THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2014
Vol. 67, No. 2, 394–416, http://dx.doi.org/10.1080/17470218.2013.816750

Retrieval dynamics in self-terminated memory search

Erika K. Hussey[1], Michael R. Dougherty[1], J. Isaiah Harbison[1], and Eddy J. Davelaar[2]

[1]Department of Psychology, Program in Neuroscience and Cognitive Science, University of Maryland, College Park, MD, USA
[2]Department of Psychological Sciences, Birkbeck College, University of London, London, UK

Most free-recall experiments employ a paradigm in which participants are given a preset amount of time to retrieve items from a list. While much has been learned using this paradigm, it ignores an important component of many real-world retrieval tasks: the decision to terminate memory search. The present study examines the temporal characteristics underlying memory search by comparing within subjects a standard retrieval paradigm with a finite, preset amount of time (closed interval) to a design that allows participants to terminate memory search on their own (open interval). Calling on the results of several presented simulations, we anticipated that the threshold for number of retrieval failures varied as a function of the nature of the recall paradigm, such that open intervals should result in lower thresholds than closed intervals. Moreover, this effect was expected to manifest in interretrieval times (IRTs). Although retrieval-interval type did not significantly impact the number of items recalled or error rates, IRTs were sensitive to the manipulation. Specifically, the final IRTs in the closed-interval paradigm were longer than those of the open-interval paradigm. This pattern suggests that providing participants with a preset retrieval interval not only masks an important component of the retrieval process (the memory search termination decision), but also alters temporal retrieval dynamics. Task demands may compel people to strategically control aspects of their retrieval by implementing different stopping rules.

Keywords: Free recall; Memory search and retrieval; Interretrieval times; Stopping rules.

A fundamental component of many information search tasks is the decision of when to terminate the search process. Information search can focus either on the external or on the internal environment (Hills, Todd, & Goldstone, 2008). External search concerns the acquisition of information readily available in the physical environment. Internal search occurs any time memory is accessed, whether it is for the purpose of recollecting items from a list or generating a set of hypotheses (Thomas, Dougherty, Sprenger, & Harbison, 2008). Regardless of the nature of the task, people must at some point decide to terminate search. In some cases, the decision to terminate is marked by the successful retrieval of a single target memory. In other circumstances, it may not

Correspondence should be addressed to Erika K. Hussey, Department of Psychology, Program in Neuroscience and Cognitive Science, University of Maryland, 1147 Biology-Psychology Building, College Park, 20742, MD, USA. E-mail: ehussey@umd.edu

This research was supported by Grant BCS-1030831 from the National Science Foundation. We thank David Alexander, Pooja Datta, Dzi Bo, Angela Choi, Danielle Kershberg, Laura Rego, and other members of the Decision, Attention, and Memory Lab for their assistance with collecting and scoring data. We also thank Erica Yu and Jeffrey Chrabaszcz for their valuable feedback on sections of the paper. A portion of this work was presented at the 50th Annual Meeting of the Psychonomic Society in Boston, MA, the 32nd Annual Meeting of the Cognitive Science Society in Amsterdam, The Netherlands, and the 34th Annual Meeting of the Cognitive Science Society in Sapporo, Japan.

be clear when memory search should be terminated, as the amount of to-be-searched-for information might be ambiguous or large. For example, two situations that involve memory search termination are verbal-fluency tasks and free-recall tasks, in which participants must retrieve as many items as possible from either a semantic category or a previously studied list. Verbal-fluency tasks require that the participant search an ill-defined and potentially large sample space to find semantic associates of a cue. In contrast, free-recall tasks involve an objectively well-defined search space that is typically relatively small—with the size dependent on the design of the experiment—such that the participant must separate recently learned items (i.e., those on the list) from prior experiences. A shared feature of both of these tasks and many others is that the participant must decide when and on what basis to terminate search—that is, unless the experiment is designed in a way to obviate this element of the search process by, for example, providing participants with a time limit for how long search can last.

The goal of the present research is to understand the mechanisms that influence how people terminate memory search in free-recall tasks and to examine the influence of self-terminated memory search on characteristics of the retrieval process. Since most memory retrieval experiments use a predetermined recall interval where participants are given a finite amount of time to recall items from a recently presented list (e.g., Anderson & Bower, 1972; Kahana, Howard, Zaromb, & Wingfield, 2002; Rohrer & Wixted, 1994), there is an experimental void of accurate measurements of retrieval time in order to evaluate stopping decisions. This is because in most recall studies the experimenter controls the duration of the retrieval interval, such that participants typically are not given the opportunity to overtly indicate when they have stopped trying to retrieve additional items within this finite interval. As a result, it is unclear whether participants persist in search for the entire allotted time or opt to truncate search prior to the expiration of the preset interval. Even though search termination decisions are ubiquitous in nonlaboratory memory tasks, and many of the computational models of free recall assume that participants make termination decisions during the retrieval process (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998; Raaijmakers & Shiffrin, 1981), few studies (Dougherty & Harbison, 2007; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Klein, Addis, & Kahana, 2005; Shiffrin, 1970; Unsworth, Brewer, & Spillers, 2011) have examined memory processes in contexts in which memory search is self-terminated.

The difference between self-terminated (open-interval) and predetermined (closed-interval) memory search tasks is illustrated in Figure 1. For the purposes of this example, assume that participants are engaged in a free-recall task in which they are instructed to retrieve multiple items from memory. In a closed-interval paradigm (Figure 1A), the experimenter decides the total amount of time (e.g., 10 s, 45 s, or 5 min) allowed for retrieval. In contrast, in an open-interval paradigm (Figure 1B), participants are allowed to terminate search on their own and can, in principle, spend an indefinite amount of time on any particular list. Thus, open-interval designs allow the participant—rather than the experimenter—to determine the total time spent in search.

Although the closed-interval paradigm allows for precise control of recall trials and makes it possible for participants to demonstrate asymptotic levels of recall, one major disadvantage of the design is that there is no latency measurement of participants' search termination decisions. In contrast, the open-interval paradigm licenses the measurement of total time spent in search, making it possible for additional latency measures to be computed (e.g., exit latency, or the time between the final retrieval and the termination of memory search, see Dougherty & Harbison, 2007, and Figure 1). While total time and exit latency represent fine-grained assessments of the retrieval process (compared to recall rates and errors), they have proven to follow a surprisingly lawful pattern across experiments and participants. For example, Harbison and colleagues (2009) found that 85–95% of individual subjects in two separate experiments showed the same basic pattern of data for total time and exit latencies:

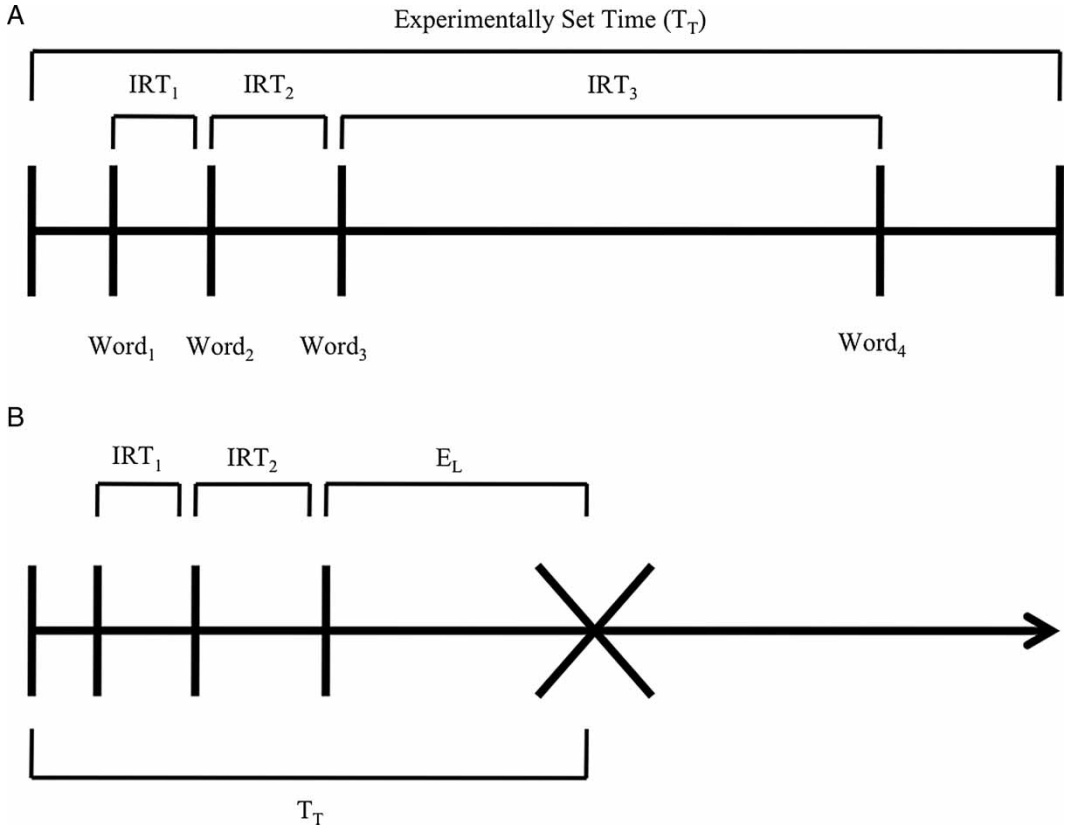

Figure 1. *(A) Closed-interval free-recall trial and (B) open-interval retrieval free-recall trial. X indicates the time when a participant decides to terminate memory search; hash marks indicate the time associated with the latency onset of words recalled. T_T = total time searching; E_L = exit latency; IRT = interretrieval time. From "Motivated To Retrieve: How Often Are You Willing To Go Back To The Well When The Well Is Dry?", by M. R. Dougherty and J. I. Harbison, 2007,* Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, *p. 1109. Copyright 2007 by American Psychological Association.*

Total time increased and exit latency decreased as a function of number of words retrieved (see also Unsworth et al., 2011). Importantly, this behavioural regularity was anticipated by only one of the four stopping rules tested—namely, the total number of failures rule (cf. Raaijmakers & Shiffrin, 1980), which posits that memory search is terminated based on the total number of retrieval failures. Additional evidence for a number of failures rule comes from Unsworth and colleagues (2011) who demonstrated that recall errors tended to precede search terminations. Related to this, our earlier simulations demonstrate that error patterns *correlate* with exit latencies, such that exit latencies decrease as more retrieval errors are committed (see Harbison, Davelaar, Yu, Hussey, & Dougherty, in press). Therefore, total amount of search time and the exit latency may be diagnostic for assessing the mechanisms (i.e., stopping rules) governing the termination of memory search.

