: the observed data are re-plotted for ease of comparison.

12. Right panel: Predictions from the basic item-response learning model of the mean probability of errors in the pure-lists experiment, plotted as a joint function of condition (CM,

VM, AN), probe type (old, new) and memory set size (3, 6, 9). Left panel: the observed data are re-plotted for ease of comparison.

13. Right panel: Predictions from the basic item-response learning model of the mean correct response times (msec) in the pure-lists experiment, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9). Left panel: the observed data are re-plotted for ease of comparison.

14. Right panel: Predictions from the basic item-response learning model of the mean probability of errors for old test probes in the pure-lists experiment, plotted as a joint function of condition (CM, VM, AN), lag, and set size. Left panel: the observed data are re-plotted for ease of comparison.

15. Right panel: Predictions from the basic item-response learning model of the mean correct response times (msec) for old test probes in the pure-lists experiment, plotted as a joint function of condition (CM, VM, AN), lag, and set size. Left panel: the observed data are re-plotted for ease of comparison.

S1. Top panels: Mean probability of errors in the mixed experiment, shown separately for the cohort-1 (left panel) and cohort-2 (right panel) participants, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9). Bottom panels: the analogous mean correct response-time data.

S2. Right panels: Predictions from the familiarity-only model of the mean probability of errors (top) and mean correct response times (bottom) in the mixed-list experiment, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9). Left panels: the observed data are re-plotted for ease of comparison.

S3. Right panels: Predictions from the basic item-response learning model of the mean probability of errors (top) and mean correct response times (bottom) in the mixed-list experiment, plotted as a joint function of condition (CM, VM, AN), probe type (old, new) and memory set size (3, 6, 9). Left panels: the observed data are re-plotted for ease of comparison.

Figure 1.

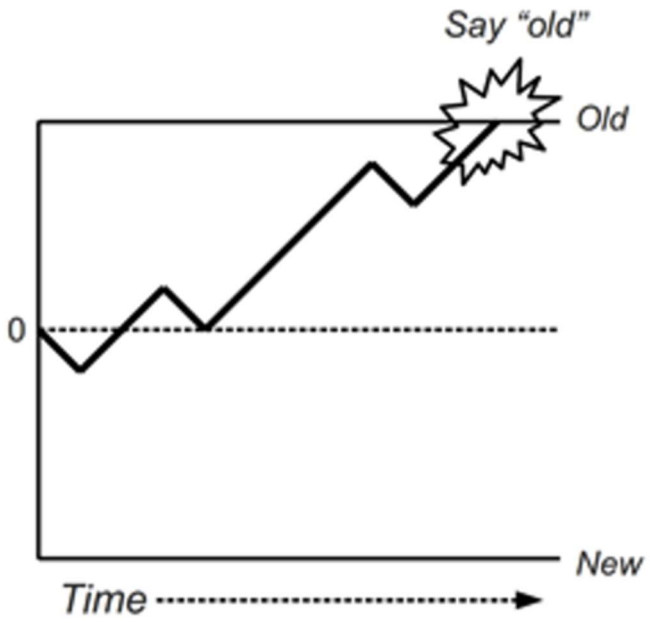


Figure 2.

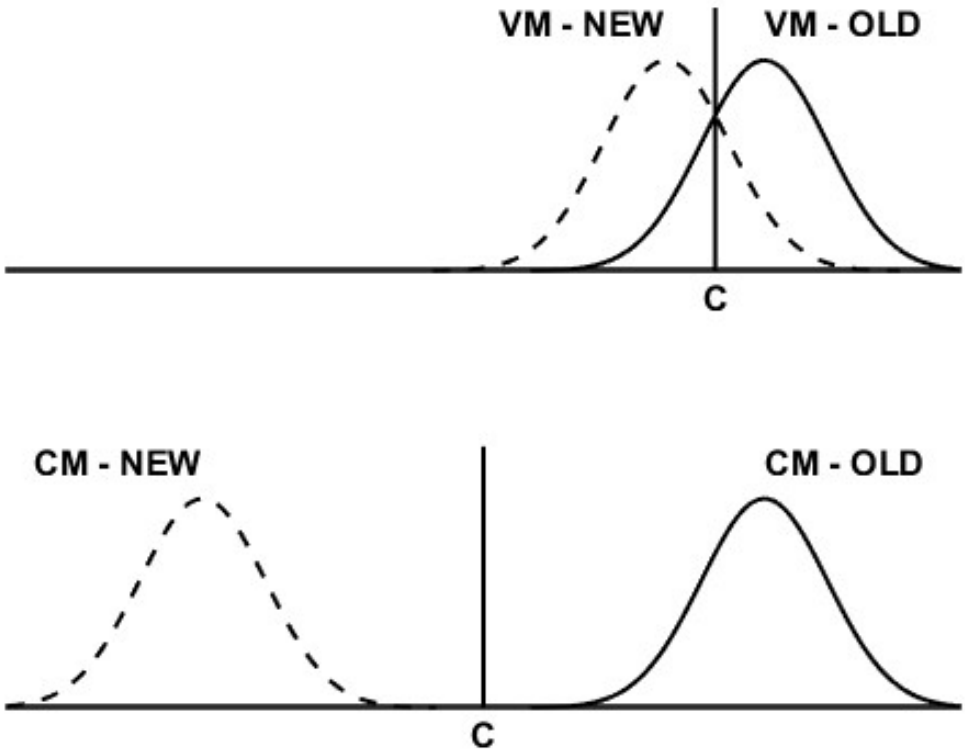


Figure 3.

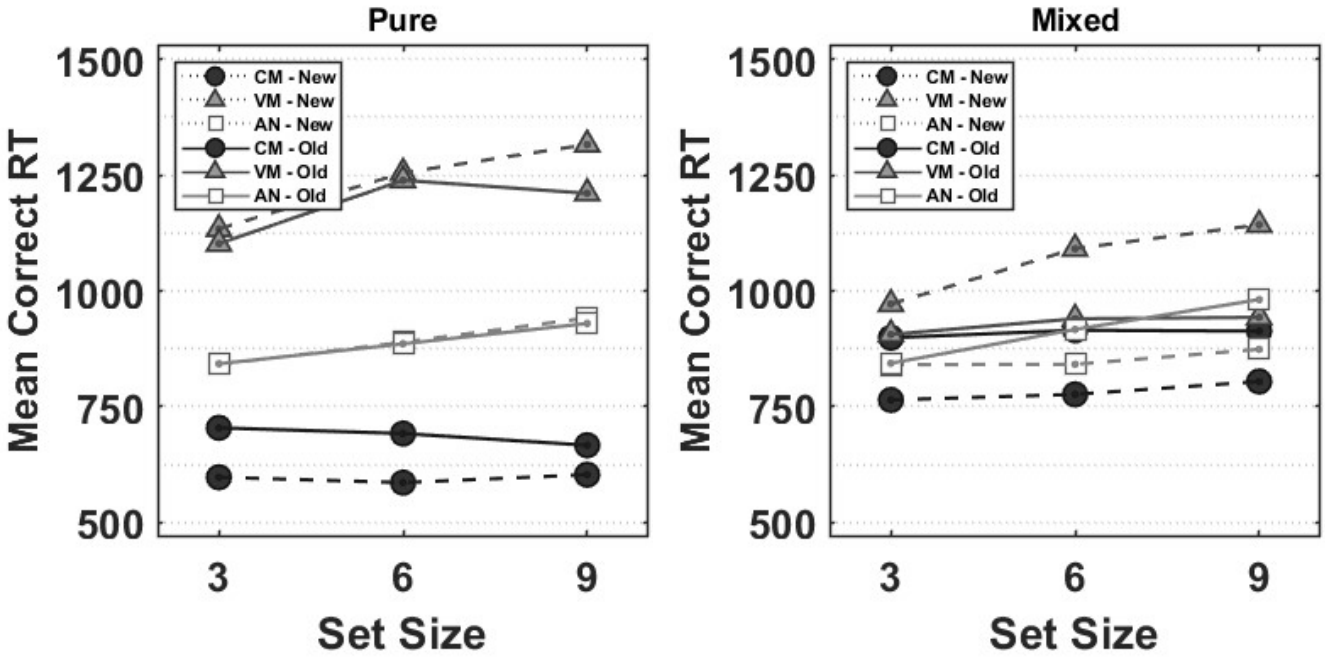

Pure
Mixed
Mean Correct RT
1500
1250
1000
750
500
3
6
9
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old

Figure 4.

