Abstract

Episodic memory, the processes by which information about experienced events is encoded into some long-term store and retrieved, has in recent years been studied in terms of retrieval tasks. Typically, researchers consider how experimental manipulations affect performance on recall tasks or recognition tasks, but rarely both. This dividing line came into being following the discovery of a null list strength effect in recognition. In free recall tasks, memory performance for an item is harmed if memory for other items studied on the list are strengthened, but not so for recognition tasks. Although the models that resulted from this dissociation represent a significant advance, that there is a dissociation at all between models of recognition and models of recall is not a desirable outcome. Efforts should be made to return towards models of memory that can account for a wide variety of test tasks. A consideration of cued recall, a task that incorporates elements of both recognition and free recall, may help advance the field in that regard. To that effect, in a series of experiments, list strength effect in cued recall was measured. In broad terms, the list strength effect in cued recall was found to be very small and largely indistinguishable from a null effect. We apply the REM model to these findings and demonstrate that the inclusion of context as a test cue accounts for these findings. This places cued recall, both in the REM model and the data, as a point of contact between the context-dominant free recall task and the item-dominant single item recognition task along the dimension of a critically diagnostic effect for models of episodic memory.

THE LIST STRENGTH EFFECT IN CUED RECALL: ESTIMATION, IMPLICATIONS, AND MODELS

by

Jack Wilson

B.S. Psychology, College of Charleston, 2011
B.S. Biology, College of Charleston, 2011

Thesis
Submitted in partial fulfillment of the requirements towards a Master of Science in Experimental Psychology

Syracuse University
June, 2015

Table of Contents

List of Tables

List of Figures

Introduction

Episodic memory, often conceived of as a process by which experienced events are encoded into a long-term store, maintained, and retrieved, is often studied in terms of the latter: how manipulations affect performance on some specific set of retrieval tasks. These tasks are commonly either recognition, identifying whether or not an item was experienced in some specific context, or recall, the generation of what items were experienced within some context and/or associated with some other item, but rarely both. This major dividing line is reflected by the models used, some dedicated to recognition tasks (e.g.: Bind-Cue-Decide Model of Episodic Memory, BCD-MEM, Dennis & Humphreys, 2001), some more commonly associated with recall tasks (e.g.: Temporal Context Model, TCM, Howard & Kahana, 2002). This division arose following the decline of the multi-task global matching models.

The global matching models (GMMs, Humphreys, Pike, Bain, & Tehan, 1989) are a class of models that were, for quite some time, able to successfully account for long-term episodic memory effects as observed in a variety of test tasks, including recognition and free recall. These models, which include the Search of Associative Memory (SAM; Raaijmakers & Shiffrin 1981) model and the Theory of Distributed Associative Memory model (TODAM; Murdock, 1982), among others, have the common property that retrieval depends on a "global match," hence the name, to the contents of long-term memory. As such, retrieval depends primarily on memory for the specific to-be-remembered item and other items studied within the same contexts. Specifically, these models made the strong prediction that as the strength of other list items increased, accuracy decreased. This prediction was tested in the list strength paradigm.

The list strength effect (Tulving & Hastie, 1972) proved to be a critical test for this class of models (Shiffrin, Ratcliff, & Clark, 1990). Paradigmatically, participants study and are tested on a “pure strong” list where their memory for all the items are strengthened, a “mixed” list where their memory for only half of the items are strengthened, and a “pure weak” list where their memory for none of the items are strengthened, testing their memory after the presentation of each list. Memory is typically strengthened by repetitions, study time, or encoding task. The effect is then computed by comparing the difference in performance for strong items and weak items on pure versus mixed lists. Three qualitative outcomes are possible. In a positive list strength effect, strong items show better performance on mixed lists than pure lists and weak items show better performance on pure lists than mixed lists. In a negative list strength effect, the reverse occurs: strong items show better performance on pure lists than mixed lists, and weak items show better performance on mixed lists than pure lists. In a null list strength effect, the strength of the other items on the list has no effect on memory performance.

Critically, Ratcliff, Clark and Shiffrin (1990) observed that the measured effect of list strength, when memory is strengthened via spaced repetitions, depends on how memory is tested. In free recall tasks, they observed a positive list strength effect. A smaller positive list strength effect was observed in cued recall. However, when participants were tested with single item recognition, a null list strength effect occurred. In short, when memory for items is strengthened via spaced repetitions, the effect of mixing lists depends upon the test condition.

