Self-Terminated vs. Experimenter-Terminated Memory Search

J. Isaiah Harbison (isaiah.harbison@gmail.com)
Erika K. Hussey (erikahussey@gmail.com)
Michael R. Dougherty (mdougher@umd.edu)
Department of Psychology, University of Maryland at College Park
College Park, MD 20742 USA

Eddy J. Davelaar (eddy.davelaar@gmail.com)
Department of Psychological Sciences, Birkbeck University of London
Malet Street, WC1E 7HX, London, UK

Abstract

In most free-recall experiments, participants are given a preset amount of time to search memory. Recently, several studies have examined retrieval in an open-interval design in which the participant, not the experimenter, determines when to terminate memory search. The present study performs the first direct comparison between participant-terminated and experimenter-terminated retrieval. No difference was found in the number of items retrieved from memory; however, inter-retrieval times (IRTs) did differ, such that the participant-terminated paradigm did not show the hyperbolic function typically found when using the experimenter-determined, closed-interval design. We were able to account for this result by equipping a simple relative sampling model with a memory search stopping rule that assumes that giving participants a pre-set retrieval interval causes them to search longer (and tolerate more search failures) than they would in the open-interval design.

Keywords: memory; free-recall; stopping rules.

Terminating Memory Search

At some point, any search of memory must end. Several recent studies have begun to examine how this decision is made in free recall tasks, where search is often terminated while there remain potentially retrievable items unreported (Dougherty & Harbison, 2007; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Unsworth, Brewer, & Spillers, 2011). These studies used a slightly modified version of the standard list recall paradigm; the only difference is that participants terminate their own memory search. This open-retrieval interval design (henceforth, open-interval design) is in contrast to the standard, closed-retrieval-interval design (closed-interval design) that gives participants a pre-determined retrieval interval. Both the open-interval and closed-interval designs have strengths and weaknesses, but to date, they have not been directly compared. The goal of the present study is to perform this comparison and evaluate how allowing or not allowing participants to terminate their own search influences the variables used to describe memory retrieval.

Open- and Closed-Interval Designs

The difference between the open-interval and closed-interval recall design is illustrated in Figure 1. The closed-interval design, shown in Panel A, is restricted, in that the retrieval interval participants are given is pre-determined by the experimenter. After the interval has expired, search is terminated for the participant by the experimenter or by the software program used for the experiment. One reason this design has been used is that it allows greater focus on basic processes of memory retrieval, attempting to eliminate individual differences in how long participants spend searching memory. However, it is not necessarily the case that participants search during the entire pre-determined interval. In fact, many process models that have been proposed to account for the retrieval results from the closed-interval design have assumed a stopping decision to be part of the memory search process (e.g., Raaijmakers & Shiffrin, 1981). Moreover, the closed-interval design might induce participants to search memory longer than they normally would, potentially leading to results that do not replicate when participants are free to retrieve and self-terminate memory search.

The open-interval design (panel B of Figure 1) gives participants an unlimited amount of time to retrieve. The principle strength of this design is that it allows for the measurement of memory search termination decisions, including the total time spent in search—determined by the participant—and the exit latency, or the time between the final retrieval and the decision to terminate search. Both measures have proven diagnostic for evaluating memory search stopping rules (Harbison et al., 2009). The design also, arguably, has greater ecological validity: Individuals are unlikely to have a fixed external time limit when searching memory during most everyday tasks outside the lab (for an examination of termination decision in response to external demands, see Davelaar, Yu, Harbison, Hussey, & Dougherty, 2012). However, the open-interval design too has potential weaknesses: Self-termination might prime participants to put less effort into retrieval and therefore provide inadequate data for the purposes of theory testing. If participants lacked sufficient motivation to adequately search memory, they might recall fewer items in the open-interval paradigm. However, to the best of our knowledge, this account has not been evaluated. Moreover, as far as we are aware, there has been no comparison between open- and closed-interval designs more generally. In what ways do retrieval data obtained from an open-interval design differ from those obtained in the closed-interval design? And, what can the open-interval design tell us about memory retrieval that cannot be discerned from the close-interval design?