While previous work has made progress in describing the behavioural regularities involved with self-terminated memory search (Dougherty & Harbison, 2007; Harbison et al., 2009; Unsworth et al., 2011), there has not been a direct comparison between open- and closed-interval memory retrieval. Anecdotally, inspection of data presented in Dougherty and Harbison (2007) and Harbison et al. (2009) suggests that

many of the data patterns generated using an open-interval retrieval paradigm are consistent with previously collected data from closed-interval retrieval. However, there is one apparent discrepancy: The pattern of interretrieval times (IRTs) appears to be qualitatively distinct in the open-interval paradigm.

An IRT refers to the amount of time that has passed between two successive retrievals (e.g., the time between the first and the second retrieval, the second and the third, etc.; see Figure 1). IRTs have played an important role in explaining memory retrieval (Rohrer, 1996; Rohrer & Wixted, 1994; Wixted & Rohrer, 1994) and are generally well described by the equation:

$$\mathrm{IRT}_i = \frac{\tau}{N - i} \tag{1}$$

for $i = 1, 2, \ldots, N-1$, where i is the interresponse interval starting with the interval between the first and second retrievals; tau (τ) is the estimated mean retrieval latency; and N is the total number of items retrieved. Equation 1 captures the key empirical result that IRTs typically follow a hyperbolic function, such that the time between successive retrievals increases (Murdock & Okada, 1970; Polyn, Norman, & Kahana, 2009; Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). More specifically, if reciprocally transformed IRTs (1/IRT) are plotted from the final retrieval interval to the first retrieval interval, the result is a line with slope $1/\tau$ that crosses the intercept at zero during closed-interval retrieval (see Rohrer, 1996). These results are taken as strong support for a relative-strength model of recall, because Equation 1 is predicted by the equal memory strength variant of the model, and in the presence of a recovery threshold, the unequal memory strength version of the model predicts the same pattern (Rohrer, 1996). Contrary to Rohrer's findings, Figure 2 shows that the 1/IRT data generated from participants in an open-interval retrieval task from Dougherty and Harbison (2007) and from Harbison et al. (2009, Experiment 1) are *not* well fitted by Equation 1. In particular, the intercept is consistently and substantially greater than 0 (ranging from .275 to .495). The discrepancy is particularly prevalent at the final IRT, where participants did spend significantly more time retrieving the final item than what Equation 1 predicts.

Are these open-interval retrieval results at odds with the relative-strength model or do they reflect separate search termination rules implemented by participants performing an open-interval task? To test this, we adopted the simulation methodology used by Rohrer (1996) to establish the predictions of the relative-strength model. The implemented model randomly sampled items from memory with replacement based on their relative activation. A sampled item was deemed recovered if its activation exceeded an absolute activation threshold of 0.5. Our simulations used the same deterministic threshold as that of Rohrer (1996), as well as the same activation distributions (see below). The only change was the inclusion of a memory search stopping rule; in particular, we implemented the total number of retrieval failures rule previously found to account for open-interval retrieval data (Harbison et al., 2009). According to this rule, participants are sensitive to the total number of retrieval failures they experience during recall, where a retrieval failure is any retrieval attempt that does not produce a new memory item. In the case of the reported simulations, a retrieval failure was defined as one of the following: (a) sampling an item that did not meet or exceed the recovery threshold or (b) resampling an item (i.e., repetitions). Once the maximum number of retrieval failures (the stopping threshold) was reached, the model terminated search. Note that we included two sets of simulations, each of which maintained different assumptions regarding the recovery threshold: As sketched above, following from Rohrer (1996), the first set of simulations assumed a *deterministic* threshold such that activations must exceed some specified value (0.5) to be recovered. The second set of simulations maintained a *probabilistic* threshold akin to that implemented in Search of Associative Memory model (SAM; Raaijmakers & Shiffrin, 1980, 1981), wherein items are recovered in proportion to their activation levels.

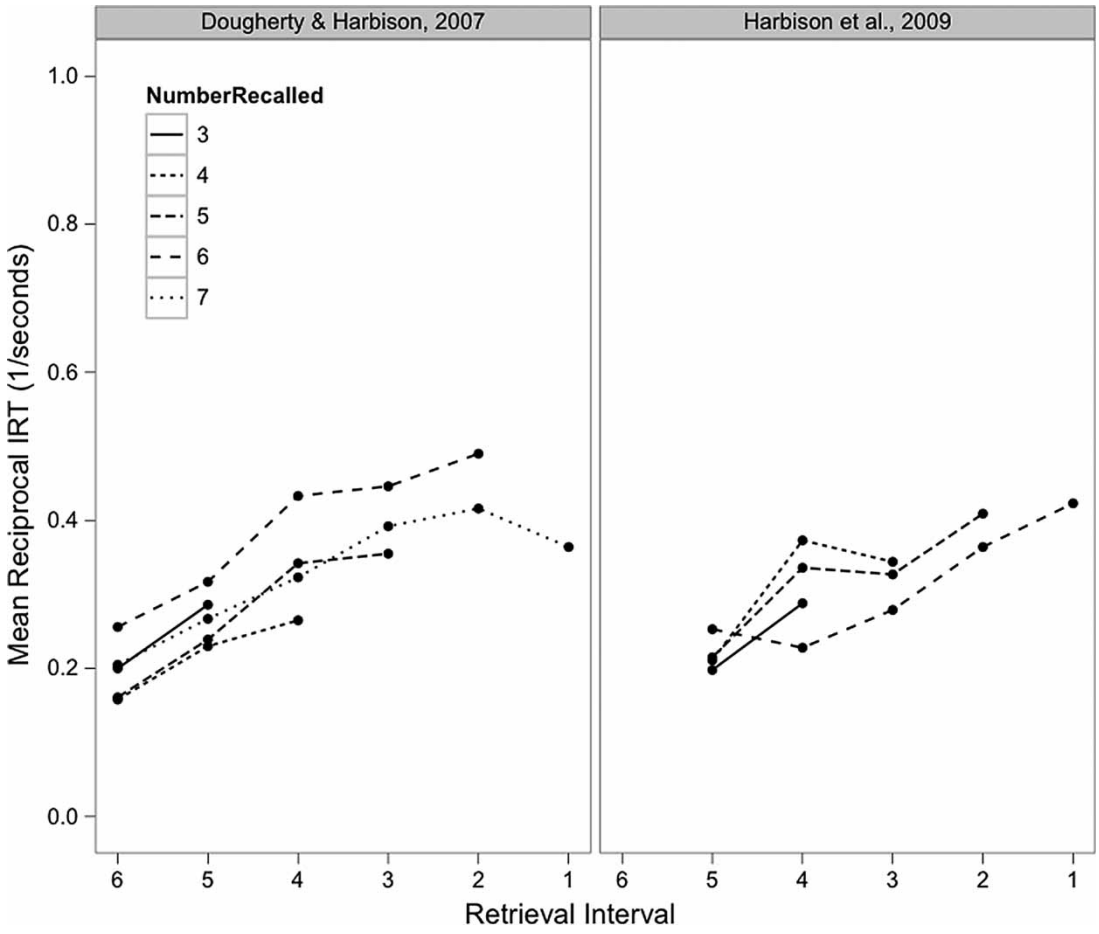

Figure 2. *Mean reciprocal interretrieval time (1/IRT) from Dougherty and Harbison (2007; left panel) and Harbison et al. (2009; right panel) for each number recalled (3, 4, 5, 6, and 7 items retrieved), with 1 on the* y*-axis representing the final interval and 0 representing the intercept. The* x*-axis is the retrieval interval plotted in reverse order. Left panel: From "Motivated to retrieve: How often are you willing to go back to the well when the well is dry?", by M. R. Dougherty and J. I. Harbison, 2007,* Journal of Experimental Psychology: Learning, Memory, and Cognition, 33. *Right panel: From "On the lawfulness of the decision to terminate memory search", by J. I. Harbison, M. R. Dougherty, E. J. Davelaar, and B. Fayyad, 2009,* Cognition, 111.

Deterministic recovery threshold

Figure 3A illustrates the reciprocal IRT results for two memory item activation distributions (0.5, 0.6, 1.0, 1.2, 1.2, 1.5 and 0.4, 0.5, 0.6,1.0, 1.5, 2.0; from Figures 2a and 5a of Rohrer, 1996) for stopping thresholds from 10 to 30 in increments of 10 with each activation distribution and stopping threshold pairing run 10,000 times. Note that all of the activations tested in Rohrer (1996) showed the same pattern of results. In our simulations, the reciprocal IRTs, and in particular the final reciprocal IRT, were found to vary as a function of the stopping threshold. When the stopping threshold was higher, the intercept was closer to zero, mimicking the effects seen during a closed-interval retrieval period, consistent with Rohrer's 1996 model. Moreover, by comparing Figures 2 and 3A, it is apparent that the results reported from the open-interval designs of Harbison et al.

(2009) and Dougherty and Harbison (2007) pattern closely to those associated with stopping rules assuming lower stopping thresholds.

As shown in Figure 3B, the untransformed IRT data demonstrate a larger impact of variation in the stopping threshold than the reciprocal IRTs in Figure 3A. Perhaps surprising, despite the large fluctuations in the last IRT, adjusting the stopping threshold had a relatively minor impact on the number of items retrieved. Number recalled varied from 5.35 to 5.97 for one activation pattern and 4.35 to 4.97 for the other with slightly more items anticipated for the closed block, whereas changes in the final IRT varied from 3.35 to 8.12 and 3.38 to 8.43, respectively, with longer IRTs predicted for larger thresholds of a number of failures stopping rule within an open-interval retrieval period. This demonstrates that the number of failures stopping rule influences temporal properties more than recall rates.