This finding that the list strength effect was test-dependent even under the same study conditions was a critical problem for the GMMs. For this class of models, strengthening memory for an item increases the activation strength for when that item is probed at test, but also increases the activation strength for when other items are probed to some extent. As such, strengthening an item adds signal in that it increases activation for that item, making it more likely to be recalled or labeled as “old” in a recognition test, but also adds noise, in that increasing the activation of other items can make them more difficult to distinguish from the current item, or, in the case of recognition, from unstudied items presented at test (foils). In SAM, this noise comes in the form of additional strengthening of associations between items and contexts. In TODAM, this noise is a direct byproduct of adding the repeated item multiple times in memory. Regardless, in all cases the outcome is the same. In pure strong lists, the increase in signal is greater than the increase in noise, and performance improves relative to that in pure weak lists. However, in a mixed list, there is more noise than that in a pure weak list and less than that in a pure strong list. The end result is a greater signal-to-noise ratio for strong items on mixed lists than pure lists, and a smaller signal-to-noise ratio for weak items on mixed lists than pure lists. In other words, the GMMs naturally predicted a positive list strength effect, and did so due to the way in which items were encoded and therefore held regardless of the nature of the memory test. Overcoming this effect proved difficult for the GMMs. Shiffrin, Ratcliff, and Clark (1991) provide a detailed review of the difficulties involved.

This effect, along with the observation of a word frequency mirror effect (that, in a recognition paradigm, less common stimuli are more likely to be correctly identified as studied and more likely to be correctly identified as unstudied than more common stimuli) and other empirical findings led to the eventual abandonment of the GMMs as general models of episodic memory. In their place, differentiation and context noise classes of models were created. The differentiation class of models, amongst them the Subjective Likelihood in Memory (SLiM; McClelland & Chapell, 1998) and Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) models, account for the null list strength effect by adding more information about a stimulus to its representation upon strengthening, reducing confusability between items. As such, strengthening memory for dissimilar items does not add additional noise to memory, and in fact removes a small amount. Note that this differentiation process reverses when the information stored in memory is highly similar in both REM and SLiM, see Criss (2006) and Criss and McClelland (2006). This history gives the list strength effect a place of special importance for memory modeling: rarely does one effect cut against a fundamental component of so many models at once or, for that matter, herald a paradigm shift in the field.

While this development represented a significant advance in memory theory, it also led to something of a division in the literature, with models, experiments, and articles primarily focusing on issues pertinent to recognition memory or recall, but rarely both (c.f. Criss & Howard, 2015). These two topics, of course, are not the true interest of the scientific exploration of human long-term episodic memory. As a field, we want models, theories, and approaches that can explain how information is encoded, stored, and retrieved, in the many ways in which these processes can be manipulated. It seems incumbent, then, to return to models of memory that can account for arrays of effects across multiple test tasks.

Cued recall may be a means by which we may return to the prominence of multi-task memory models. This is because, in a fashion, cued recall is a hybrid of recall and recognition tasks. Like recognition, a cue is provided during cued recall and like free recall, an item must be generated from memory and output. Elements of each task are incorporated into a cued recall task. In contrast to simply accounting for effects of recognition with recognition theories and effects of recall with recall theories, a merger of both sets of theories will likely be needed to account for cued recall data. This makes cued recall a natural choice as a task by which multi-task memory models may be evaluated.

Given the list strength effect's history as a critical component in evaluating multi-task models of memory, it is important to have a firm grasp of precisely the effect's size and direction if one wishes to study multitask models though the lens of cued recall. In the following set of experiments, we seek get a more precise measure of the size and direction of the list strength effect in cued recall, and then apply the REM model to the findings. To foreshadow and summarize our findings, we observe a very small positive list strength effect in cued recall that was, in each experiment, indistinguishable from a null effect by hypothesis tests. The list strength effect in cued recall can be accounted for in the REM model if context information is used alongside item information in the cue.

Overview of Experiments

Each of the following experiments employs some variation on the list strength effect paradigm[1]. As such, every participant completed at least three study-test blocks, consisting of a minimum of one pure strong block, one pure weak block, and one mixed block. Each study-test block consisted of a study list followed by a distractor task and a test of memory. In Experiments 1 through 3, this was always a test of cued recall, but in Experiments 4 and 5 participants were additionally given free recall and single item recognition tests for some blocks, with memory task post-cued. Additionally, each study-test block used a unique and randomized set of words, such that no participant would study a word that had appeared in a previous block. In pure weak blocks, all word pairs were presented just once. In pure strong blocks, all word pairs were strengthened via multiple spaced presentations of the pair during study. In mixed blocks, half the pairs were strengthened via repetition, and half were not. Details of the timing and nature of repetition and mixing vary for each experiment and will be described for each individual experiment.

[1] Much of this paper, specifically the methods section, compares and contrasts our methods to that Experiment 6 of Ratcliff, Clark, and Shiffrin (1990), resulting in a ubiquity of in-line citations for said experiment and said publication. To minimize hassle, references to the publication will be initialized to RCS(1990), and references to Experiment 6 will be initialized to RCS(1990)e6.