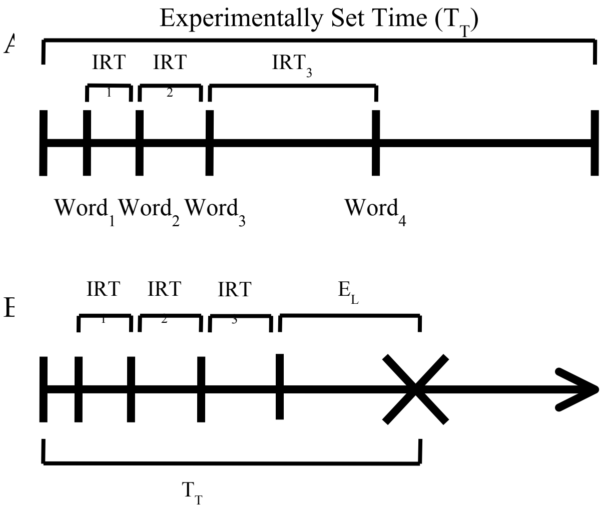

Figure 1. (A) Closed-interval and (B) open-interval retrieval designs, adapted from Harbison & Dougherty (2007). X indicates the time when a participant decides to terminate memory search; hash marks indicate the time associated with the latency onset of words recalled. T_T=Total Time Searching; E_L = Exit Latency; IRT = Inter-Retrieval Time.

Previous Results

As no experiment has yet directly compared the open- and closed-interval designs, it is not possible to draw firm conclusions from existing data. However, there is a wealth of data from closed-interval experiments suggesting specific patterns in the temporal dynamics of recall. For our purposes, we focus on the inter-retrieval times (IRTs), or the time between successive retrievals. IRTs have played an important role in constraining theories of memory retrieval (Rohrer, 1996; Wixted & Rohrer, 1994), and they are generally well described by the equation:

$$IRT_i = \frac{\tau}{N-i} \tag{1}$$

for i=1,2,…N-1, where i is the inter-response interval starting with the interval between the first and second retrieval, 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 as the number of items retrieved increases (Murdock & Okada, 1970; Polyn, Norman, & Kahana, 2009; Raaijmakers & Shiffrin, 1980; Wixted & Rohrer, 1994).

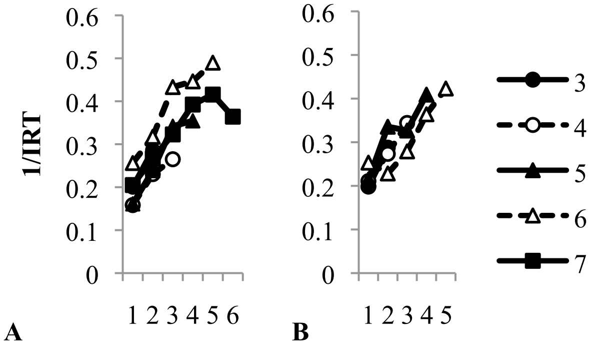

Figure 2. 1/IRT data from (A) Dougherty and Harbison, 2007 and (B) Harbison et al., 2009. The x-axis is the retrieval interval in reverse order, with 1 representing the final interval. The legend indicates the number of items recalled.

A particularly informative way of looking at IRT data is by inverting the IRT (1/IRT) and plotting the results in reverse order along the x-axis, such that the final IRT is in the first position. When plotted this way, Equation 1 predicts that the intercept should be zero. Rohrer (1996) found support for this prediction using the closed-interval design. However, a reanalysis of data from two open-interval experiments (Dougherty & Harbison, 2007; Exp. 1, Harbison et al., 2009) revealed a different pattern: Instead of an intercept of zero, the data from these experiments were best fit by lines with intercepts greater than zero (ranging from .103 to .210), as shown in Figure 2.

What does this result mean for the comparison of open- and closed-interval designs? In particular, could this indicate that something differs in the search process when participants make their own stopping decisions? To answer these questions, we conducted a simulation evaluating the predictions of the relative-strength model of retrieval when stopping decisions were included.