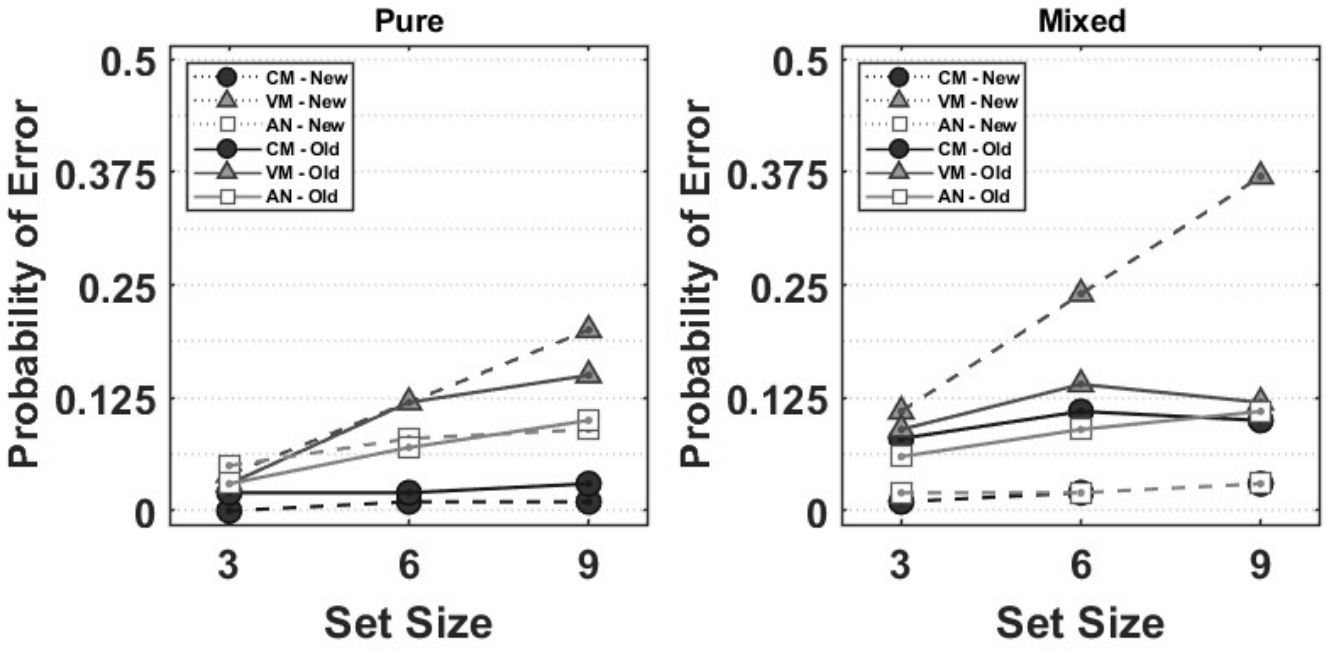

Pure
Mixed
Probability of Error
0.5
0.375
0.25
0.125
0
3
6
9
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old

Figure 5.

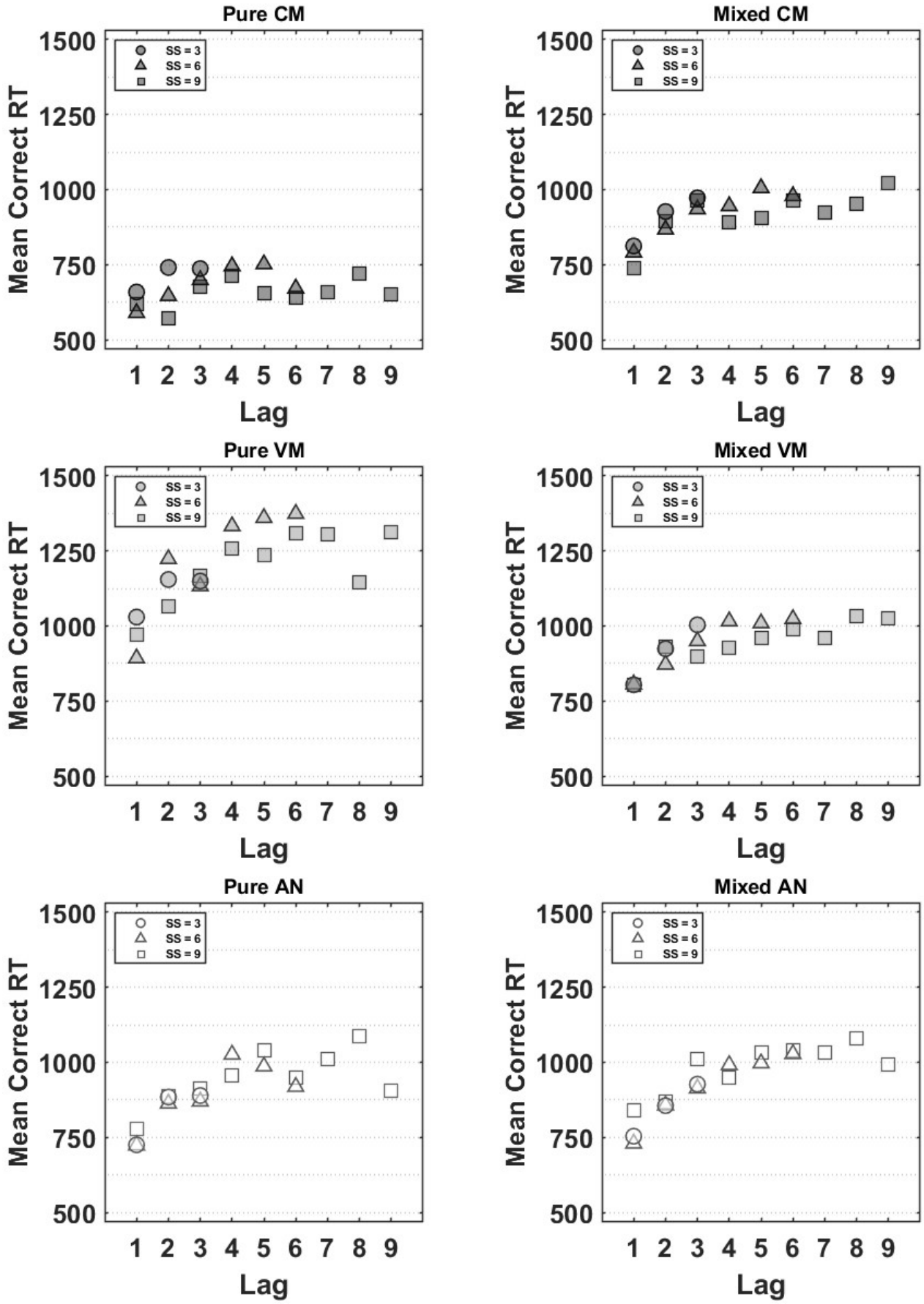

Pure CM
Mixed CM
Pure VM
Mixed VM
Pure AN
Mixed AN
SS = 3
SS = 6
SS = 9
Mean Correct RT
Lag
1500
1250
1000
750
500
1 2 3 4 5 6 7 8 9

Figure 6.