On cued recall trials for all reported experiments, participants were instructed to type out the word they had studied alongside the test cue, but were given the option to type out the phrase "idk" if they did not know the answer. As such, any given participant's response could be scored as a correct response (matching the target word, with errors in spelling, tense, and pluralization allowed), an intrusion (a response that is incorrect), or a response failure if they responded with an "idk" or left the question blank. In the case of free recall, participants were instructed to type out as many words as they could recall, as such the number of correct responses was the number of unique correct outputs. Intrusions in free recall tasks were computed by the number of outputted items that were not on the studied list. In single item recognition, a yes-no decision was forced for a series of targets and foils intermixed and performance was measured primarily by *d'* with the loglinear correction (Hautus 1995, Stanislaw & Todorov 1999):

$$d' = \Phi\left(\frac{H + \frac{1}{2}}{H_{max} + 1}\right) - \Phi\left(\frac{FA + \frac{1}{2}}{FA_{max} + 1}\right) \tag{1}$$

where H and FA are the number of hits and false alarms, respectively, in a condition, and H_{max} and FA_{max} are the maximum possible hits or false alarms, respectively, in a condition. On mixed lists, because it is impossible to segregate false alarms by strength condition, the total false alarms given on a mixed list will inform the *d'* for both weak and strong conditions on a mixed list, and FA_{max} will be twice that of H_{max} (half of a mixed list is strong, half is weak). Similarly, intrusions in the free recall of mixed lists cannot be segregated. It should also be noted that discriminability measures do not change independently of bias with strength manipulations (Balakrishnan & Ratcliff, 1996; Hirshman, 1995; Stretch & Wixted, 1998). Interpretations of either measure between conditions should therefore be made with some degree of caution.

Analysis Plan

We employ both null hypothesis significance testing (NHST) and Bayesian methods to measure the list strength effect in each of the following experiments. For NHST analyses, a classic repeated measures ANOVA was used to determine significant deviations from the point null hypothesis of a null list strength effect (no interaction of item strength and list type). Bayes factors were then used to find evidence for or against this null hypothesis using the Jeffery-Zellner-Siow prior with an assumed effect size scaling of $r = 1$, as recommended by Rouder et al. (2009, http://pcl.missouri.edu/bf-one-sample). The Bayes factor BF_{01} may be interpreted as the ratio of evidence for the null hypothesis H_0 to the evidence for its alternate H_1. For example, a Bayes Factor of $BF_{01} = 10$ may be thought of as stating that it is 10 times more likely that this data came from a distribution centered around H_0 than H_1. We map Bayes factors to a verbal account for or against H_0 using the modified classification scheme of Jeffreys (1961) as described by Wetzels, et al. (2011).

Finally, we use a difference of difference score as a summary statistic and as the critical contrast of the Bayesian cross-experimental analyses of the magnitude of the list strength effect. If *MW*, *MS*, *PW*, and *PS* refer to the proportion of correctly recalled words (or, in the case of single item recognition, the discriminability index) for mixed weak, mixed strong, pure weak, and pure strong items, the difference of difference score is:

$$DoD = (MS - MW) - (PS - PW) \quad (2)$$

A positive list strength effect is signaled when DoD tends to be greater than 0 and a null list strength effect is signaled when the DoD approximates 0. Note that the list strength effect has previously been computed in two other ways: a ratio of ratios, (*MS*/*MW*)/(*PS*/*PW*), and a difference of ratios (*MS*/*MW* - *PS*/*PW*). Specifically, RCS(1990)e6 used the ratio of ratios as a summary statistic and reported paired-samples *t* tests of *MS*/*MW* and *PS*/*PW* to determine significance (this is mathematically identical to one-sampled *t* tests of differences of ratios). We chose the difference of differences over these other metrics primarily because its computation does not result in loss of data or necessitate corrections when participants fail to make a correct response on some condition.

Experiment 1

Methods

Participants

40 students from Syracuse University completed this experiment to obtain class credit.

Stimuli

Each participant studied 144 words randomly sampled from a pool of 1643 words of letter length 4 to 8 and various word frequencies (KF: range = 1 to 500, mean = 44.8; logHAL: range = 2.89 to 13.7, mean = 8.69). The words were randomly assigned into three groups of 48 words, forming 24 unique word pairs per studied list.

Procedure

In the pure weak block, the study phase consisted of 24 word pairs presented on the screen for 3s. Participants were instructed to place each word pair into a mentally-generated scene during this time. Immediately following presentation of each pair, participants were cued to rate the difficulty of completing this scene generation task on an integer scale of 1 to 9. In the pure strong block, each of the 24 word pairs was presented twice: once during the first half of the list, once during the second half. In the mixed block, half of the word pairs, the strong pairs, were presented twice: once during the first third of the study list, once during the final two-thirds of study. The other half of the pairs, the weak pairs, were presented once during the final two-thirds of the study list. All word pairs were thus studied once during the final 24 study trials and, as such, the mean lag between the final presentation of a word pair and its subsequent test of cued recall was controlled across both list type and pair strength. Between study-test blocks, participants were afforded the opportunity to take a quick break.