Simulation

For simplicity and ease of comparison with previous research examining IRT results from the closed-interval design, we followed the same simulation procedure as Rohrer (1996). We used the same relative-strength model, which is nearly identical to the sample and recovery processes of the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1981). We also used the same activation patterns tested by Rohrer (1996). The model randomly sampled items based on their relative activation and attempted to recover the sampled item based on its absolute activation. An iteration of random sampling and attempted recovery is referred to as a retrieval attempt and each retrieval attempt could either succeed or fail. For the current simulations, potential retrieval failures are (1) re-sampling an already-outputted item or (2) failing to recover a sampled item due to its absolute activation not meeting the recovery threshold. As we used the same activation patterns as Rohrer (1996), we also used the same recovery threshold, .5.

The one difference between the model tested here and the model used by Rohrer was our use of a stopping rule that terminated the search process. Our model included a stopping rule based on the number of retrieval failures. This stopping rule, native to the SAM model, is the only rule tested so far that has been able to account for both the total time and exit latency data from open-interval experiments (Harbison et al., 2009).

Other than the recovery threshold, the only parameter in the model was the stopping threshold, the number of retrieval failures the model allowed before memory search was terminated. The stopping threshold was varied between 10 and 40 in steps of 10. Each activation pattern was run with each stopping threshold 10,000 times. The dependent variables of interest included the number of items retrieved, the IRTs, and the intercept of the best fitting line for the 1/IRT data.

Table 1. Mean Simulation Results by Activation Pattern.

Act. Pattern	Variable	Stopping Threshold			
		10	20	30	40
1,1,1,1,1,1	Num Ret	5.60	5.94	5.99	6.00
	Last IRT	3.30	5.17	5.80	6.00
	Intercept	.16	.03	0	0
.5,.6,1,1.2,1.2,1.5	Num Ret	5.35	5.79	5.92	5.97
	Last IRT	3.35	5.54	7.03	8.12
	Intercept	.14	0	-.05	-.06
.4,.5,.6,1,1.5,2	Num Ret	4.35	4.78	4.92	4.97
	Last IRT	3.37	5.73	7.43	8.42
	Intercept	.13	-.02	-.07	-.09

Results and Discussion

The results are reported in Table 1 and a representative sample of the model's behavior over various activation patterns. The results of the second activation pattern in Table 1 are shown in Figure 3. First, it should be noted that the variation in stopping threshold did not have a large impact on the mean number of items retrieved (see Figure 3a). However, Figure 3c illustrates that the intercept of the inter-retrieval rates did show substantial variation. Consistent with Rohrer (1996), when the stopping threshold was larger (e.g., 40 failures), the intercept was indeed near zero. However, there was a negative correlation between stopping threshold and the intercept, such that at smaller stopping thresholds the intercept was greater than zero. Therefore, the 1/IRT predictions of the relative-strength model appear to be consistent with closed-interval experiments at greater stopping thresholds and open-interval experiments at smaller stopping thresholds.

Investigating the IRT data more closely, Figure 3b shows the last (or K) IRT, the second to last (K-1) IRT, the third to last (K-2) IRT, and so forth for each stopping threshold value tested. The last IRT showed the greatest variation due to changes in the stopping threshold. Weaker relationships between IRT and stopping threshold were observed the further the IRT was from the final IRT. Importantly, there were large variations in the last IRT even when there were only minute changes in the number of items retrieved. For example, while the mean final IRT more than doubled in size when going from a stopping threshold of 10 to 40, the mean number retrieved varied by only ten percent.