Probabilistic recovery threshold

Next, we conducted simulations assuming a probabilistic recovery threshold similar to that implemented in SAM. In particular, items were deemed recoverable proportionate to their activation levels. Figure 3C depicts the untransformed IRT results of these simulations for stopping thresholds from 10 to 30 in increments of 10. Similar to the deterministic thresholds (in Figures 3A and 3B), stopping thresholds had significant effects on the last IRT, but minor influences on

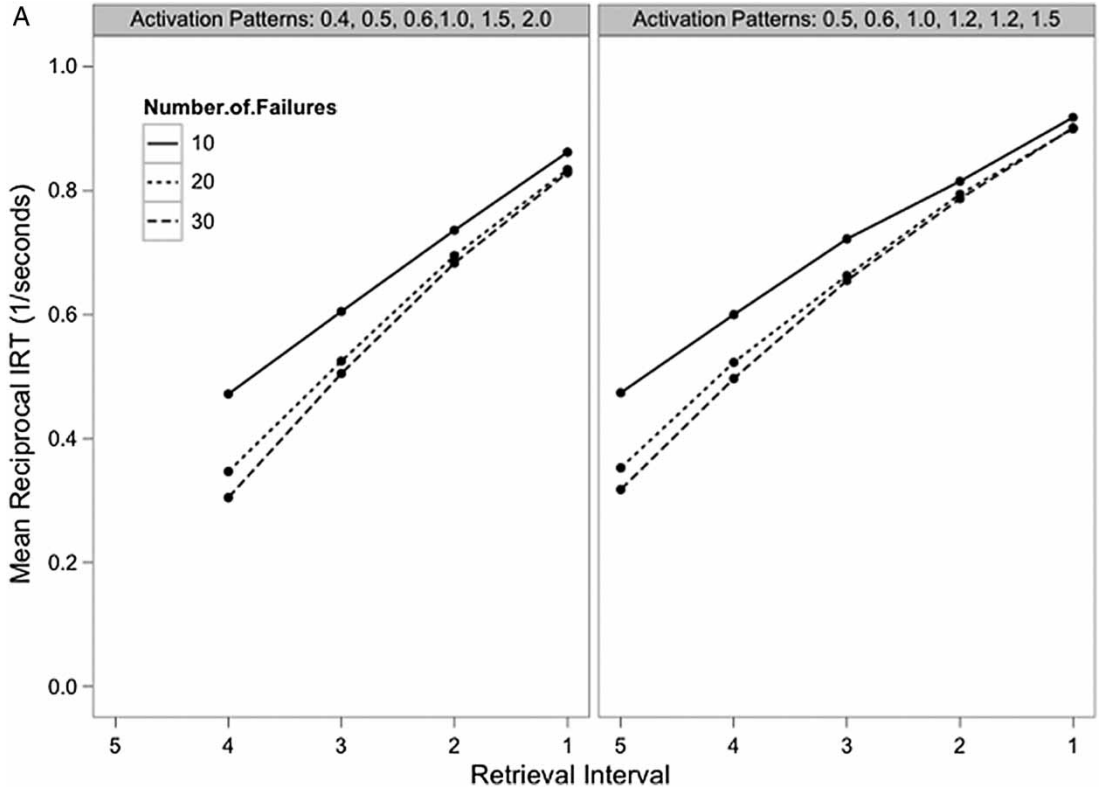

Figure 3. *Simulation results using a random sampling-with-replacement model with a recovery threshold of 0.5, sampling from the activation patterns 0.4, 0.5, 0.6,1.0, 1.5, 2.0 (left panels of Figures A and B) and 0.5, 0.6, 1.0, 1.2, 1.2, 1.5 (right panels of Figures A and B). (A) Mean reciprocally transformed and (B) raw, untransformed reciprocal interretrieval time (IRT) results for 10, 20, and 30 failures for models with a deterministic recovery threshold of 0.5 and (C) a probabilistic recovery threshold. The* x*-axis denotes retrieval intervals plotted in reverse order.*

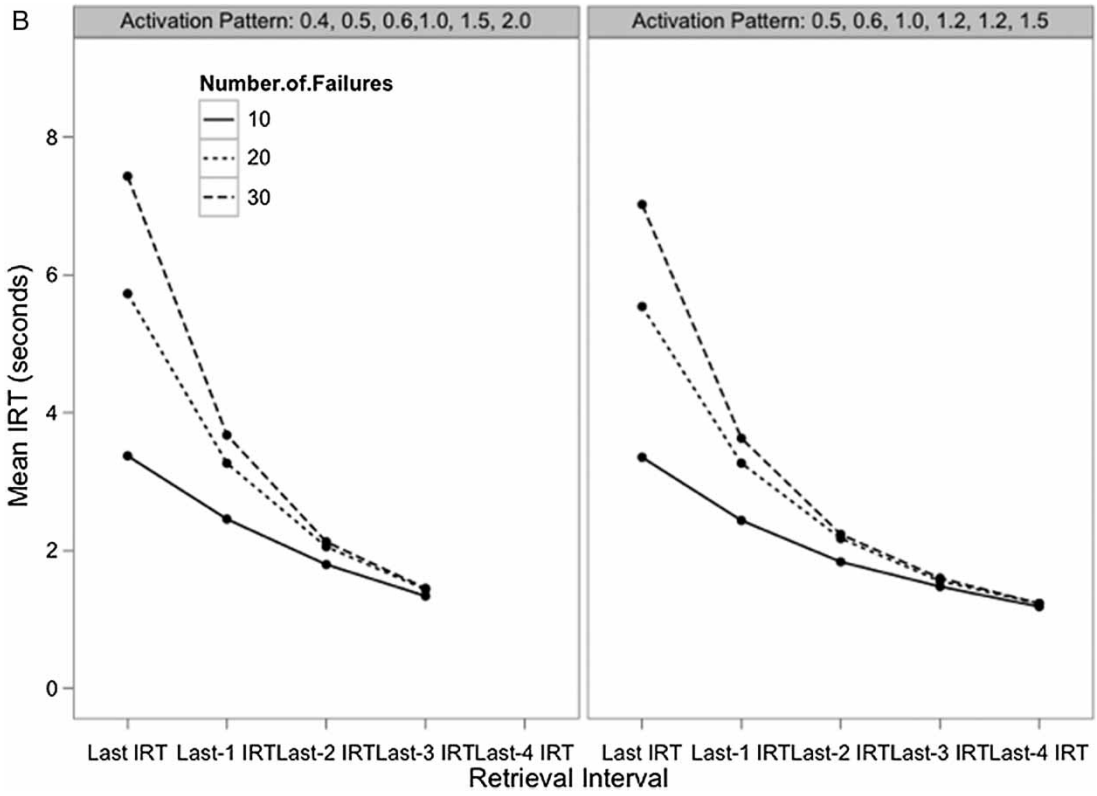

Figure 3. *Continued.*

number recalled. Namely, the last IRT varied from 3.16 s to 6.90 s, such that longer IRTs were present for larger number of failures thresholds (see Figure 3C). Alongside the deterministic threshold above, the underlying predictions of the role of a number of failures stopping rule were largely unaffected by recovery assumptions put forth by each model; specifically, we varied the recovery threshold—as deterministic versus probabilistic—and observed the same IRT patterns as a function of the stopping threshold, such that fewer number of failures resulted in flatter overall IRT functions than did greater number of failures.

The results of the present simulations suggest that a sampling-with-replacement model equipped with a stopping rule accounts for the differences between the closed- and open-interval paradigms, but only if we assume that the stopping threshold is larger during closed-interval retrieval. This assumption is not unreasonable provided that having additional time drives participants to find more items before ending memory search. Furthermore, note that by not including a stopping rule, we would not expect to see a difference in the IRTs in the open- and closed-interval retrieval conditions because the only factor that discriminates these conditions is the termination threshold. Rather, if there is no stopping rule, the final IRTs would be exaggerated in both conditions, given that search would proceed until every list item was retrieved. In fact, the mathematical approximation of such a model would predict that the intercept would cross zero (akin to Equation 1), because higher probabilities of retrieving the final item on a list—and in this case, all list items—correspond to lower intercepts.

In light of these patterns, it is important to note that the empirical evidence for a difference in IRTs

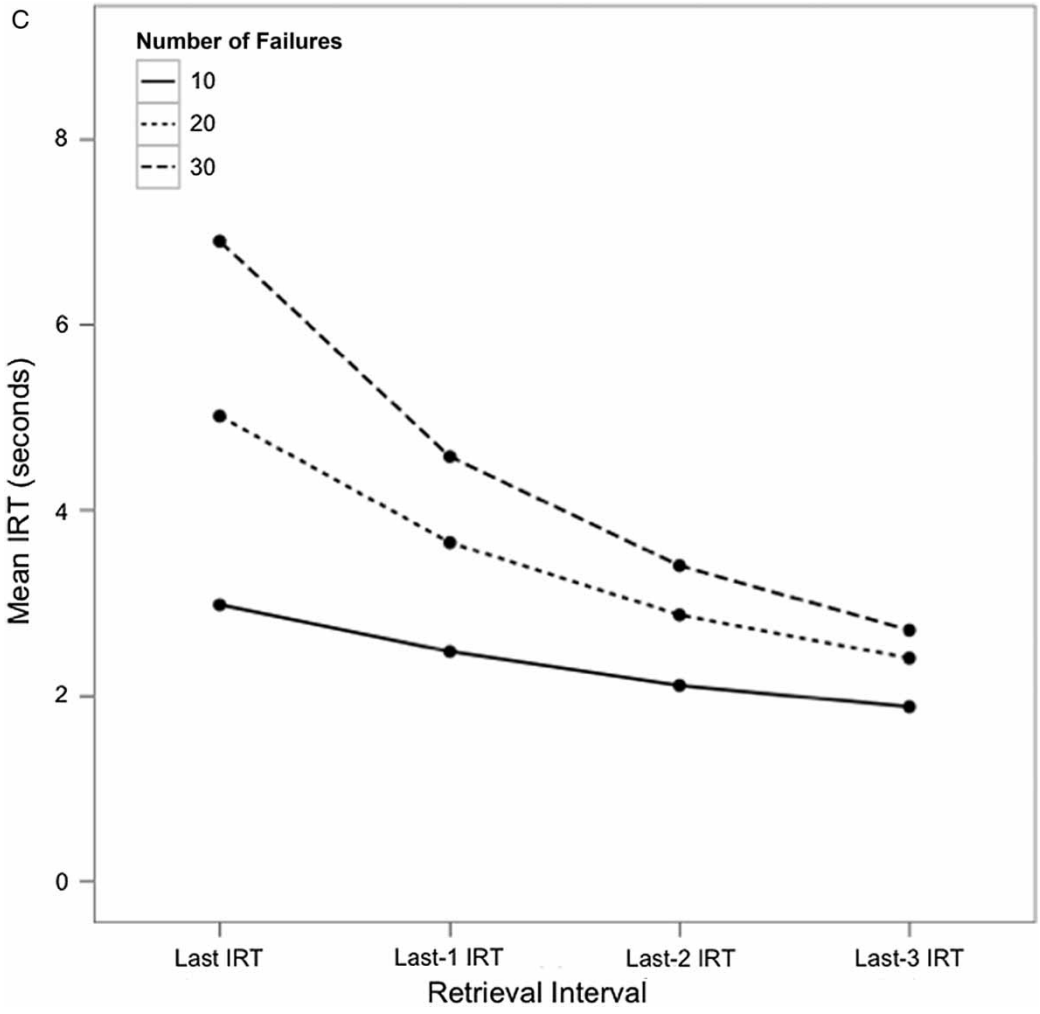

Figure 3. *Continued.*

between open- and closed-interval paradigms is only anecdotal. Therefore, we conducted a within-subjects experiment where participants completed blocks of both paradigms to directly compare the temporal dynamics of memory search in each. Based on the model simulations presented above, we predicted that by allotting a set period of time in the closed-interval retrieval block, we would sway participants to adopt a larger threshold on their stopping rule, such that more failures (e.g., 30) are required to render a search termination decision. Further, we anticipated this stopping rule to manifest as an exaggerated final IRT compared to that of the open-interval retrieval block. Put differently, allowing participants to terminate search on their own should compel them to quit after fewer failures (e.g., 10), leading to a more temporally uniform retrieval trend.