Results

Means and standard errors are provided in

Figure *1* for correct responses, and in Table 1 for all three response measures. Data were analyzed in a 2 (pure list vs mixed list) x 2 (strong pair vs weak pair) repeated measures ANOVA. Critically, no list type by item strength (henceforth shortened to list strength) interaction was observed for correct responses, $F(1,39) = 1.05$, $p = .312$. The data provide substantial evidence for the null hypothesis of a null list strength effect, $BF_{01} = 4.89$. A list strength interaction was present for both intrusions, $F(1,39) = 5.10$, $p = .030$, and response failures, $F(1,39) = 5.96$, $p = .019$. Strong pairs were significantly more likely to be correctly recalled than weak pairs, $F(1,39) = 72.03$, $p <$

$.001$, and significantly less likely to be associated with a response failure, $F(1,39) = 46.85$, $p <$ $.001$. No main effect of strength was observed for intrusions, $F(1,39) = 3.54$, $p = .068$. No main effect of list type (mixed vs pure) was observed for correct responses $F(1,39) < 1$, intrusions $F(1,39) = 2.00$, $p = .164$, or response failures $F(1,39) = 6.20$, $p = .436$.

Discussion

This conceptual replication of RCS(1990)e6 found a null list strength effect for cued recall, which does not match the qualitative outcome of prior experiments (RCS(1990)e6, Kahana, Rizzuto, & Schneider, 2005). It should be emphasized at this point that this is only a conceptual replication RCS(1990)e6's test of the list strength effect in cued recall, with the focus

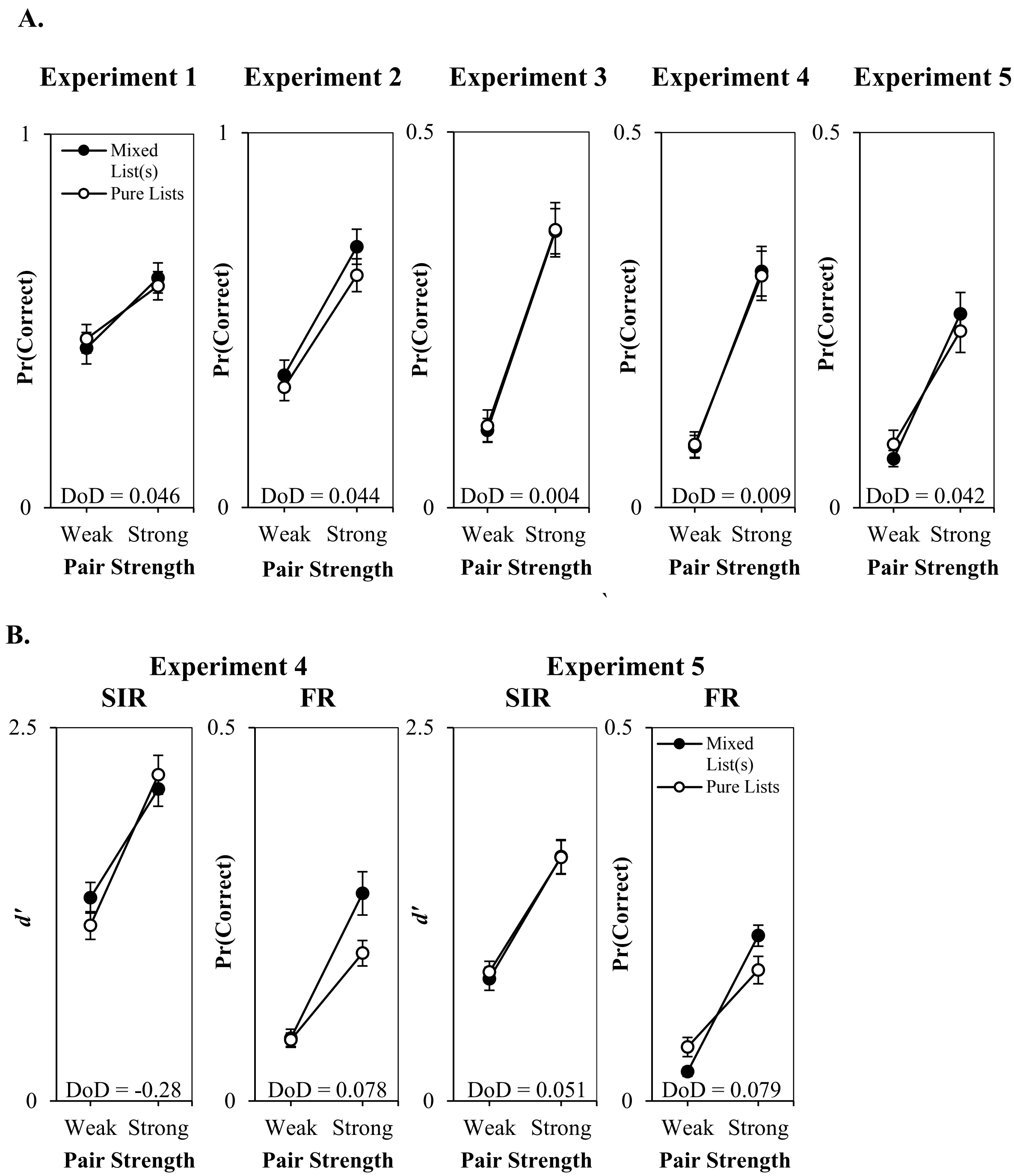


Figure 1: List strength effect means +/-1 SEM for A) cued recall in all five experiments and B) single item recognition and free recall in Experiments 4 and 5. DoD: difference of difference score.