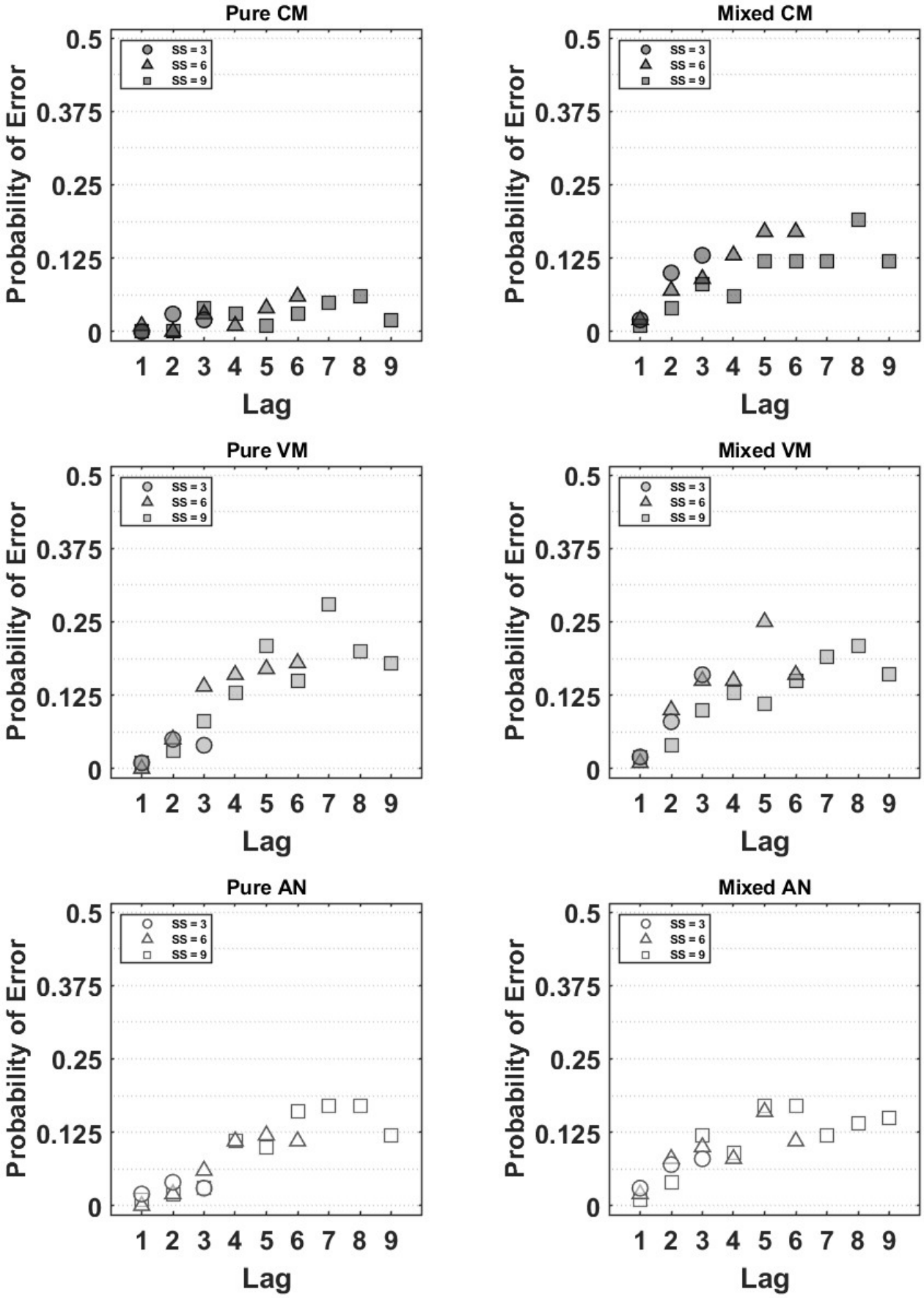

Pure CM
Mixed CM
Pure VM
Mixed VM
Pure AN
Mixed AN
SS = 3
SS = 6
SS = 9
Probability of Error
0
0.125
0.25
0.375
0.5
1 2 3 4 5 6 7 8 9
Lag

Figure 7.

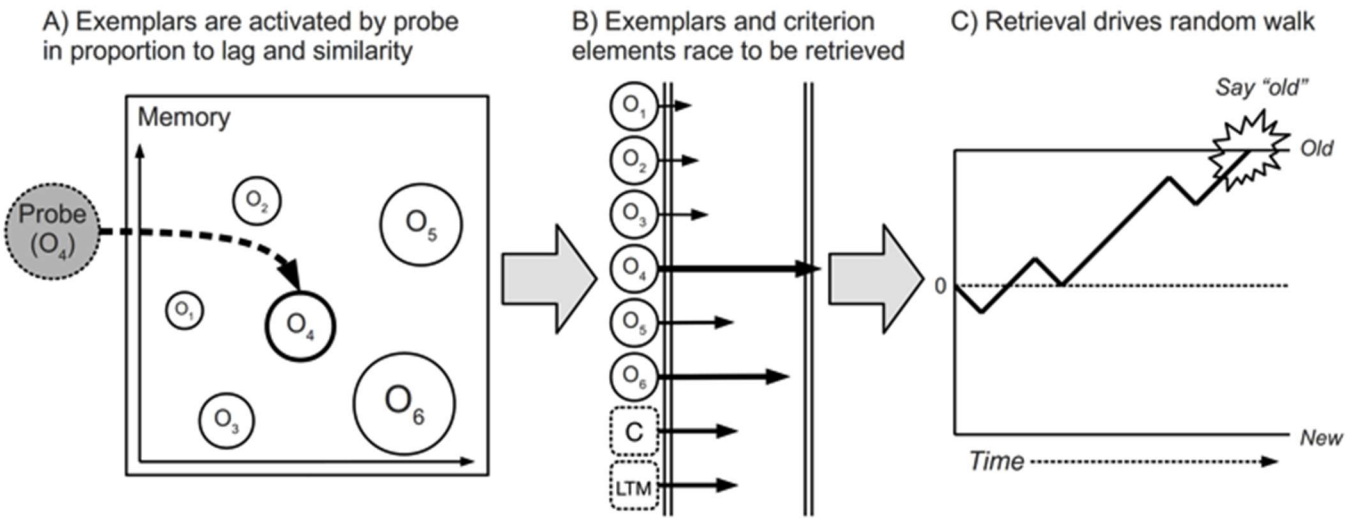

A) Exemplars are activated by probe in proportion to lag and similarity
B) Exemplars and criterion elements race to be retrieved
C) Retrieval drives random walk
Memory
Probe (O_4)
O_1
O_2
O_3
O_4
O_5
O_6
C
LTM
Say "old"
Old
0
New
Time

Figure 8.