Method

Participants

Forty-nine University of Maryland psychology undergraduate students participated in the

experiment for course extra credit. Block order was counterbalanced between participants such that 27 participants received the open-interval block first, and 22 participants received the closed-interval block first.

Stimuli

The resulting experiment comprised a 2 (block type: open vs. closed) × 4 (list length: 5, 7, 9, 11 words) within-subjects design. List length was varied randomly such that all subjects were given four study lists of each of four lengths evenly and randomly within each block. List length was systematically varied for two reasons: First, the unpredictable length of various lists may prevent participants from learning exactly how many items were on each list, thereby preventing them from using their knowledge of list length as a guide for how long to continue in search; indeed, through a quasi-experimental analysis, Unsworth and colleagues (2011) demonstrated that total search time increases with knowledge of set size, compared to cases when memory list length is not explicitly specified. Second, we sought to ensure sufficient variability in the number of items retrieved to examine IRTs and exit latency as a function of number recalled. This is particularly relevant given that more items are recalled from longer list lengths, and recall time is proportional to study list length (Roberts, 1972). Note that because previous studies investigating the mechanisms of retrieval latencies tend to implement a uniform list length (e.g., Murdock & Okada, 1970; Rohrer, 1996; Rohrer & Wixted, 1994), we caution the reader against comparing the present findings directly to previous work because of this critical design difference. Despite this, we feel that by looking at each list length level separately, we will be able to glean insight into the effects of list length—and, thus, proportion recalled—on retrieval dynamics.

Thirty-six word lists were randomly generated for each participant from a list of 280 high-imageability (M = 577/700), high-concreteness (578/700), moderate-to-high-frequency (Kucera–Francis frequency = 54), single-syllable nouns drawn from the MRC psycholinguistic database (Coltheart, 1981; Wilson, 1988).

Procedure

In each block (open vs. closed), participants were given a total of 18 lists consisting of two practice trials followed by 16 test trials. During the list presentation of each trial, words were presented sequentially at a rate of 3 s per item. Following a study list, participants were given a distractor task that consisted of two simple, timed maths problems. Each problem contained three digits and two operands (e.g., 3 * 2 + 1). Each component of the problem was presented sequentially at a rate of 1 s per item. After viewing the final digit of the problem, participants saw an equals sign with a question mark. This prompted them to provide an answer using the keypad (one of 10 single-digit numbers, 0–9). The range of average maths accuracy was 64.1–98.4, well above the chance threshold of 10% accuracy; therefore, we assume that the maths problems sufficiently distracted participants from actively rehearsing list items. Additionally, on average, this maths distraction task occupied a 14.85-s interval between the presentation of the final word on the list and the retrieval prompt. Time spent on the distractor task did not vary as a function of block, block order, or number recalled in the experiment (ps > .45). Importantly, prior work (Postman & Phillips, 1965) has shown that a 15-s filled interval is sufficient to eliminate the contribution of short-term memory for long-term memory retrieval. Consistent with this, we find attenuated recency effects, such that the probability of recalling later list items is no different from recalling intermediate list items, $t(186) = -1.76$, $p > .08$; however, the probability of recalling earlier list items is greater than that for intermediate list items, $t(194) = 13.21$, $p < .001$, compatible with a classic primacy effect.

During open trials, participants were given control over when to end the retrieval interval. Participants were told to press the spacebar when they could no longer retrieve additional items from the most recently studied memory list. An unlimited amount of time was provided for

participants to verbally recall the words presented on each word list (cf. Dougherty & Harbison, 2007). The following typed instructions were presented prior to starting: "During the retrieval period, please try to recall as many words as you can. You will be given as much time as you need to recall the words. When you feel as though you cannot recall any additional words, please press the spacebar to terminate the trial."

During closed retrieval blocks, participants were given 45 s to retrieve the memory list items. The following typed instructions were presented prior to starting: "During the retrieval period, please try to recall as many words as you can. You will be given a set amount of time to recall the words. When the time has expired, you will see a screen asking you to press a button in order to begin the next trial." Based on prior research, we anticipated that a 45-s retrieval interval would provide ample time for most participants to complete the recall task. Moreover, this retrieval interval has been used in other research (e.g., Kahana et al., 2002, Experiment 2) and is well beyond the average duration that participants in our prior studies spent retrieving when provided with the option to self-terminate search (see Dougherty & Harbison, 2007; Harbison et al., 2009). All subjects were presented with both block types to ensure a proper comparison of IRTs between the open and closed intervals, and the order of block presentation was counterbalanced across participants.

All retrievals were made verbally by speaking into a microphone and were digitally recorded for later scoring. The responses for each list were hand-coded to extract the items recalled and the time-to-word associated with each.

Coding

Using Audacity© audio-analysis software, we retrospectively analysed all aspects of the retrieval data with millisecond accuracy. Two separate coders analysed the verbal recall data to include: (a) all words that were produced by each participant on each trial, (b) the time stamps of the verbal onset of all generated words, and (c) the time stamps indicating retrieval termination for open-interval trials (i.e., times associated with spacebar presses). Total number of items recalled, the average number of intrusions (i.e., repetitions and previous and extraexperimental false alarms), and interretrieval times were analysed separately at each list length for the open- and closed-retrieval blocks for each participant. Exit latencies were also computed at each list length for each participant on open blocks only. IRTs were examined as a function of list length, number recalled, and block type.

Results

We conducted Jeffreys–Zellner–Siow (JZS) Bayes-factor (BF) tests to verify the results of each *t* test reported below using R's ttest.Quad function (BayesFactorPCL library, Morey & Rouder, 2010; see Rouder, Speckman, Sun, Morey, & Iverson, 2009, for a detailed explanation of BFs of *t* tests). JZS BF tests include a parameter, *r*, used to index expected effect sizes; because we have no hypotheses with respect to effect size, *r* was set a priori to a default value of 1.0. Cauchy priors were assumed for all BF tests implemented for each reported balanced one-way analysis of variance (ANOVA) model below using R's onewayAOV.Quad function (BayesFactorPCL library, Morey & Rouder, 2010; see Masson, 2011; Rouder, Morey, Speckman, & Province, 2012); note that where an ANOVA model is unbalanced or requires more than one factor of interest, BFs are not reported. Some comparisons are expected to support the null hypothesis, and JZS BFs provide a means to assess the degree to which this is indeed the case. Bayes-factor tests reflect the likelihood of support for the alternative hypothesis over support for the null hypothesis, such that for *t* tests, coefficients less than 0.1 index strong support for the null hypothesis and those less than 0.3 index substantial support for the null hypothesis, while those greater than 3 index substantial support for the alternative hypothesis, and those greater than 10 strongly support the alternative hypothesis. All analyses are conducted with respect to two variables of interest: block type (open vs. closed) and list length (5 vs. 7 vs. 9 vs. 11).

Block order

We first conducted a manipulation check to determine whether there was an effect associated with block order (open interval first vs. closed interval first). A repeated measures ANOVA with factors block order, block type, and list length revealed no effect for average total number of items recalled, average number of correctly recalled items, or total number of intrusions (Fs < 0.963; ps $> .41$). Exit latency cannot be computed for trials in the closed-interval block, so we examined the effect of block order and list length on exit latency only on open-interval trials and found no main effect, $F(1, 192) = 0.4069$, $p > .52$. Finally, there were no reliable effects of block order, block type, and list length for IRTs at any level of number recalled (ps $> .09$). Because these early analyses suggest that there are no effects of block order, all subsequent analyses collapsed across this factor.

Number recalled

We next conducted an ANOVA with factors block type (open vs. closed) and list length (5, 7, 9, 11) on the total number of words recalled. This test failed to reach significance, $F(3, 388) = 0.011$, $p > .92$; however, as can be seen in Figure 4A, there was a main effect of list length, $F(1, 194) = 31.33$, $p < .001$, BF > 100, but no significant difference in the number of items retrieved between the closed- and open-interval trials, $F(1, 96) = 0.596$, $p > .44$, BF $= 0.149$; see Table 1). Note, however, that quantitatively the closed block resulted in slightly more items recalled than the open block. The same series of patterns held when only *correctly* recalled items were considered, such that the interaction of block type and list length failed to reach significance, $F(3, 388) = 0.011$, $p > .92$, but there was a main effect of list length, $F(1, 194) = 14.215$, $p < .001$, BF > 100, such that more items are recalled with longer list lengths, though no effect of block type, $F(1, 96) = 0.224$, $p > .64$, BF $= 0.125$; see Figure 4B and Table 1). That is, participants generated more correct words in the closed- than open-interval trials, though this difference did not reach significance. While there was no significant difference in number recalled as a function of block type, note that the simulations presented above predicted that the closed interval would render slightly more items retrieved than the open interval, as we observed.