on maintaining the critical elements that should theoretically and on their own produce a positive list strength effect in cued recall—namely, strengthening through spaced repetition and employing mixed and pure study lists. Taken in isolation, one may identify a number of specific methodological differences between this experiment and RCS(1990)e6 that could potentially account for, or complicate, the findings of Experiment 1. Perhaps some differences in the way the lists were mixed, the number of repetitions, the strength of the items, the list length, or encoding strategy are leading to the failure to observe a positive list strength effect here. These are addressed in the experiments that follow.

Table 1

Means and standard deviations for cued recall performance by Experiment and condition

	Mixed		Pure	
Output	Weak	Strong	Weak	Strong
	Experiment 1			
Corrects	0.427 (0.042)	0.615 (0.040)	0.452 (0.038)	0.594 (0.038)
Intrusions	0.127 (0.027)	0.125 (0.026)	0.176 (0.031)	0.115 (0.022)
Don't Knows	0.446 (0.043)	0.260 (0.034)	0.372 (0.031)	0.292 (0.033)
	Experiment 2			
Corrects	0.353 (0.040)	0.696 (0.047)	0.321 (0.036)	0.620 (0.044)
Intrusions	0.199 (0.039)	0.154 (0.035)	0.194 (0.032)	0.159 (0.031)
Don't Knows	0.449 (0.042)	0.151 (0.033)	0.486 (0.037)	0.221 (0.028)
	Experiment 3			
Corrects	0.103 (0.016)	0.368 (0.030)	0.109 (0.021)	0.370 (0.036)
Intrusions	0.223 (0.037)	0.231 (0.032)	0.249 (0.038)	0.229 (0.035)
Don't Knows	0.642 (0.039)	0.402 (0.032)	0.674 (0.039)	0.397 (0.036)
	Experiment 4			
Corrects	0.081 (0.017)	0.315 (0.033)	0.087 (0.015)	0.309 (0.033)
Intrusions	0.213 (0.032)	0.257 (0.034)	0.174 (0.029)	0.237 (0.032)
Don't Knows	0.706 (0.035)	0.428 (0.034)	0.742 (0.031)	0.454 (0.035)
	Experiment 5			
Corrects	0.065 (0.011)	0.258 (0.028)	0.085 (0.019)	0.235 (0.028)
Intrusions	0.143 (0.022)	0.190 (0.026)	0.181 (0.029)	0.189 (0.035)
Don't Knows	0.792 (0.023)	0.552 (0.034)	0.734 (0.030)	0.575 (0.038)

Experiment 2: Repetition, Study Time, and List Length

This next experiment sought to check whether the observed qualitative differences between our Experiment 1 data and that of RCS(1990)e6 were due to differences in methodological details–list length, study time, and number of repetitions–by replicating these three dimensions.

Methods

Participants

39 students from Syracuse University completed this experiment to obtain class credit.

Stimuli

Each participant studied 96 words randomly sampled from the same word pool as used in Experiment 1. The words were randomly divided into three lists of 16 word pairs.

Procedure

The following modifications were made from Experiment 1: Instead of presenting strong pairs twice, strong pairs were presented four times. Study time in all lists was reduced from 3s to 1.25s, and the word pair was removed from the participants' monitor during the rating task to reduce the potential for residual study while rating. The mixed list was constructed as in Experiment 1, such that the entire set of strong words was presented three times, in a random order each time, followed by a fourth presentation of the set mixed with the single presentation of the weak items. Finally the list length was reduced from 24 unique pairs to 16 unique pairs. The list length, study time, and repetition values were chosen because they are identical to RCS(1990)e6.

Results

The statistical analyses used in Experiment 1 were used again in Experiment 2 and the same critical set of effects for correct responses (

Figure *1*, Table 1) was observed. No list strength interactions were present, $F(1,38) < 1$ for all three response measures. The data provide substantial evidence in favor of the null hypothesis of a null list strength effect (correct responses: $BF_{01} = 5.15$). Strong word pairs were associated with more correct responses, $F(1,38) = 131.45$, $p < .001$, fewer intrusions, $F(1,38) = 4.87$, $p = .033$, and fewer response failures $F(1,38) = 114.11$, $p < .001$. No main effects of list type were observed (corrects: $F(1,38) = 2.97$, $p = .093$; intrusions: $F < 1$; response failures: $F(1,38) = 2.73$, $p = .106$).

Discussion

A positive list strength effect is once again absent from the observed effects, suggesting that the absence of the effect in Experiment 1 cannot be attributed to insufficient repetitions or differences in list length or presentation time.