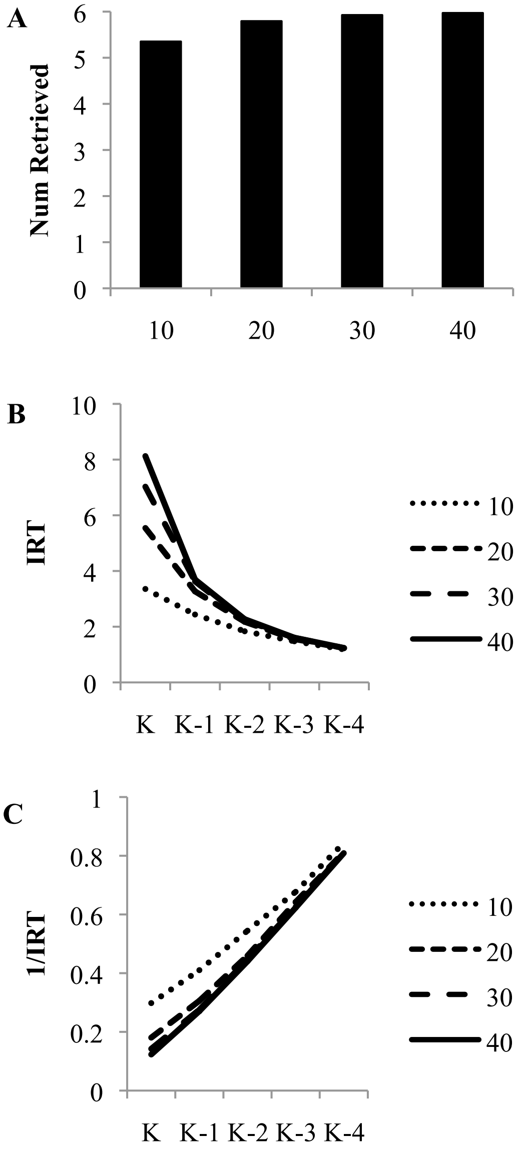

Figure 3. Simulation results for (A) number of items retrieved, (B) IRT, and (C) 1/IRT functions for 4 stopping thresholds (10, 20, 30, or 40 failures).

What is the source of the relationship between IRTs and the stopping threshold? When using the total failures stopping rule, the lower the stopping threshold, the fewer the number of possible attempts for retrieving each item, on average. If an item is not retrieved with a minimal number of failures, it will not be retrieved. For example, if the fifth item is not retrieved before 10 retrieval failures have occurred, then search will terminate with only four items retrieved. Since the probability of retrieval failure increases with each item retrieved, the limit on the number of allowable retrieval failures plays a larger role towards the end of retrieval and particularly for the final item retrieved.

Note that these predictions are particular to the total failure stopping rule. Although not presented here, of the four stopping rules tested in Harbison et al. (2009) the only other stopping rule that correctly predicts the general form of the IRTs is the total time rule, though this rule *cannot* account for the systematic variation in total time spent in search. When equipped with the time-since-last-success or the last-IRT stopping rules, the relative-strength model produces IRT predictions that vary substantially from the results of both open- and closed-interval experiments.

The apparent difference in IRT data between the open- and closed-interval designs are accounted for by the relative-strength model as long as the model includes a stopping rule based on total retrieval failures. According to the model, the difference between the open- and closed-interval designs is expected if the designs induce subjects to use different stopping thresholds. What is left to determine is if the pattern is indeed real. Testing this requires a direct comparison between the designs.

Experiment

Forty-nine participants were randomly assigned into one of two counterbalancing conditions: open-then-closed or closed-then-open. List length was also varied within participant, resulting in a 2 (retrieval block: open vs. closed) x 4 (list length: 5, 7, 9, vs. 11 words) within-subjects design. List length was varied randomly such that all participants were given four study lists of each of four lengths evenly and randomly within each block. List length was systematically varied primarily to prevent participants from learning exactly how many items were on each list and using that information to determine stopping decisions.

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

Procedure For both the open- and closed-interval blocks, participants were given a total of 18 lists consisting of 2 practice trials followed by 16 test trials. During the list presentation of each trial, words were presented sequentially at a rate of 3 seconds per word. Following the list, participants were given a distracter task that consisted of two simple, timed math 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 one second per item. After viewing the final digit of the problem, participants saw an equal sign with a question mark, prompting them to respond with the correct answer.

Participants were then given the opportunity to verbally recall items from the most recent word list. During open-interval trials, participants were told to press the spacebar when they could no longer retrieve additional items from the current memory list; hence, they were given control over when to end the retrieval interval. During closed-interval blocks, participants were given 45 seconds to retrieve the study list items. Based on prior research, we anticipated that a 45-second retrieval interval would provide ample time for most participants to complete the recall task.

All participants 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 recorded in an audio file and hand coded to extract the time-to-word for each item recalled.