Analyses of recall errors (i.e., intrusions and repetitions) are also consistent with this conclusion: The average number of intrusions did not differ as a function of list length and block type, $F(3, 384) = 0.382$, $p > .77$. Intrusion rates did not change as a function of time spent in the experiment for any list length in either block (ps $> .09$; BFs < 0.432). Moreover, the interaction of list length and block type failed to reach significance when intrusions were split into three types (Fs < 1.916; ps $> .13$): (a) extraexperimental false alarms, or items recalled that were not presented during any prior studied lists in the experiment (left panel of Figure 4C); (b) previous-list false alarms, or items that were incorrectly output that occurred on previous memory lists within the experiment (centre panel of Figure 4C); and (c) repetitions (right panel of Figure 4C). There was a reliable main effect of list length for previous-list false alarms where more intrusions were observed with longer list lengths, $F(3, 192) = 4.130$, $p < .01$; see Table 1), but no main effects reached significance for the remaining intrusion measures (ps $> .18$). Despite this, the mean number of previous-list intrusions never exceeded 0.25 in either block at any list length, indicating that although this main effect exists, it is negligible with respect to interpreting how these intrusions may manifest as a function of list length. Given these results, we are comfortable concluding that the open-interval paradigm does not differ substantially from the closed-interval paradigm in terms of number and type recalled, even in the presence of various list lengths.

Exit latency, total time, and time-to-last

The open-interval design provides the unique opportunity to measure total self-paced search time and exit latency during a retrieval period. We found the current data to be consistent with previous findings using an open-interval paradigm,

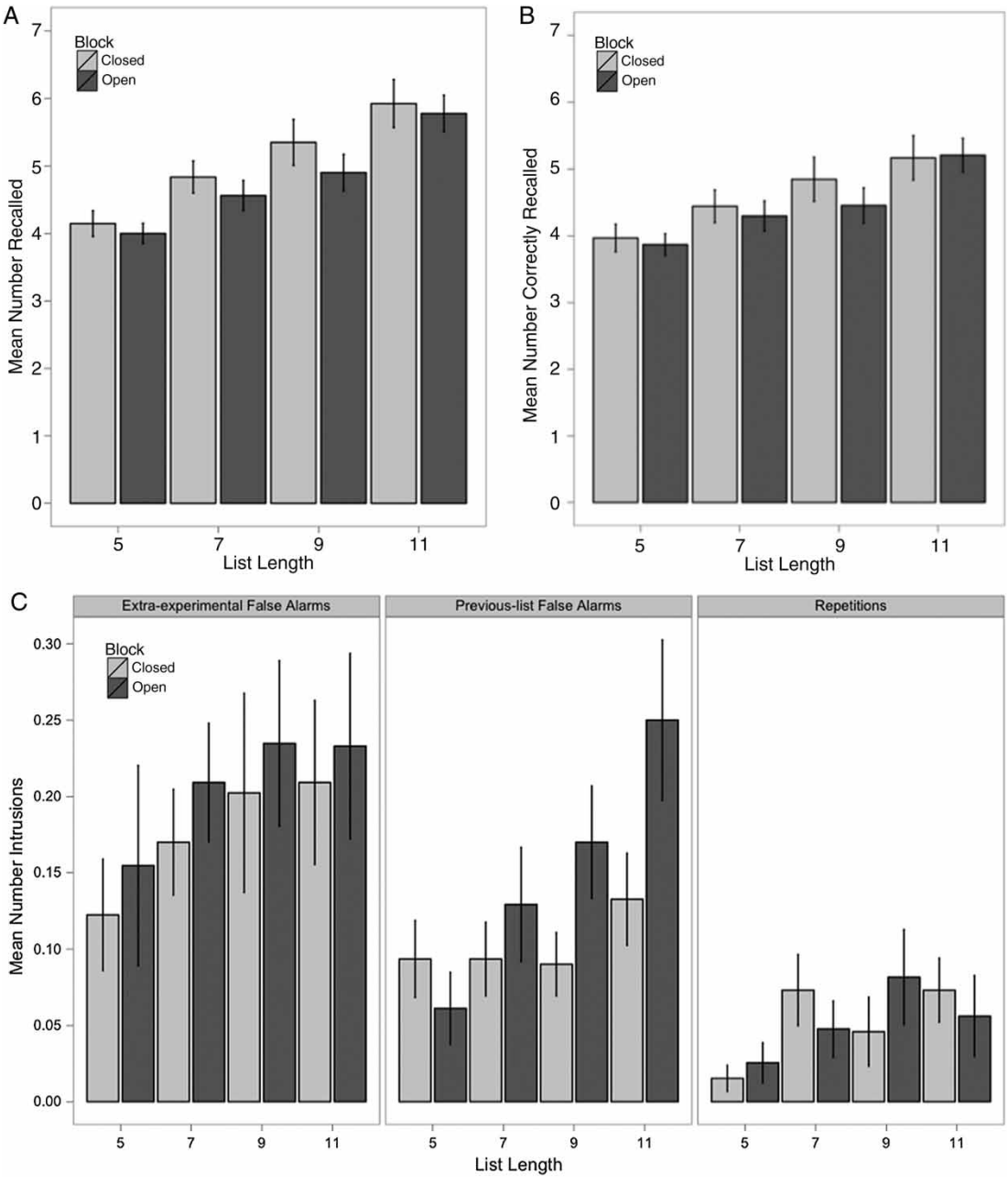

Figure 4. *(A) Mean total number of items (hits, repetitions, and false alarms) retrieved by block and list length. (B) Mean number of correct items (i.e., hits only) retrieved by block and list length. (C) Mean number of errors by block and list length: The left panel depicts mean number of extraexperimental false alarms, the centre panel illustrates mean number of previous-list false alarms, and the right panel shows mean number of repetitions. Error bars* $= \pm 1$ *standard error of the mean.*

Table 1. *Mean and standard deviation for number of items recalled split by total, hits, extraexperimental false alarms, previous-list false alarms, and repetitions for each block type collapsed across list length and for each list length collapsed across block type*

		Total items recalled		*Correct items recalled (hits)*		*Extraexperimental intrusions*		*Previous-list intrusions*		*Repetitions*	
Block type	*List length*	M	SD	M	SD	M	SD	M	SD	M	SD
Open	All	4.89	1.47	4.46	1.39	0.20	0.32	0.15	0.20	0.05	0.12
Closed	All	5.11	1.84	4.62	1.82	0.18	0.25	0.10	0.11	0.05	0.08
All	5	4.08	0.98	3.92	1.15	0.12	0.23	0.08	0.12	0.02	0.06
All	7	4.69	1.44	4.39	1.52	0.19	0.18	0.11	0.16	0.06	0.12
All	9	5.11	1.91	4.62	1.92	0.21	0.31	0.13	0.16	0.06	0.13
All	11	5.87	1.96	5.16	1.89	0.22	0.30	0.19	0.22	0.06	0.12

Note: Block type: open versus closed. List length: 5, 7, 9, 11.

such that exit latency decreased as a function of number recalled (Dougherty & Harbison, 2007). Mean within-subject gamma (γ) correlation coefficients for exit latency and number recalled ($M = -.139$, $SD = .26$) indicate that participants take longer to terminate search after the final item is recalled when fewer words are output in a trial (one-sample t test of γ), $t(48) = -3.719$, $p < .001$, BF = 43.731). Furthermore, one-sample t tests of gamma at each list length indicated that this pattern holds for all lengths greater than 5 (see Table 2).

Also consistent with previous data, participants spent more overall time searching at longer list lengths, borne out by a reliable main effect of list length, $F(1, 194) = 37.301$, $p < .001$, BF > 100; see Table 3). Moreover, participants spent significantly less total time searching in the open-interval condition at all list lengths than in the 45-s closed-interval period ($ts > 18.74$; $ps < .001$).

We next examined time-to-last, or the amount of time it takes from first starting a retrieval interval to generating the final recalled item. We observed a significant main effect of list length, $F(1, 194) = 46.475$, $p < .001$, BF > 100, but no effect of block type, $F(1, 96) = 1.0358$, $p > .31$, BF = 0.184. That is, the closed-interval condition resulted in a longer time-to-last than the open-interval condition, $t(48) = 29.43$, $p < .001$, BF > 100. Additionally, longer list lengths resulted in longer time-to-last, consistent with the patterns observed for total time searching and exit latency (see Table 3).

Table 2. *Mean and standard deviation for gamma correlations for exit latency as a function of number recalled*

Block type	*List length*	M_γ	SD_γ	t	*BF*
Open	All	–.14	.26	–3.719***	43.73
Closed	All	—	—	—	—
Open	5	–.13	.68	–1.37	0.31
Open	7	.31	.59	–5.81***	>100
Open	9	–.27	.55	–3.45***	20.35
Open	11	–.30	.57	–3.65***	35.71

Note: One-sample t-tests of each gamma correlation are reported for each block type (open vs. closed) collapsed across list length and for each list length (5, 7, 9, 11) collapsed across block type. BF = Bayes factor.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3. *Mean and standard deviation for total retrieval time, time to recall the final item (time-to-last), and final interretrieval times for each block type collapsed across list length and for each list length collapsed across block type*

		Total time (s)		*Time-to-last (s)*		*Last IRT (s)*	
Block type	*List length*	M	SD	M	SD	M	SD
Open	All	13.88	6.90	9.64	4.23	2.79	1.69
Closed	All	45	—	11.54	5.57	4.46	2.40
All	5	9.26	4.97	6.18	3.83	1.30	0.71
All	7	13.23	6.58	8.79	5.31	1.61	1.06
All	9	16.77	9.49	11.06	6.76	2.38	1.41
All	11	18.47	9.91	12.87	7.37	2.69	1.98

Note: Time-to-last = time to recall the final item; IRT = interretrieval time. Block type: open versus closed. List length: 5, 7, 9, 11.

Interretrieval times

Interretrieval times (IRTs) were computed by taking the difference between the verbal onset times for each subsequent item recalled in a trial. We conducted these irrespective of the identity of the item recalled (i.e., IRTs were computed to incorporate trials containing intrusions to boost the number of observations per cell). We reported the data including intrusions for consistency with Rohrer (1996). Moreover, it is unlikely that subjects who output intrusions are aware of such errors during the experiment. Therefore, these items should have little effect on temporal dynamics associated with recall. Including or excluding intrusions did not result in different patterns of the IRT results presented below; thus all subsequent analyses include intrusions to increase the number of observations per cell.