Experiment 3: Mixing, Word Frequency, and Encoding Task

In this experiment, we removed three more methodological differences between our procedure and RCS(1990)e6. First, RCS(1990)e6 used high-frequency words. Second, multiple mixed lists were used in RCS(1990)e6, each one varying from the mixing procedure we used in Experiments 1 and 2. RCS(1990)e6 placed all the weak pairs at the beginning of one list, all the weak pairs at the end of another list, and randomly mixed in the weak pairs throughout the remaining list. They observed a list strength effect when comparing the pure lists to the mixed list where weak pairs were studied last (although not for direct comparisons of the other lists) alongside a list strength effect when the three mixed lists were aggregated. Third, RCS(1990) did not report the use of a ratings task. In addition to the changes made in Experiment 2 (list length, study time, and number of repetitions), we match the word frequency, procedure for building mixed lists, and eliminate the rating task to almost exactly replicate[2] the cued recall tests of RCS(1990)e6.

Methods

Participants

47 students from Syracuse University completed this experiment to obtain class credit.

Stimuli

To match the use of high-frequency words in RCS(1990)e6, 160 words were randomly sampled for each participant from a pool of 800 high-frequency words (KF frequency ≥ 50, logHAL ≥ 9, 4-11 letters long).

Procedure

Within-subjects, participants now ran through 5 study-test blocks including 1 strong, 1 weak, and 3 mixed lists. Rather than providing the opportunity for a quick break, participants were informed between each study-test block that they were moving on to a new set of words and instructed to forget the contents of the previous lists[2]. The three mixed blocks are arranged as follows. In the "weak first" list, all weak pairs are presented fist, followed by strong pairs. In the "weak last" list, all presentations of the strong pairs are presented first, followed by the presentations of the weak pairs. In both of these lists, inter-item spacing between strong pair presentations was controlled to between 4 and 10 intervening presentations, matching the possible range of inter-item spacing for the equivalent lists in RCS(1990)e6. In the "weak

[2] RCS(1990)e6 does not specify the manner in which lists are segregated.

shuffled" list presentations of the weak pairs were randomly shuffled into the arrangement of strong pairs. All other details matched those of Experiment 2.

Results

Data (Figure *1*, Table 1) from the three mixed lists conditions were averaged into a single mixed list condition[3] and the same statistics used in the prior two experiments were performed here. As in Experiment 2, no list strength interaction was observed, $F(1,46) < 1$, for all three response measures. There is substantial evidence in favor of the null hypothesis of a null list strength effect (correct responses: $BF_{01} = 8.69$). Strong items were more likely to elicit correct responses, $F(1,46) = 130.81$, $p < .001$, and less likely to elicit response failures, $F(1,46) = 102.56$, $p < .001$, than weak items, but no main effect of strength was observed on intrusions, $F(1,46) < 1$. No main effect of list was observed, $F(1,46) < 1$ for all three response metrics.

Table 2

Means and standard errors for cued recall response metrics for each mixed list type by strength condition in Experiments 3 and 5

	Weak First		Weak Last		Weak Shuffled	
Output	Weak	Strong	Weak	Strong	Weak	Strong
			Experiment 3			
Corrects	0.073 (0.019)	0.346 (0.042)	0.112 (0.024)	0.354 (0.041)	0.122 (0.026)	0.407 (0.043)
Intrusions	0.213 (0.041)	0.228 (0.037)	0.221 (0.041)	0.237 (0.036)	0.237 (0.042)	0.229 (0.039)
Don't Knows	0.731 (0.041)	0.425 (0.042)	0.668 (0.044)	0.410 (0.042)	0.641 (0.043)	0.364 (0.040)
			Experiment 5			
Corrects	0.076 (0.013)	0.238 (0.033)	0.047 (0.014)	0.292 (0.040)	0.074 (0.018)	0.245 (0.033)
Intrusions	0.130 (0.028)	0.218 (0.032)	0.159 (0.028)	0.181 (0.031)	0.140 (0.026)	0.170 (0.030)
Don't Knows	0.794 (0.028)	0.544 (0.041)	0.794 (0.029)	0.527 (0.043)	0.787 (0.031)	0.586 (0.040)

[3] In a 3 (mixed list types) x 2 (weak vs strong) repeated measures ANOVA, no main effect of list (corrects: $F(2,90) = 1.54$, $p = .221$; intrusions: $F(2,90) < 0$; don't knows: $F(2,90) = 2.40$, $p = .097$) or list strength interaction ($F(2,90) < 0$ for all three measures,) was observed.

Discussion

The failure to observe a positive list strength effect after matching list length, presentation time, study task, measured item strength, and word frequency suggests that these differences between RCS(1990)e6 and Experiments 1 and 2 are an insufficient explanation for the null list strength effect observed, now, across three experiments.