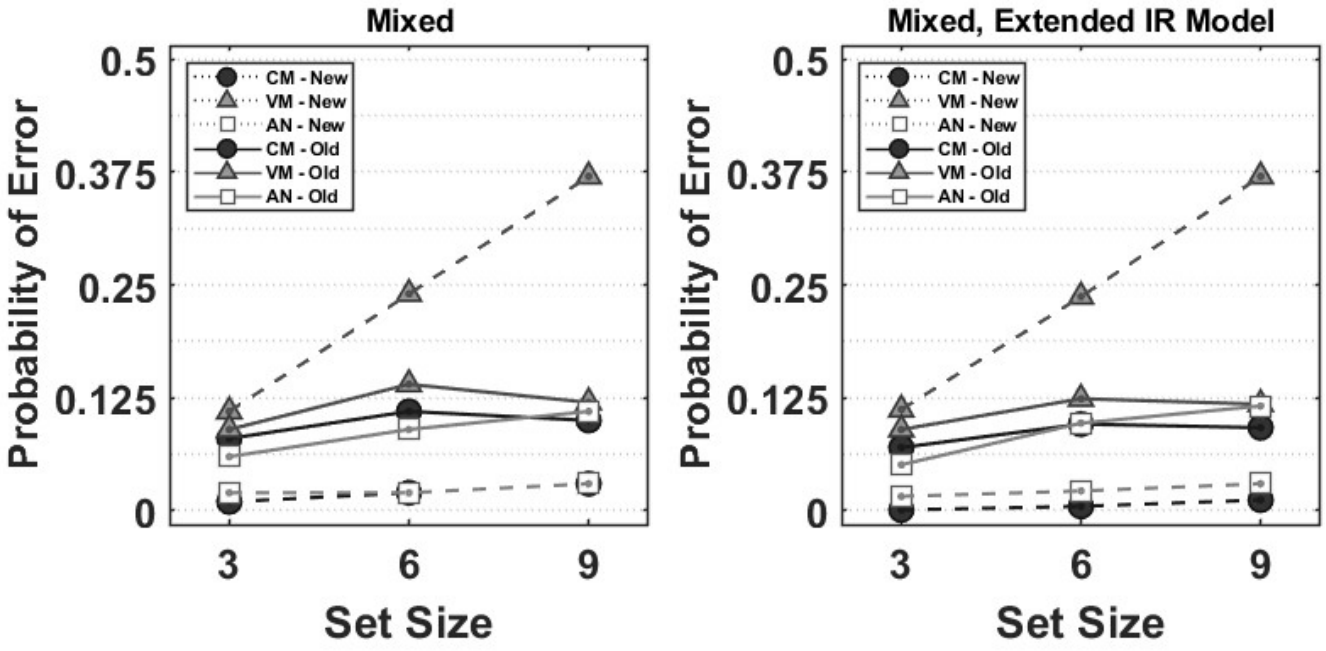

Mixed
Mixed, Extended IR Model
Probability of Error
0.5
0.375
0.25
0.125
0
3
6
9
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old

Figure 9.

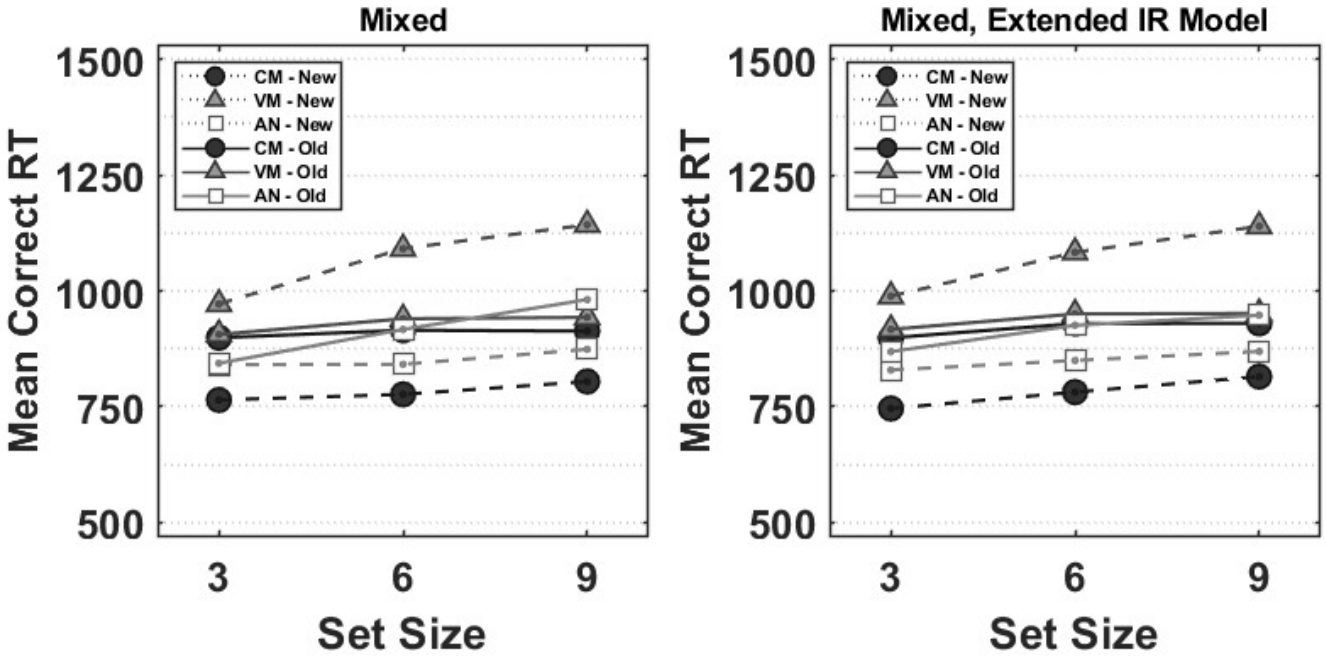

Mixed
Mixed, Extended IR Model
Mean Correct RT
Set Size
1500
1250
1000
750
500
3
6
9
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old

Figure 10.

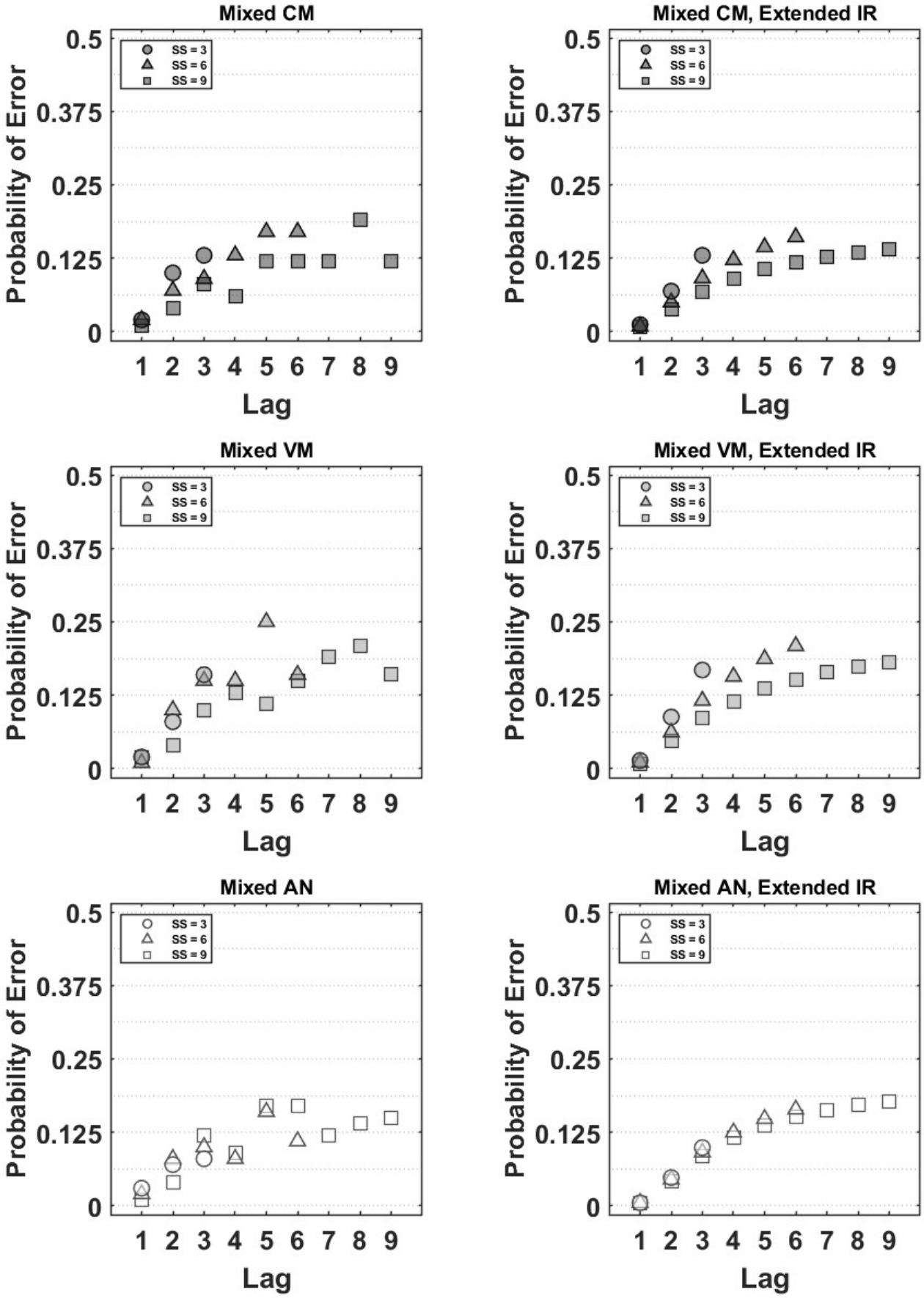

Mixed CM
Mixed CM, Extended IR
Mixed VM
Mixed VM, Extended IR
Mixed AN
Mixed AN, Extended IR
Probability of Error
Lag
SS = 3
SS = 6
SS = 9
0
0.125
0.25
0.375
0.5
1 2 3 4 5 6 7 8 9

Figure 11.