We first examined IRTs as a function of list length (5 vs. 7 vs. 9 vs. 11) and block type (open vs. closed) for each level of number recalled for each participant. Since many participants did not output a full range of number recalled levels across both open- and closed-interval blocks, pairwise comparisons only included data from subjects that could contribute to both levels of block type for a given list length and number recalled level. For example, it was possible that a participant recalled three items in two separate trials with a list length of 7 of the open-interval block and never recalled three items in any trials with a list length of 7 of the closed-interval block. Because of this variation in observations, we only report IRT averages when at least 10 subjects contributed data to both open and closed intervals for a given list length. Also, to be consistent with the simulation findings reported above, Figure 5 only includes data for 3, 4, 5, 6, and 7 items recalled.

Figure 5A illustrates the IRTs for open and closed blocks as a function of number recalled, block type, and list length. Upon visual inspection, two patterns that separate open and closed blocks are apparent: First, a more linear trend is seen for IRTs of open-interval trials (left panels) than for those of closed-interval trials (right panels), which instead follow a more hyperbolic trend (akin to that seen in Rohrer, 1996). Similar results were found using SAM: By assuming various list lengths, we simulated IRTs as a function of number of failures (10 or 20) for the list lengths of 5, 7, 9, and 11, consistent with the design of the present study. Figure 5B illustrates that assuming a larger stopping threshold (e.g., 20 failures; right panels), the final IRT is exaggerated across all list lengths compared to cases when fewer failures are needed to terminate search (e.g., 10; left panels). Note also that this pattern holds regardless of the number of items recalled—a factor that is plotted as separate lines within each panel of Figure 5A—and list length. Moreover, by examining and comparing the leftmost panels of Figures 5A and 5B, we see a consistent trend across the simulation patterns and observed data, where lower stopping thresholds (as are expected within open-interval blocks) yield flatter IRT curves than

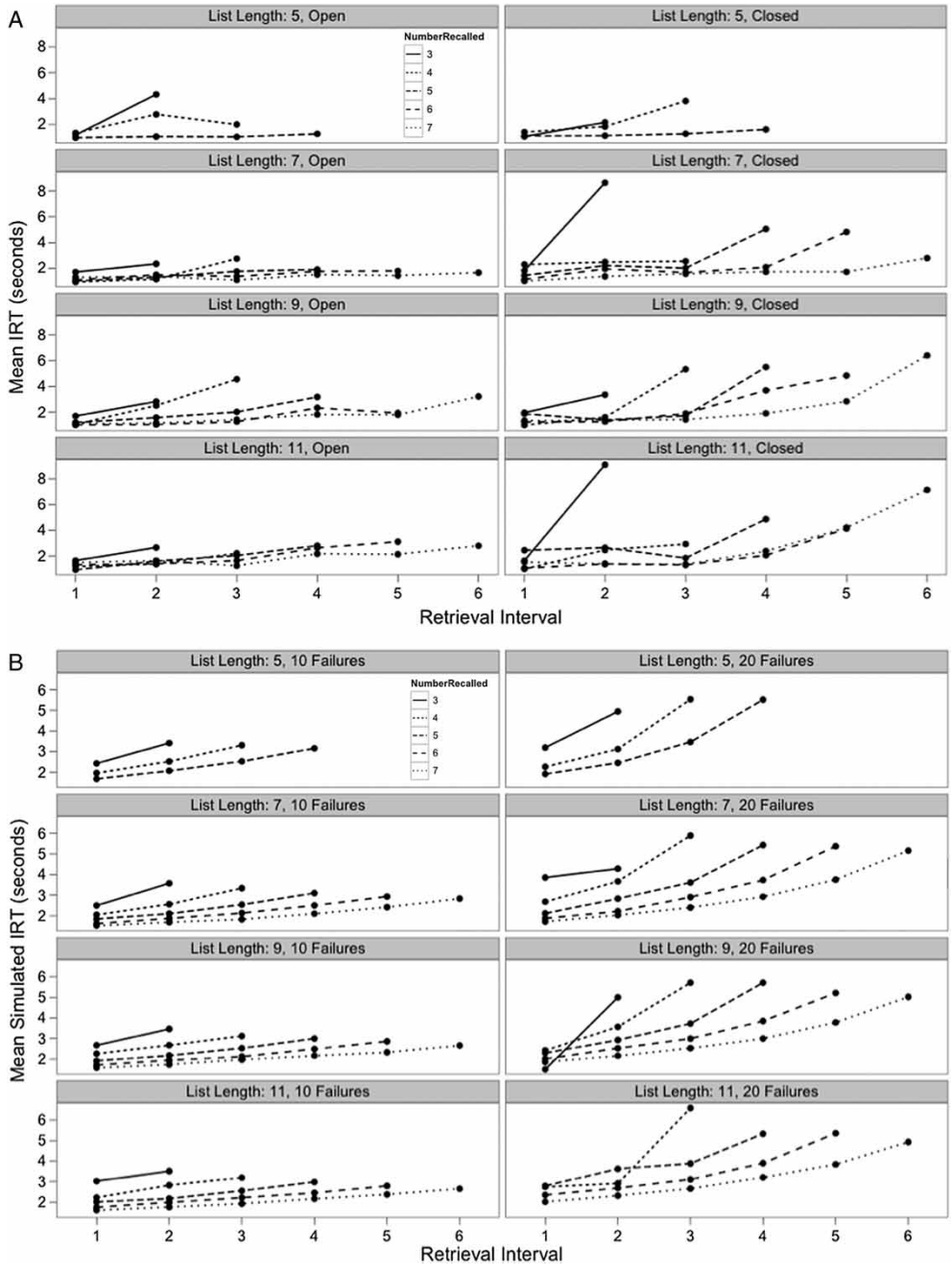

Figure 5. *(A) Mean interretrieval times (IRTs) plotted as a function of number recalled (3, 4, 5, 6, and 7 items retrieved) for each block type (open in left panels and closed in right panels) and list lengths 5, 7, 9, and 11 in each row. (B) Simulation results using a random sampling-with-replacement model with a recovery threshold consistent with the probability of recovery implemented by the SAM model. Activation patterns were all set to 1 (note that this uniform activation did not deviate from cases where a variety of patterns were sampled, as in Figure 3). Mean IRT results for 10 and 20 failures are illustrated in the left and right panels, respectively, for list lengths of 5, 7, 9, or 11 items in each row.*

Table 4. *Mean and standard deviation for final interretrieval time (Last IRT), second-to-last IRT (Last-1), third-to-last IRT (Last-2), and the fourth-to-last IRT (Last-3) for each block type and list length*

		Open		*Closed*		*Test of block type*	
IRT	*List length*	M	SD	M	SD	t	*BF*
Last IRT	5	1.62	1.48	1.73	2.24	0.25	0.08
	7	1.60	1.17	3.12	3.27	2.82**	3.22
	9	2.70	2.37	3.55	3.41	2.44*	1.35
	11	2.50	2.02	4.19	2.97	3.07**	5.95
Last-1 IRT	5	1.33	1.24	1.23	0.71	–0.48	0.10
	7	1.36	0.75	1.75	1.01	2.13*	0.74
	9	1.93	1.63	1.92	1.32	–0.01	0.08
	11	1.81	1.02	2.31	1.48	1.91*	0.51
Last-2 IRT	5	1.10	0.45	1.08	0.39	–0.22	0.09
	7	1.09	0.45	1.63	1.06	3.18**	7.80
	9	1.48	1.10	1.43	0.83	–0.26	0.09
	11	1.46	0.83	1.56	0.84	0.55	0.96
Last-3 IRT	5	0.96	0.43	1.11	0.49	1.42	0.24
	7	1.13	0.61	1.57	1.40	1.91	0.57
	9	1.28	0.56	1.31	0.77	0.23	0.09
	11	1.26	0.62	1.50	0.75	1.68	0.34

Note: Times in s. IRT = interretrieval time. Block type: open versus closed. List length: 5, 7, 9, 11. BF = Bayes factor.
*$p < .05$. **$p < .01$. ***$p < .001$.

high thresholds, which are anticipated for closed-interval blocks (see the rightmost panels of Figures 5A and 5B). One discrepancy, however, involves IRTs at list length 5: The model predicts a slight upward inflection on the last IRT of the closed block, an effect that we do not see in the observed data. Given that recalling five items is a relatively easy task for most participants, that the closed block follows the same trajectory as the open block when only five items must be recalled may not be terribly surprising. Rather, this might reflect a case in which a smaller stopping threshold is utilized in the closed block because few failures are made when fewer items must be retrieved.

Especially important to relationship between interretrieval time and the retrieval interval is the last IRT (see Figure 6). Although there was not an interactive effect of block type and list length for the last IRT, $F(1, 375) = 3.230$, $p > .07$, there was a main effect of block type, $F(1, 96) = 8.399$, $p < .01$; BF = 5.635, such that closed-interval trials led to longer final IRTs than open-interval trials. There was also a main effect of list length, where longer final IRTs accompanied longer list lengths, $F(1, 194) = 32.21$, $p < .001$; BF > 100; see Table 3. Corrected pairwise comparisons of block type for the last IRT at each list length revealed in the following pattern: List length of 5 did not show a reliable difference for last IRT, but list lengths greater than 5 resulted in a clear effect of block (see Table 4). Upon reversing the retrieval interval and examining the last-1, last-2, and last-3 IRTs, we observed no reliable block type and list length interactions (ps > .19), but all three measures demonstrated significant main effects of list length (Fs > 6.26, ps < .01, BFs > 20.87), and only the last-3 IRT resulted in a main effect of block type, $F(1, 94) = 4.444$, $p < .05$, BF = 0.926; other IRT measures: ps > .08; see Table 4 for a comprehensive breakdown of the means for each IRT measure).