Results and Discussion

We conducted Jeffreys-Zellner-Siow (JZS) Bayes-factor (BF) tests to verify the results of each significance test. Moreover, some comparisons reported below 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 null hypothesis over support for the alternative hypothesis, such that coefficients less than 0.3 index strong support for the alternative hypothesis and those greater than 3 index strong support for the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

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 analysis of variance (ANOVA) revealed no effect of Block Order x Block Type for average number of items recalled (F(1,47)=0.002, p>0.96, BF=4.68) or total number of intrusions (F(1,47)=0.442, p>0.50, BF=3.84). Because exit latency cannot be computed for trials in the closed-interval block, we examined the effect of Block Order on exit latency only on open-interval trials, and found no main effect (F(1,47)=2.30, p>0.13, BF=1.70). Finally, there were no reliable effects of Block Order for IRTs at any level of number recalled (p's>0.56). Because these early analyses suggest that there are no effects of Block Order, all subsequent analyses will be collapsed across this factor to increase statistical power.

List Length We next examined the effect of list length on number recalled. Replicating earlier work, we showed a main effect of number recalled (F(3,291)=60.39, p<0.0001).

Scheffé post hoc analyses revealed that longer study lists resulted in the retrieval of additional items (M5=4.14, M7=4.75, M9=5.16, M11=5.92; p's<0.05). Thus, all subsequent analyses will be collapsed across list length.

Number Recalled There was not a significant difference in the number of items retrieved between the closed- (M=5.11) and open-interval trials (M=4.89; t(48)=0.947, p>0.33, BF=5.78), indicating that people's decisions to terminate retrieval did not impact recall rates. Analyses of intrusion and repetition errors are also consistent with this conclusion: The average number of intrusions did not differ as a function of Block Type (Mclosed=0.351; Mopen=0.441; t(48)=-0.972, p>0.33, BF=5.64). An effect of Block Type did not emerge when intrusions were split into 3 types: a) repetitions (t(48)=0.467, p>0.64, BF=8.04); b) extra-list false alarms, or items recalled that were not presented in any prior study lists (t(48)=0.502, p>0.61, BF=7.91); and, c) intra-list false alarms, or items that were incorrectly output that occurred on previous lists (t(48)=1.642, p>0.10, BF=2.47). Also, intrusion rates did not change as a function of time spent in the experiment (p>0.47, but see Unsworth et al., 2011). Given these results, we are comfortable concluding that the open-interval design does not differ from the closed-interval design in terms of number and type recalled.

Exit Latency and Total Time We found that the current exit latency and total time data were consistent with previous results using an open-interval paradigm. 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 (mean γ = -0.139) indicate that participants spend more time deciding 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<0.001). Also consistent with previous data, the total time spent in search was positively correlated with the number recalled (mean γ = 0.268; t(48)=5.775, p<0.0001).

Inter-Retrieval 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). We first examined IRTs as a function of Block Type (open- vs. closed-interval) 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 were only conducted for subjects that could contribute data to both levels of Block Type for a given Number Recalled level. For example, it was possible that a participant recalled three items in two separate trials of the open-interval block and never recalled three items in any trials of the closed-interval block. Because of this variation in observations, we only report IRT averages when at least 15 subjects contributed data to both open- and closed-intervals.

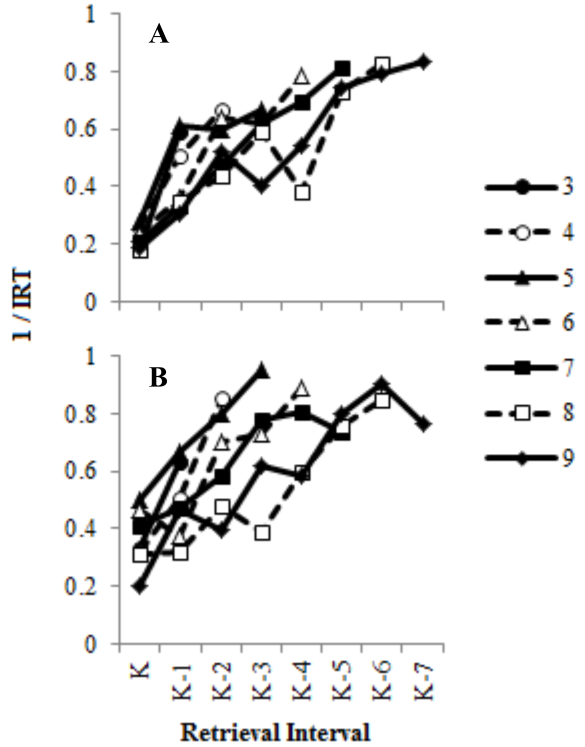

Figure 4. Mean IRTs as a function of Number Recalled for (A) Closed- and (B) Open-interval retrieval trials.