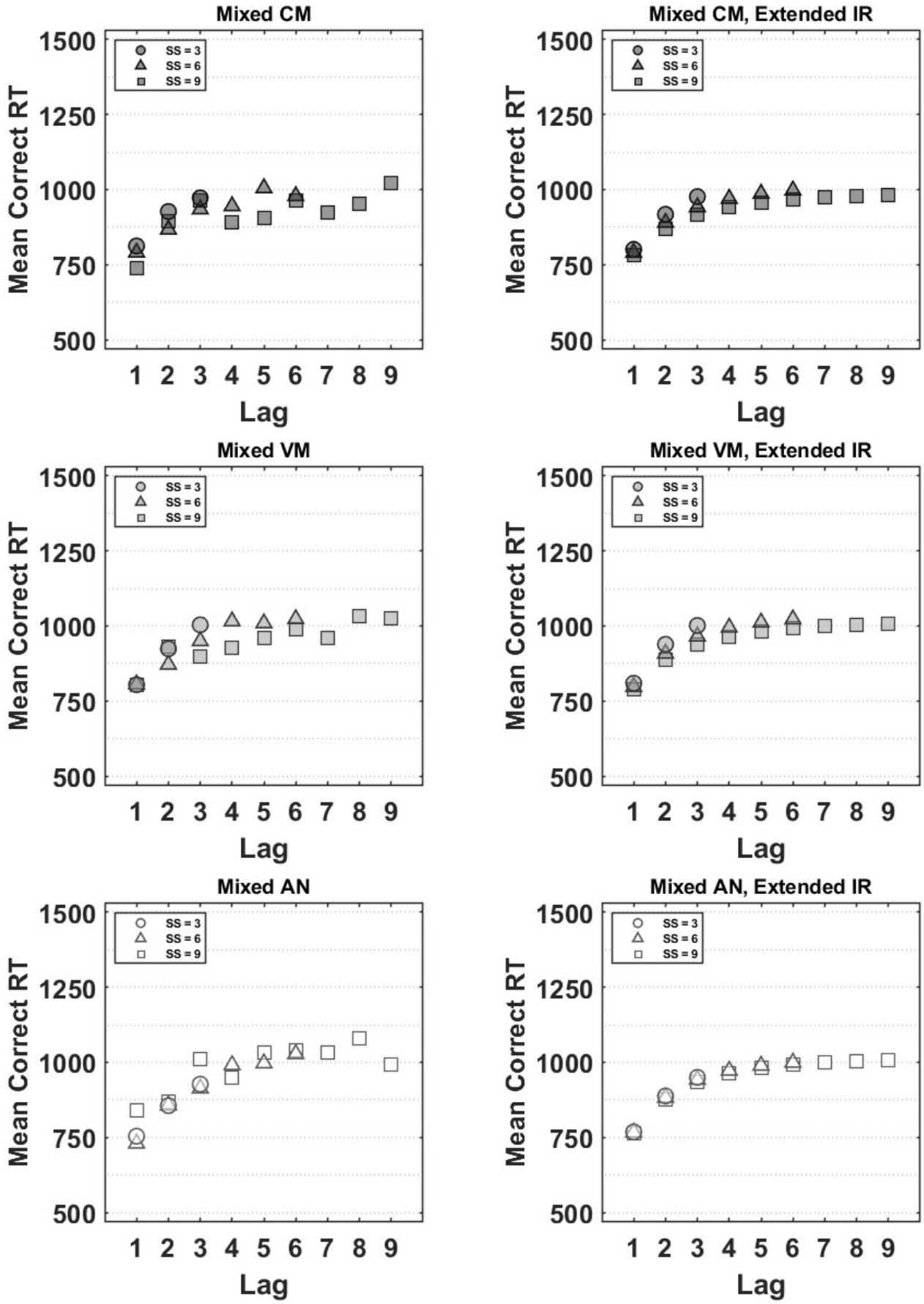

Mixed CM
Mixed CM, Extended IR
Mixed VM
Mixed VM, Extended IR
Mixed AN
Mixed AN, Extended IR
SS = 3
SS = 6
SS = 9
Mean Correct RT
1500
1250
1000
750
500
1 2 3 4 5 6 7 8 9
Lag

Figure 12.

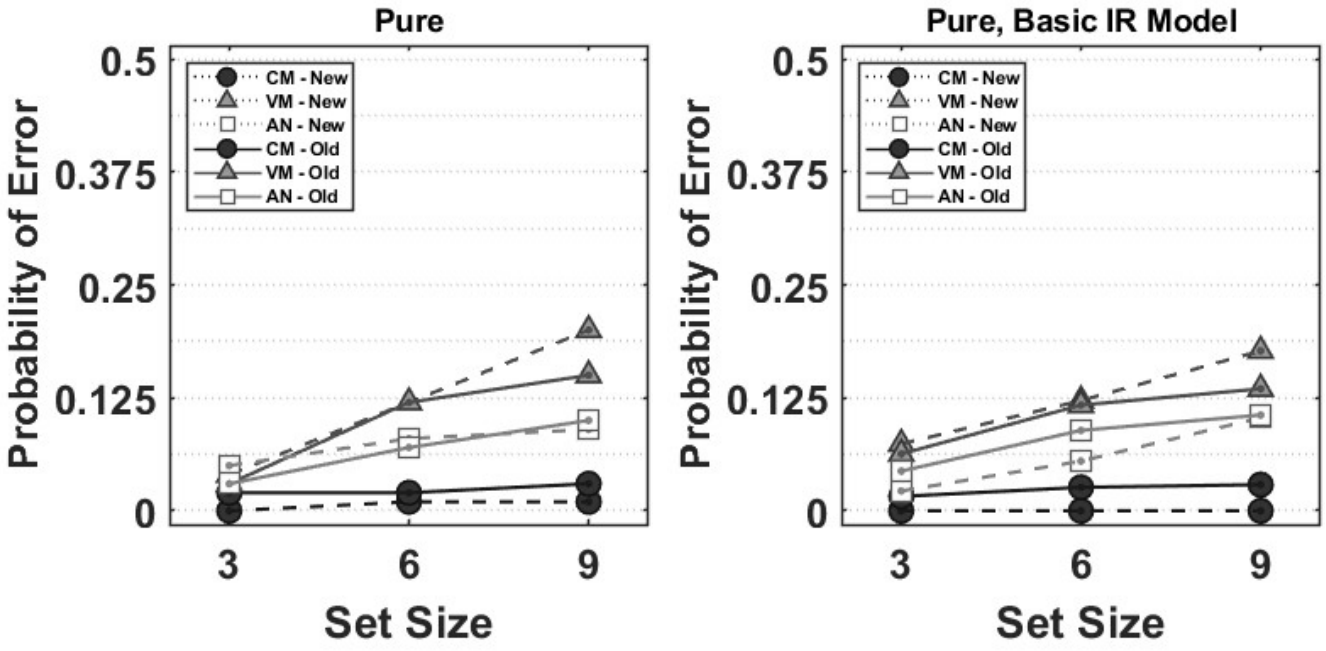

Pure
Pure, Basic IR Model
Probability of Error
0.5
0.375
0.25
0.125
0
3
6
9
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old

Figure 13.