Furthermore, the pattern of IRTs plotted in Figure 6 follows that predicted by the model as shown in Figure 3B. In particular, assuming that

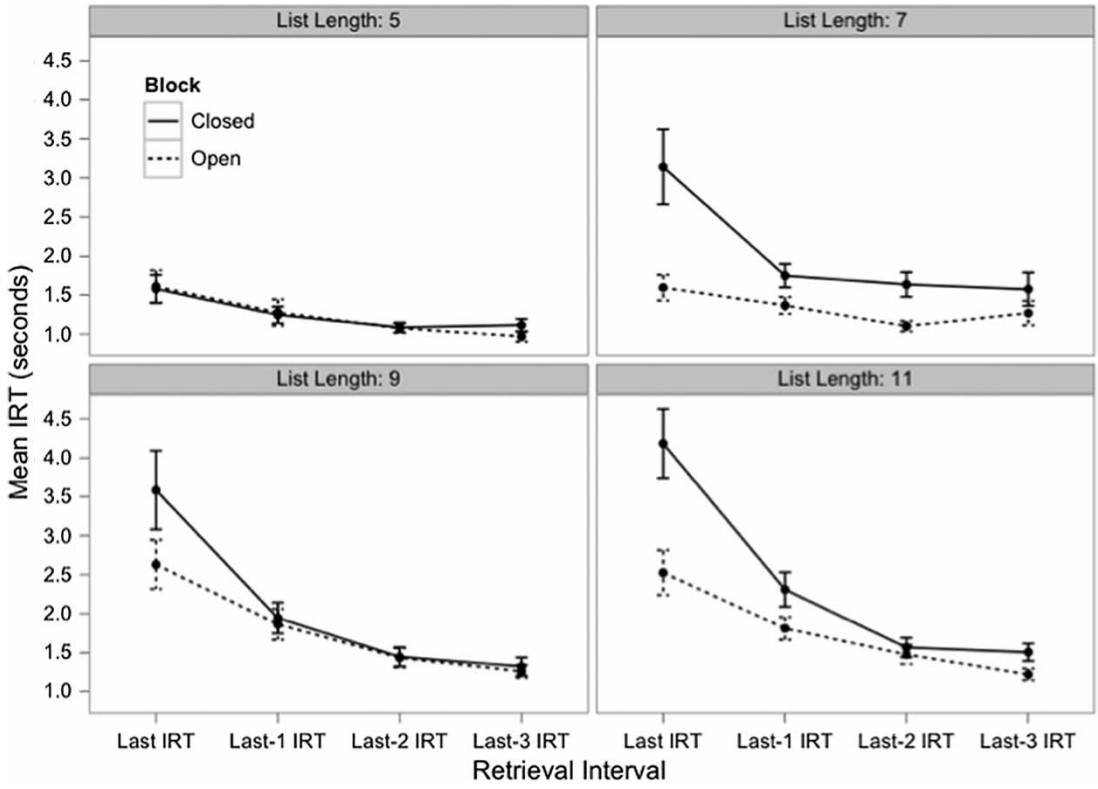

Figure 6. *Mean retrieval times as a function of interretrieval time (IRT) position (final, second-to-last, third-to-last, and fourth-to-last IRTs) by block type for each list length. Error bars = ±1 standard error of the mean.*

participants in the open-interval (relative to the closed-interval) block adopt a lower threshold of number of failures, IRT patterns in the open-interval condition show less curvature than what is predicted by the IRT function (Equation 1). Taken together, despite that there were no significant differences in number recalled (albeit a small, statistically unreliable advantage for the closed block), there do appear to be differences in the temporal dynamics between open- and closed-interval blocks, an effect mostly driven by the final IRT. Specifically, the observed block type differences reflect the comparison of the previous open-interval data (see Figure 2 depicting patterns from Dougherty & Harbison, 2007, and Harbison et al., 2009) with data from closed-interval experiments (see Rohrer, 1996), as well as the predictions of the sampling-with-replacement models detailed above. Namely, the closed block results in similar IRT functions to the simulations assuming larger stopping thresholds. In addition to list length and number recalled, IRTs are also influenced by stopping threshold, a characteristic that is probably sensitive to retrieval interval type.

In addition to examining the mean IRTs for block type and list length, we also conducted a model-based analysis of the IRT functions. Rohrer and Wixted (1994) showed that Equation 1 (above) provided a reasonable fit to IRT data collected using a closed-interval design. We used this equation to estimate retrieval latencies (τ) separately for the open- and closed-interval blocks at each list length for individual participants. Tau provides an index of IRT magnitude and, thus, primarily reflects the magnitude of later IRTs, because according to Equation 1, these later IRTs are the longest. An

Table 5. *Mean and standard deviation for retrieval latency and modeling fit for each block type collapsed across list length (see upper two rows) and each list length collapsed across block type (see bottom four rows)*

		Retrieval latency (τ)		*Modelling fit (RMSE)*	
Block type	*List length*	M	SD	M	SD
Open	All	3.26	1.55	0.87	0.35
Closed	All	4.73	2.14	0.98	0.42
All	5	2.41	1.99	0.81	0.24
All	7	3.17	2.51	0.92	0.29
All	9	4.28	3.38	0.96	0.31
All	11	4.25	2.84	1.00	0.34

Note: Retrieval latency in s. IRT = interretrieval time. Block type: open versus closed. List length: 5, 7, 9, 11. RMSE = root mean square error.

individual fit was obtained for each subject. The interaction of block type and list length did not reach significance, $F(1, 352) = 2.260$, $p > .13$; however, we observed main effects of block type, $F(1, 94) = 16.481$, $p < .001$, BF > 100, and list length, $F(1, 194) = 16.481$, $p < .001$, BF > 100. The average retrieval latency was shorter for open-interval than for closed-interval trials (see Figure 7 and Table 5), indicating that participants spent less time searching memory between retrievals in the open- than in the closed-interval condition. As shown in Figure 6, much of this reduction is due to the final IRT. The average latency generally increased as a function of list length (see Figure 7 and Table 5). Indeed, pairwise comparisons of block for τ at each list length corroborated this finding at list lengths 7, 9, and 11 (ts > 1.94, ps < .05, BFs > 0.49), but not at length 5, $t(42) = 1.819$, $p > .07$, BF = 0.418.

As an aside, the average modelling fit indices, or root mean square errors (RMSEs) suggest that Equation 1 better accounts for the closed-interval than the open-interval condition, $F(1, 94) = 6.572$, $p < .05$, BF = 2.469. The RMSEs were also larger for longer list lengths, $F(1, 190) = 10.178$, $p < .01$, perhaps indicating that Equation 1 provides a better characterization of IRT functions when not all items are easily recalled, which is generally the case at longer list lengths. This explanation is also consistent with the predictions related to number of failures in the simulations described above. Although the interaction of block type and list length were not reliable, $F(1, 352) = 0.816$, $p > .36$, the main effects of both factors indicate that the closed-interval trials result in longer estimated retrieval latencies as given by later retrieval intervals (the last IRT) and thus provide a better fit of the IRT function given by Equation 1. Finally, pairwise comparisons of block type for the RMSEs at each list length patterned with the comparisons reported above for last IRT (see Figure 7 and Table 5). To summarize, list length of 5 did not show a reliable difference for last IRT, $t(43) = 0.342$, $p > .73$, BF = 0.089, but list lengths greater than 5 resulted in effects of block (ts > 2.074, ps < .05, BFs > 0.68).

Discussion

The present experiment directly compared a closed-interval design, in which the experimenter determines the length of the retrieval interval, to an open-interval design, in which participants are allowed to terminate their own memory search. Though number of items recalled did not vary significantly across retrieval designs, the IRT functions did: Compared to their closed-interval counterparts, open-interval trials resulted in overall shorter final average IRTs, lower estimated mean retrieval latencies (τ), and nominally poorer fits to a classic IRT function (Equation 1), a set of effects that generally held for all list lengths greater than 5 items.

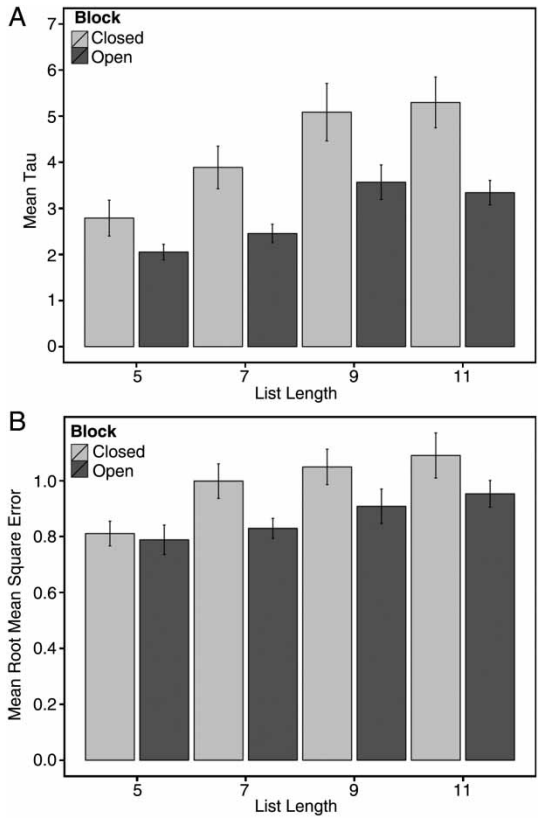

Figure 7. *Mean modeled retrieval latencies (τ panel A, top) and mean model fits (root mean square errors; panel B, bottom) as predicted by Equation 1 plotted as a function of list length for each block type.*

What do these differences in the IRTs mean? According to the simulations presented above, these results suggest a difference in the stopping threshold between open- and closed-interval retrieval. As shown in Figure 3, the relative-strength model, when equipped with the stopping rule supported by existing patterns (see Harbison et al., 2009; Unsworth et al., 2011), predicts a positive correlation between the final IRT and the stopping threshold. To explain the present results, we must assume that a stopping threshold changes; that is, in the absence of a stopping rule, the model is liable to show exaggerated final IRTs as it searches until all list items are recovered. In other words, the present simulations suggest that a stopping rule with variable thresholds must be included to capture the retrieval interval effect: When the stopping threshold is sufficiently large (e.g., 20–30 failures), the final IRT is also large, and the IRT predictions are consistent with Equation 1; specifically, the intercept of the inverse of the IRTs is 0 when plotted in reverse order (Rohrer, 1996). However, when the stopping threshold is set to smaller values (e.g., 10 failures), the final IRT is smaller, decreasing the predicted slope (or $1/\tau$) and increasing the intercept. Thus, with small stopping thresholds, the results are similar to the open-interval design and inconsistent with Equation 1. These patterns are consistent with our hypothesis that participants adopt a larger stopping threshold when the experimenter allots a long interval for retrieval in a closed-interval design than the threshold that is adopted when termination is under the control of the participant. Interestingly, this effect holds despite the open-interval paradigm actually allowing for *more* time for retrieval (theoretically infinite) than the closed-interval design, suggesting that stopping thresholds are actually reflective of the experimental paradigm, not the time needed for retrieval.