Figure 4 illustrates the 1/IRTs for open- and closed blocks as a function of Number Recalled. Closed-interval intercepts were less than the open-interval intercepts for 6 of the 7 different total number of items retrieved. Especially important to this functional relationship is the final, or K, IRT (see Figure 5); there was a main effect of Block Type on the K IRT (F(1,47)=28.48, p<0.0001, BF=1.53x10-4), such that closed-interval trials (M=4.462s) led to longer final IRTs than open-interval trials (M=2.792s). In fact, the mean K IRT on closed-interval trials was significantly larger than that on open trials for all but two levels of Number Recalled (i.e., 3 and 8, p's > 0.38). A sign test of the binomial relationships for the number recalled of all final IRTs reveals that six of the six comparisons favor closed-retrieval intervals to have longer IRTs than open-interval intervals; a one-tailed test assessing the probability of this pattern occurring yielded a p-value of 0.016. Thus, despite the fact that there were no significant differences in overall number recalled, there do appear to be differences in the temporal dynamics between open- and closed-interval results.

General Discussion

The present experiment directly compared the closed-interval design, in which the experimenter determines the length of the retrieval interval, to the open-interval design, in which participants are allowed to terminate their own memory search. The IRT functions differed between these designs: Compared to their closed-interval counterparts, open-interval trials resulted in overall shorter final average IRTs.

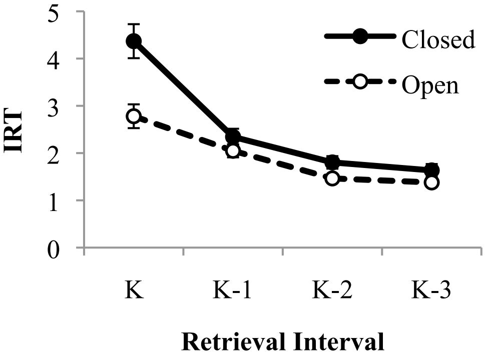

Figure 5. Mean retrieval times for the final (K) IRT, second-to-last (K-1) IRT, third-to-last (K-2) IRT, and fourth-to-last (K-3) IRT.

What do these differences in the IRTs mean? According to the present simulations, these results suggest a difference in the stopping threshold between open- and closed-interval retrieval. As shown in Figure 3B, the relative-strength model, when equipped with the stopping rule supported by the existing data (Harbison et al, 2009), predicts a positive correlation between the final IRT and the stopping threshold. When the stopping threshold is sufficiently large, 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, the final IRT is also smaller. This decreases the predicted slope (or mean retrieval latency, τ) and increases the intercept, producing results similar to those observed under the open-interval design and inconsistent with Equation 1.

The present research finds systematic differences in the temporal characteristics of memory retrieval between open- and closed-interval designs. These differences are predicted by the relative strength sampling model when equipped with a memory search stopping rule if it can be assumed that the type of retrieval interval influences memory search stopping decisions. In the terms of the model, participants appear to use the same stopping rule but set a higher stopping threshold for closed-interval retrieval than for open-interval retrieval. Participants persist in search longer; however, they do not retrieve more items in the closed-interval design than in the open-interval design. The temporal differences were predicted by the model and observed in the data despite no appreciable differences in the number of items retrieved. These results indicate that participants do not in fact terminate search over-quickly in open- relative to closed-interval designs. Furthermore, as participants were able to retrieve the same amount of items in less time, the results suggest that the open-interval design might provide a method to measure not only how memory is searched, but also how efficiently memory can be searched.

Acknowledgments

We gratefully acknowledge NSF (grant BCS-1030831) for support of this research.

References

Davelaar, E. J., Yu, E., Harbison, J. I., Hussey, E., & Dougherty, M. R. (2012). A rational analysis of memory search termination. *Proceedings of the 11th International Conference on Cognitive Modeling.*

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.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition, 111,* 146-421.

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.

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

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., 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.

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

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.