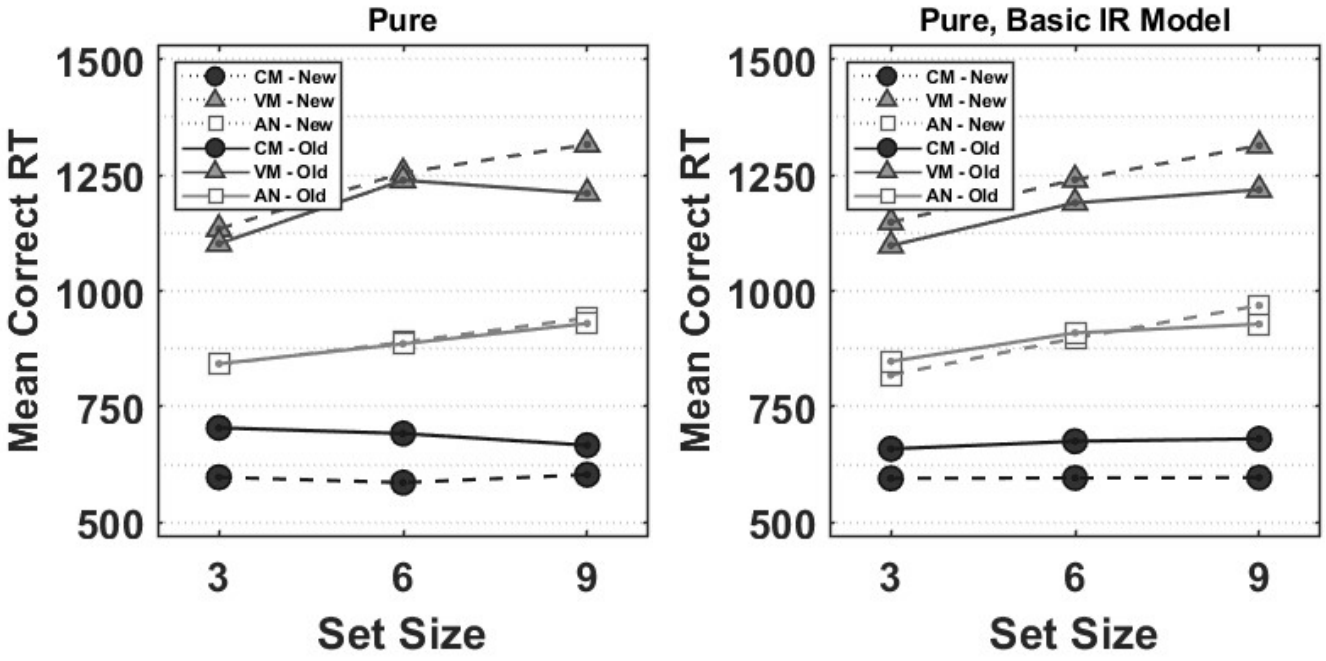

Pure
Pure, Basic IR Model
Mean Correct RT
1500
1250
1000
750
500
3
6
9
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old

Figure 14.

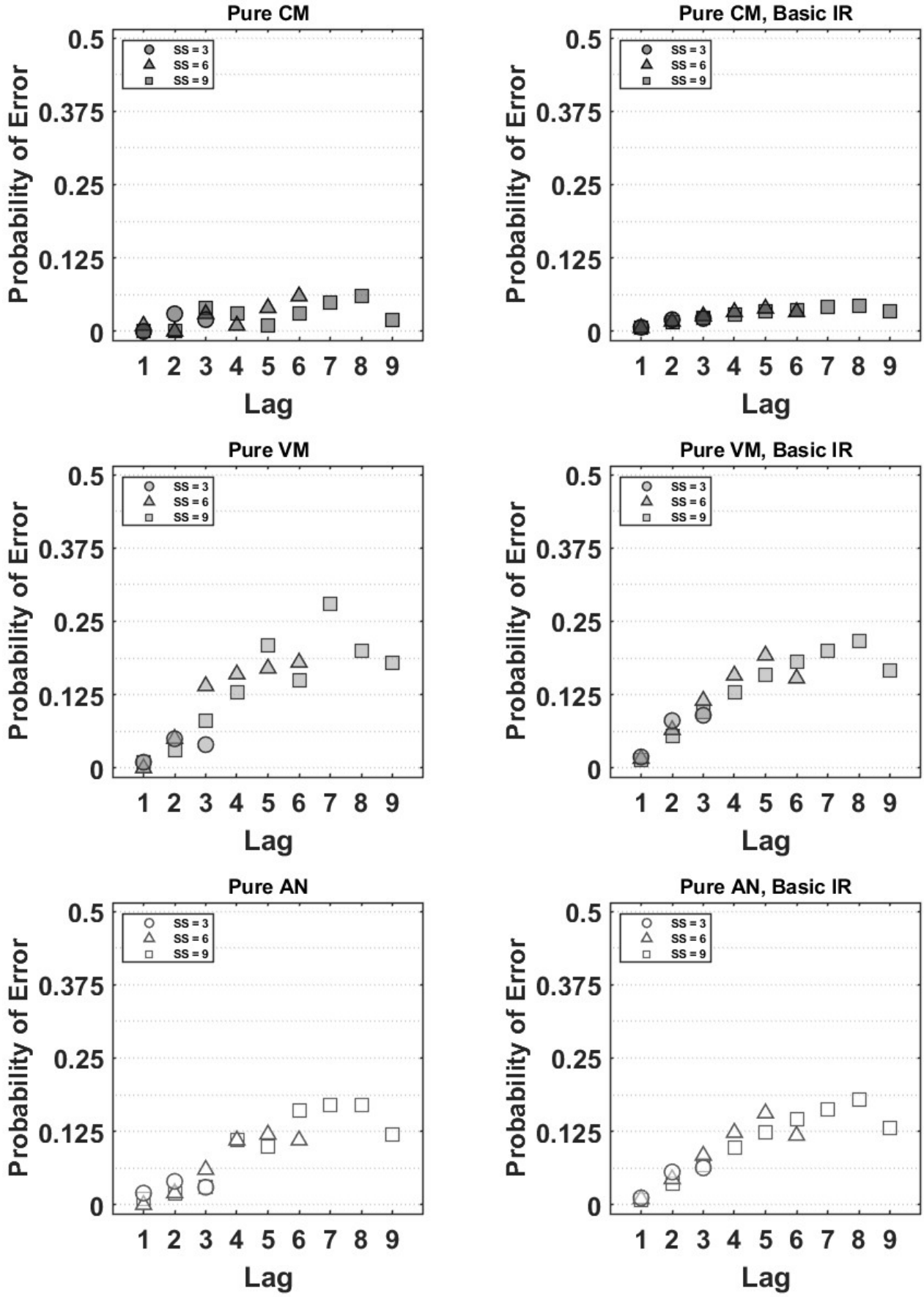
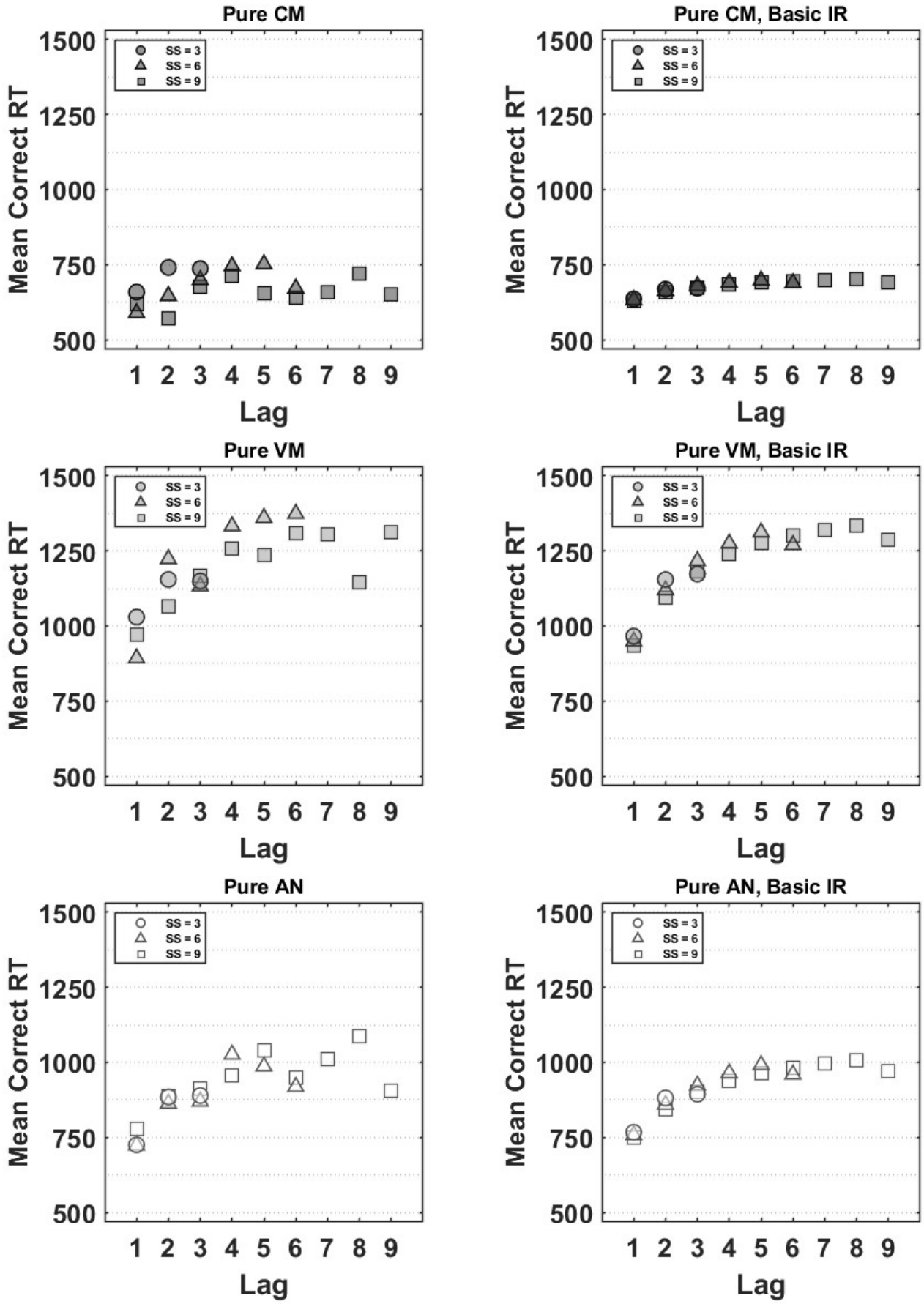