Experiment 4: Test Expectancy

The remaining difference between the prior three experiments and the original positive list strength effect experiment for cued is that cued recall blocks from RCS(1990)e6 were intermixed with blocks of single item recognition and free recall. It has been shown that people are capable of altering their encoding strategies to optimize performance on an anticipated test (Tversky, 1973). Neely and Balota (1980) demonstrated that, for instance, participants who were tested with free recall and expected a test of free recall generally outperformed those who expected a test of recognition instead. Similarly, Hockley and Cristi (1996) have demonstrated that focusing on forming associative bindings at study improves performance on associative recognition tests. The anticipation of a free recall or recognition test may have influenced participants' study strategy during cued recall blocks in a manner that could lead to a positive list strength effect by, for instance, binding studied items more closely to contexts. This experiment specifically tests that hypothesis.

Methods

Participants

41 students from Syracuse University participated in this experiment to obtain class credit.

Stimuli

Each participant studied 288 words, and were additionally exposed to 96 foils during single item recognition testing, all randomly sampled from the same high-frequency word pool as used in Experiment 3.

Procedure

Nine study-test blocks were used, three of which used a cued recall test, three of which used a single item recognition test, and three of which used a free recall test. For item recognition, one member of each studied pair served as a target. For free recall, participants were prompted to recall as many words as they could remember from the studied lists. Participants had 4 minutes to complete this task and could terminate the recall test by typing in a specific phrase to indicate they did not know any more words. The mixed list block, one per test type, were arranged in the same manner as in Experiment 2. All other details were identical to Experiment 3.

Data from cued recall were measured and analyzed as done in Experiments 1 and 2. Statistics were additionally performed on the hit rate, false alarm rate, and the *d'* measures of single item recognition performance, the last of these employing the loglinear correction for extreme rates of hits and false alarms (Hautus 1995, Stanislaw & Todorov 1999). Free recall performance was reduced to two values: a correct response proportion, the proportion of studied items in a condition that were recalled, and an intrusion rate, the number of unstudied items recalled at test divided by the number of studied items on the tested list.

Results

Cued Recall

We performed the same set of analyses for this experiment's cued recall data that we performed for Experiments 1 and 2 (Figure 1, Table 1). No list strength interaction was observed for correct responses, intrusions, or response failures, $F(1,52) < 1$ in each case. The data suggest substantial evidence for the null hypothesis of a null list strength effect (correct responses: $BF_{01} = 6.89$). A main effect of item strength was also observed, with repeated word pairs associated with more correct responses, $F(1,52) = 89.69$, $p < .001$, more intrusions, $F(1,52) = 9.64$, $p = .003$, and fewer response failures, $F(1,52) = 128.75$, $p < .001$.

Single Item Recognition

Discriminability (Figure 1 for *d'*, Table 3) was analyzed with a 2 (mixed vs pure lists) x 2 (strong vs weak items) repeated measures ANOVA. No significant list strength interaction was observed, $F(1,52) = 3.41$, $p = 0.07$. There is anecdotal evidence for the null hypothesis of a null list strength effect, $BF_{01} = 1.84$. This finding was expected, and parallels the null-to-negative list strength effect observed throughout recognition tests in RCS(1990). Strong items were more discriminable than weak items, $F(1,52) = 103.64$, $p < .001$. No main effect of list type on discriminability was observed, $F(1,52) < 1$.

For archival purposes, hits and false alarms (Table 3) were analyzed. Because false alarms on mixed lists cannot be segregated by condition, the hit rates were analyzed with a 2 (mixed vs pure lists) x 2 (strong vs weak items) repeated measures ANOVA and the false alarm rates were analyzed with a 1-way (pure strong vs pure weak vs mixed list) repeated measures ANOVA.

Table 3

Single Item Recognition Performance, by Condition

Response Measure	Mixed		Pure	
	Weak	Strong	Weak	Strong
		Experiment 4		
d'	1.361 (0.103)	2.088 (0.115)	1.175 (0.093)	2.184 (0.130)
Hits	0.611 (0.029)	0.852 (0.023)	0.647 (0.021)	0.802 (0.024)
False Alarms	0.148 (0.022) [a]		0.225 (0.022)	0.115 (0.017)
		Experiment 5		
d'	0.818 (0.076)	1.636 (0.114)	0.865 (0.071)	1.631 (0.113)
Hits	0.429 (0.022)	0.735 (0.022)	0.601 (0.018)	0.708 (0.023)
False Alarms	0.183 (0.020) [a]		0.289 (0.020)	0.166 (0.022)

	Mixed					
	Weak First		Weak Last		Weak Shuffled	
	Weak	Strong	Weak	Strong	Weak	Strong
			Experiment 5			
d'	0.791 (0.087)	1.646 (0.139)	0.778 (0.116)	1.457 (0.141)	0.911 (0.111)	1.821 (0.153)
Hits	0.370 (0.031)	0.696 (0.031)	0.480 (0.030)	0.728 (0.031)	0.445 (0.032)	0.782 (0.033)
False Alarms	0.157 (0.024) [a]		0.227 (0.024) [a]		0.167 (0.027) [a]	

[a]False alarms within a mixed list cannot be segregated by study condition

Strong items had more hits than weak items, $F(1,53) = 124.24$, $p < .001$. A positive list strength interaction was observed for hit rates, $F(1,53) = 4.96$, $p = .03$. List type had a significant effect on false alarm rates, $F(2,106) = 9.83$, $p < .001$, sphericity assumed, with the most false alarms occurring on pure weak lists, and the least occurring on pure strong lists. A post hoc t test reveals that the difference between pure weak and pure strong false alarm rates is significant, $t(53) = 4.78$, $p < .001$.