A potential strength of the closed-interval design is that it leads to an increased stopping threshold, priming participants to search memory longer and possibly for the entire predetermined retrieval interval. Consistent with this, participants often reach asymptotic levels of recall when provided ample time to retrieve items (Rohrer, 1996; Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). If measuring search performance for a set interval is an experimental goal, then the participants searching for more of that interval is certainly positive, given that recall rates and intrusions are aptly captured during closed intervals. However, in using such a design, it is difficult to ascertain exactly how long participants spend in search, other than knowing that it extends beyond the time of the final item retrieved. Perhaps, under the proper experimental demands, stopping rules may be measured using a closed-interval design wherein subjects are motivated to continue searching throughout the entirety of the retrieval period by being rewarded for every item

produced. Evidence for this stems from recent work demonstrating that employing a cost/benefit structure influences retrieval dynamics especially when correct retrievals are rewarded (Davelaar, Yu, Harbison, Hussey, & Dougherty, 2013). Despite this, another possible drawback of the closed-interval design is that under certain conditions, it may have less ecological validity than an open-ended paradigm. For example, people typically do not have a time limit (e.g., 45 s) as explicitly defined as that often provided in recall studies while searching their memories during everyday tasks. This is not to say that closed intervals are not encountered from day to day; indeed, time pressure often prompts us to strategically control aspects of our retrieval perhaps similar to that used during a closed interval. Within our rubric, time pressure should impact the threshold for deciding when to terminate search. In other words, retrieval is a dynamic process that changes as a function of task demands, and this often forces people to adopt different mechanisms to search memory. Our simulation work suggests that this is probably the case.

As a result, when making predictions about general memory use and about what role memory serves in other cognitive activities (e.g., judgement and problem solving), it is important to note that results from the closed design might not be the most representative. The present research indicates that the IRT function (given by Equation 1) is one such finding that does not appear to transfer beyond the closed-interval design, perhaps suggesting a reason for the lack of generalization. Importantly, however, although the present research indicates that the IRT function (given by Equation 1) does not appear to transfer beyond the closed-interval design, one detail to highlight is that Rohrer and Wixted's (1994) original results were based on a model that did not include a stopping rule. As the current findings indicate, the inclusion of such a threshold may be the factor driving retrieval interval effects. Additionally, Equation 1 was not originally fitted to a data set derived from variable list lengths; that is, this model did not predict effects associated with list length. That we show better fits for closed blocks at longer list lengths is a potential by-product of this important difference. One way to overcome this issue is to design a study to examine the effects of block type with list length constant.

Experimentally, the present study has important implications for researchers designing free-recall studies in terms of deciding precisely how long to allow participants to retrieve items. Knowledge of how recall is affected by the time allotted within an experimental trial is critical when deciding whether to prescribe a recall time —and if so, how long—or to leave the decision entirely up to the participant. The findings of the present study may help to identify how to make future free-recall studies more experimentally efficient. In particular, we demonstrate that providing participants with a 45-s recall period resulted in a small, but nonsignificant increase in the number of items recalled compared to the open-interval condition (which on average lasted 15.44 s), even in the face of varying list lengths (9.75 s for short, 5-word lists and 18.47 s for longer, 11-word lists). This suggests that the length of experiments utilizing closed-interval trials may be truncated to save on the time required to collect data. On the other hand, experimenters who choose to implement open-interval recall designs may need to consider whether little change in number recalled (compared to a closed-interval) is occurring at a cost to the quality of recall timing. To this end, there are important theoretical considerations when deciding on how long to allow participants to retrieve items from memory: For example, Roberts (1972) demonstrated that recall time was proportional to the number of words in a study list. By prescribing additional recall time (a natural by-product of the closed-interval design), we build on this finding by demonstrating that the final IRT increased with the number of to-be-remembered list items (see Figure 6), but did so systematically so as to maintain a difference between open and closed blocks across varying list lengths.

Next, we wish to point out that in several analyses the list length factor was shown to make a difference. Initially, we employed the random list length

procedure in order to prevent idiosyncratic retrieval strategies based on counting due to knowledge of the total number of words in each list. The main effect of list length on total number of items recalled, previous-list intrusions, exit latency, time-to-last, total time, and last IRT is in line with previous work (see Harbison et al., 2009; Murdock, 1962; Roberts, 1972). Importantly, none of the analyses reported here revealed an interaction between list length and block type, although most of the list-length effects on temporal measures were mediated by the differential number of items recalled: The shortest list length (i.e., 5) might evoke a difference in retrieval strategy, given by the lack of block effect at this list length. In a series of papers, Ward and colleagues (Grenfell-Essam & Ward, 2012; Grenfell-Essam, Ward, & Tan, 2013; Ward, Tan, & Grenfell-Essam, 2010) varied list length from 1 to 15 in a free recall paradigm and noticed that output order was sensitive to list length, such that shorter list lengths resulted in participants reporting items in a forward order (similar to serial recall), whereas items were reported in a more typical last-items-first manner at longer list lengths. Although we used a delayed free-recall procedure instead of their immediate free-recall tasks, differences in output order may influence the temporal measures. In a computational study comparing IRTs in serial recall and free recall, Davelaar (2007) showed that at a given list length, serial recall produces shorter IRTs than free recall. It is, thus, not inconceivable that in the present study, list length 5 induced a different retrieval strategy from those used at list lengths 7, 9, and 11.

One possibility that may account for the observed block type effects may have to do with different strategies implemented for each block. This is to say that perhaps the instructions provided to participants (see Method/Procedure above) during the closed block did not deter them from continuing search in a new way after outputting all initial items. Because participants were not explicitly told to stop and wait for the retrieval period to expire following any natural stopping points prior to the end of the 45-s interval, it is possible that separate strategies may have been incidentally developed during the open and closed retrieval blocks. Although we cannot rule out these possibilities entirely given the present design, we can point to a few patterns that suggest otherwise. First, and perhaps most obvious, since number recalled—or number correctly recalled—did not vary significantly as a function of block type at any list length ($ps > .30$), any strategy with the goal to output *more* items would not be different across both blocks (however, note that the small, consistent increase in number of items recalled for the closed block may be due to a real effect that we are failing to detect statistically). Perhaps related, the subtle, but non-significant trend for fewer overall recalls in the open-ended condition may be consistent with an interpretation favouring the latent encouragement to continue searching memory when more time than needed is available for retrieval (as is the case in the closed-interval block). Indeed, by this interpretation, subjects may engage in search longer than they otherwise might prefer, given by a decision they would signal by truncating an open-interval trial. As a result, longer final IRTs would reflect cases when additional items are incidentally output in the closed block.

Next, not all list lengths showed the effect of block type for last IRT: At list lengths longer than 5, we did observe an effect of block, such that the closed block had significantly longer last IRTs than the open block; however, at the shortest list length (5 words), we did not see this effect, perhaps (as suggested above) pointing to a strategy change that was sensitive to the number of items in a list. Thus, it is plausible that in the closed block when there are fewer items to remember, one strategy is implemented, but a separate strategy is used in all other instances. Regardless, if a separate strategy can account for performance differences seen across block, it is probably driven by a number of failures stopping rule given by the tight mapping between the observed data and the predictions generated by the simulations presented earlier.

In closing, the present research illustrates that providing participants with a closed-interval design yields systematic differences in the temporal characteristics of memory retrieval compared to the open-interval design. Our data and analyses suggest

that one likely cause of this difference is the threshold for terminating memory search: Participants are assumed to have a larger threshold on their stopping rule during a closed-interval task. The open-interval design allows us to capture this by making the stopping decision explicit to the subject and the experimenter, a factor important for evaluating computational models of memory. Importantly, the latencies associated with stopping decisions appear to provide valuable data for elucidating the mechanisms governing memory search termination. The open interval design is useful when specifically studying stopping rules and timing of retrieval, providing valuable information beyond recall rates and errors.

Original manuscript received 18 December 2012
Accepted revision received 7 May 2013
First published online 5 August 2013

REFERENCES

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38,* 341–380.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79,* 97–123.

Audacity. A Free Digital Audio Editor (Version 1.3) [Software]. Retrieved from http://audacity.sourceforge.net

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33, 497–505. Retrieved from http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

Davelaar, E. J. (2007). Sequential retrieval and inhibition of parallel (re)activated representations: A neurocomputational comparison of competitive queuing and resampling models. *Adaptive Behavior*, *15,* 51–71.

Davelaar, E. J., Yu, E. C., Harbison, J. I., Hussey, E. K., & Dougherty, M. R. (2013). A rational approach to memory search termination. *Cognitive Systems Research*, *24,* 96–103.

Dougherty, M. R., & Harbison, J. I. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33,* 1108–1117.

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *Journal of Memory and Language*, *67,* 106–148.

Grenfell-Essam, R., Ward, G., & Tan, L. (2013). The role of rehearsal on the output order of immediate free recall of short and long lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 317–347.

Harbison, J. I., Davelaar, E. J., Yu, E. C., Hussey, E. K., & Dougherty, M. R. (in press). Intrusions and the decision to terminate memory search. In *Proceedings of the 35th Annual Cognitive Science Society*.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition*, *111,* 416–421.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science*, *19,* 802–808.

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and lag recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28,* 530–540.

Klein, K. A., Addis, K. M., & Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition*, *33,* 833–839.

Morey, R. D., & Rouder, J. N. (2010). BayesFactorPCL package for R: Computation of Bayes factors for simple psychological designs (Version 0.8) [Software]. Retrieved from https://r-forge.r-project.org/R/?group_id=554

Morey, R. D., & Rouder, J. N. (2010). BayesFactorPCL: Computation of Bayes factors for simple psychological designs [Software]. R package Version 0.8.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Verbal Learning and Verbal Behavior*, *64,* 482–488.

Murdock, B. B., & Okada, R. (1970). Inter-response times in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, *86,* 263–267.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116,* 129–156.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, *17,* 132–138.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York, NY: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88,* 93–134.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92* (3), 365–372.

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, *24,* 188–201.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and inter-response time in free recall. *Memory & Cognition*, *22,* 511–524.

Rouder, J. N., Morey, R., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Rouder, J. N., Speckman, P., Sun, D., Morey, R., & Iverson, G. J. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16,* 225–237.

Shiffrin, R. M. (1970). Memory search. In D. Norman (Ed.), *Models of human memory* (pp. 375–447). New York, NY: Academic Press.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155–185.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Factors that influence search termination decisions in free recall: An examination of response type and confidence. *Acta Psychologica*, *138,* 19–29.

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36,* 1207–1241.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioral Research Methods, Instruments and Computers*, *20,* 6–11.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1,* 89–106.