Pure CM
Pure CM, Basic IR
Pure VM
Pure VM, Basic IR
Pure AN
Pure AN, Basic IR
SS = 3
SS = 6
SS = 9
Probability of Error
0.5
0.375
0.25
0.125
0
1 2 3 4 5 6 7 8 9
Lag

Figure 15.


Pure CM
Pure CM, Basic IR
Pure VM
Pure VM, Basic IR
Pure AN
Pure AN, Basic IR
SS = 3
SS = 6
SS = 9
Mean Correct RT
1500
1250
1000
750
500
1 2 3 4 5 6 7 8 9
Lag

Supplementary Materials

Figure S1.

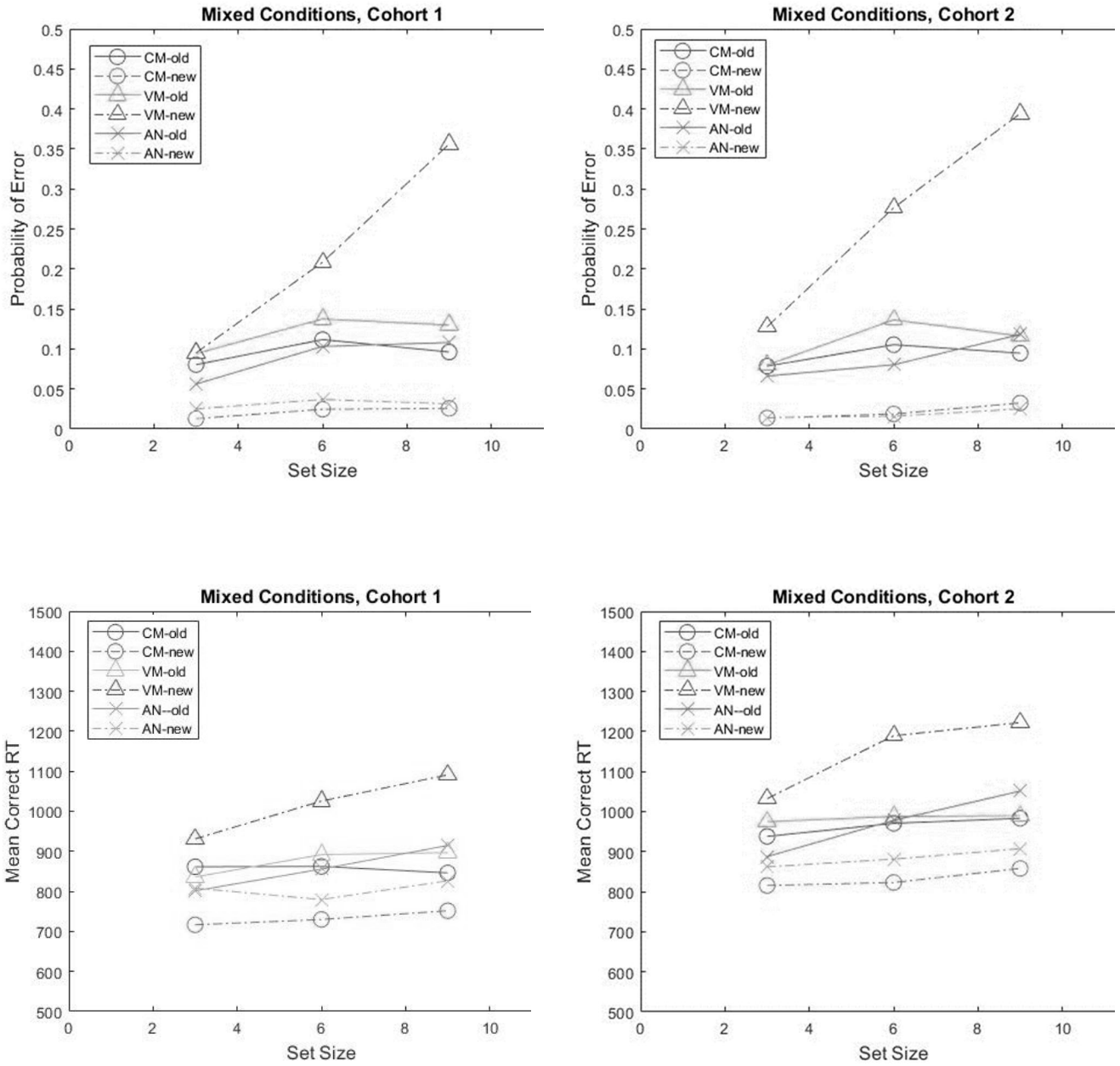


Figure S2.

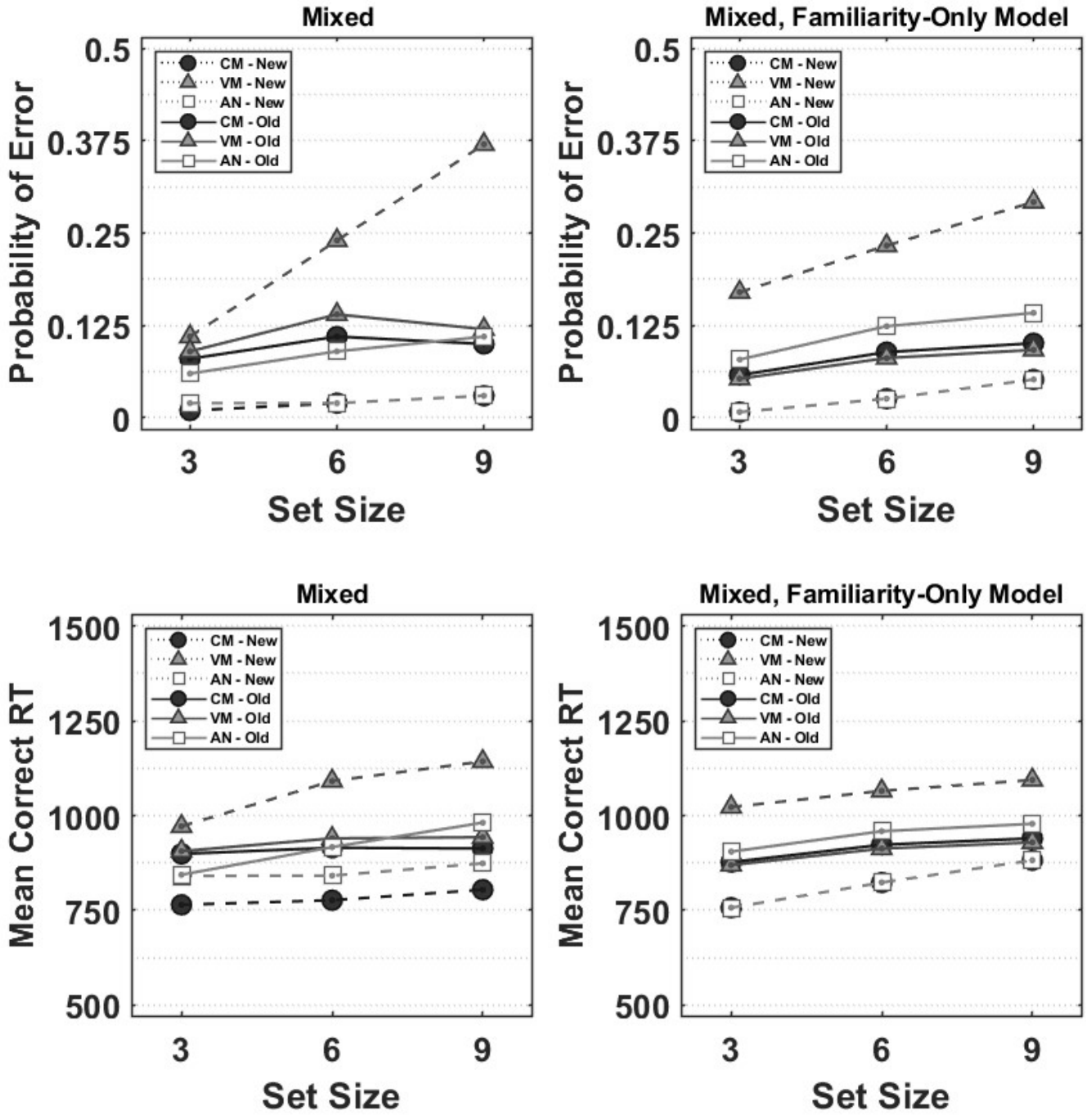
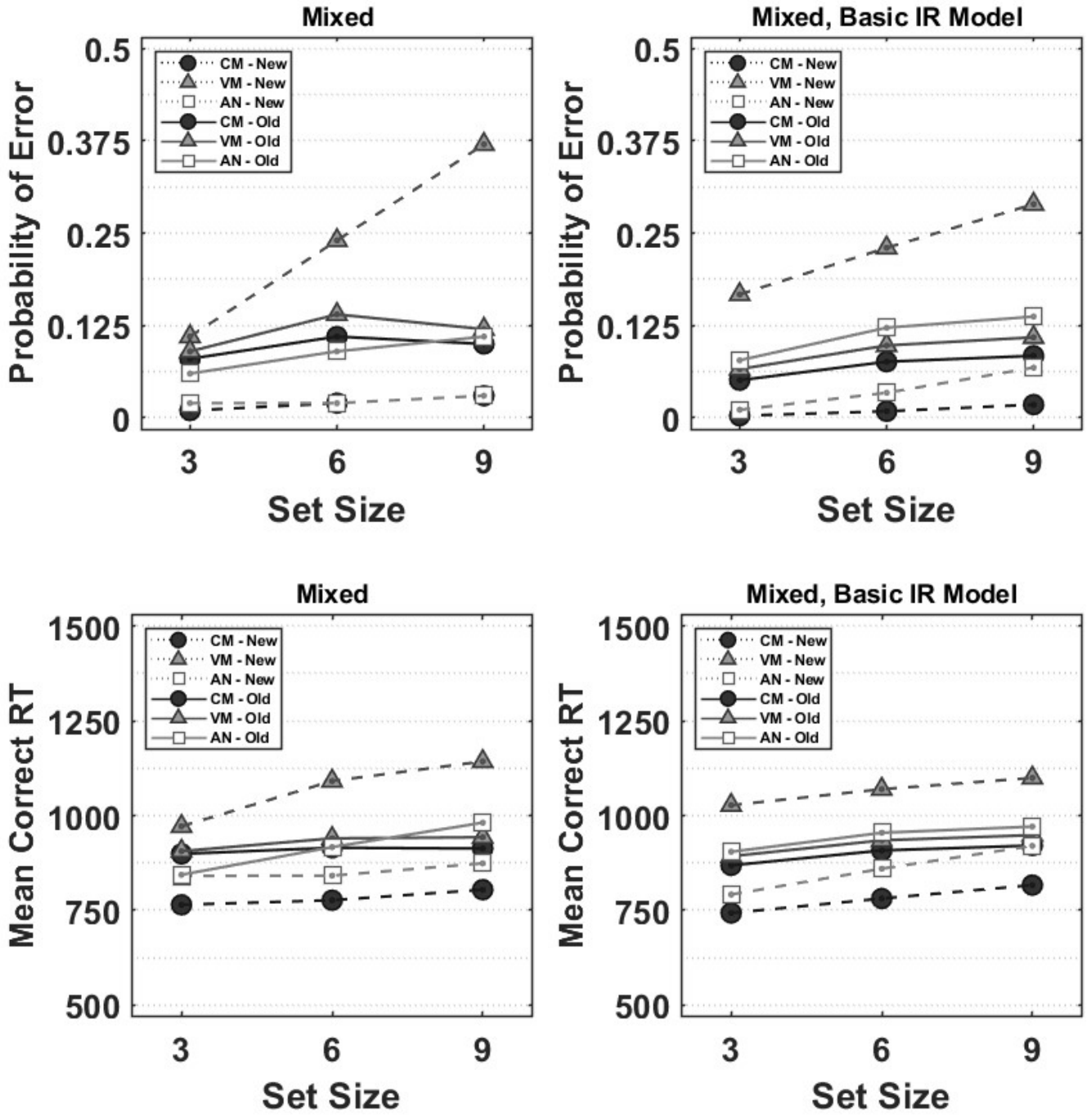

Mixed
Mixed, Familiarity-Only Model
Probability of Error
Mean Correct RT
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old
0.5
0.375
0.25
0.125
0
1500
1250
1000
750
500
3
6
9

Figure S3.


Mixed
Mixed, Basic IR Model
Probability of Error
Mean Correct RT
Set Size
CM - New
VM - New
AN - New
CM - Old
VM - Old
AN - Old
0.5
0.375
0.25
0.125
0
1500
1250
1000
750
500
3
6
9