Table 4

Free Recall Performance, by Experiment and condition

Output	Mixed				Pure			
	Weak		Strong		Weak		Strong	
				Experiment 4				
Corrects	0.084	(0.012)	0.278	(0.029)	0.082	(0.009)	0.198	(0.017)
Intrusions	0.055	(0.013)	[a]		0.051	(0.011)	0.063	(0.012)
				Experiment 5				
Corrects	0.039	(0.006)	0.221	(0.014)	0.072	(0.013)	0.175	(0.018)
Intrusions	0.079	(0.013)	[a]		0.095	(0.029)	0.096	(0.031)

	Mixed											
	Weak First				Weak Last				Weak Shuffled			
	Weak		Strong		Weak		Strong		Weak		Strong	
						Experiment 5						
Corrects	0.024	(0.005)	0.211	(0.022)	0.059	(0.010)	0.208	(0.019)	0.040	(0.006)	0.230	(0.021)
Intrusions	0.073	0.011	[a]		0.090.	0.024	[a]		0.089	0.018	[a]	

[a]Intrusions in free recall of mixed lists cannot be segregated by study condition

Free Recall

Proportion of correct responses (Figure 1, Table 4 also has intrusions) were analyzed, as in cued recall, with a 2 (mixed vs pure lists) x 2 (strong vs weak items) repeated measures ANOVA. Because intrusions cannot be segregated into strong and weak intrusions on mixed lists, these data were analyzed with a 1-way (pure strong vs pure weak vs mixed list) repeated measures ANOVA, as done for false alarms in recognition. As expected, an item strength by list type interaction was observed for correct responses, $F(1,52) = 5.74$, $p = .020$. This is anecdotal evidence against the null hypothesis of a null list strength effect $BF_{01} = 0.64$ ($BF_{10} = 1.56$). Correct responses were significantly greater for strong items than weak items, $F(1,52) = 78.99$, $p < .001$. Mixed lists had significantly more correct responses than pure lists, $F(1,52) = 5.63$, $p = .021$. No effect of list type was observed on intrusions, $F(2,106) < 1$, sphericity assumed.

Discussion

A null list strength effect in cued recall was again observed, consistent with the findings of the prior experiments reported here. Test expectations influencing participants' encoding strategy does not seem to drive the positive list strength effect observed in earlier cued recall experiments. In the single item recognition and the free recall data, we observed a pattern of data consistent, if only anecdotally, with that observed in prior experiments: a null list strength effect and a positive list strength effect, respectively (Malmberg & Shiffrin, 2005, Murnane & Shiffrin, 1991a, 1991b, Ratcliff, McKoon, & Tindall, 1994, RCS(1990), Rose & Sutton, 2006, Sahakyan, Abushanab, Smith, & Gray, 2014).

Experiment 5

As a final test for the list strength effect in cued recall, we replicated RCS(1990)e6 as closely as possible. The only notable methodological differences between this experiment and that of RCS(1990)e6 are the following, so far as we can tell: a different stimuli set was used, that nonetheless matches along the word frequency properties (KF frequency > 50) by RCS(1990); the experiment was completed over two sessions separated by 4-10 days, rather than on three consecutive days; participants were compensated with class credit rather than financially; and lastly, the order of the blocks (after the completion of the three training blocks at the beginning of the first session), kept the same for participants tested in the same test group in RCS(1990)e6, was completely randomized.

Methods

Participants

105 students from Syracuse University participated in this experiment to obtain class credit.

Stimuli

We expanded our pool of words to 1779 high frequency words (KF frequency > 50, logHAL > 9, 4-11 letters long) to accommodate the increased number of stimuli required for this design.

Procedure

Across two sessions (4-10 days apart, separated by a weekend), participants received a total of 24 study-test blocks, 8 each for cued recall, single item recognition, and free recall. For each of the 8 blocks per test type, 5 of them were arranged as they were in Experiment 3 (pure strong, pure weak, weak first (mixed), weak last (mixed), and weak shuffled (mixed). The remaining 3 study-test blocks consisted of a training block, identical in form to the pure weak block, a long list of weak pairs (long weak) consisting of 40 word pairs each presented once, and a short list of strong pairs (short strong) consisting of 10 word pairs presented four times each, with inter-item spacing restricted to between 5 and 15 intervening presentations. Memory for each of these list types was tested with single item recognition, cued recall, and free recall, post-cued for a total of 24 study-test blocks.

During session 1, participants ran though the three training blocks (1 per retrieval task, order randomized) followed by 9 other blocks selected at random. Participants were informed after the three training blocks that they would be tested using those three methods throughout the remainder of the experiment. During session 2, participants ran through the remaining 12 blocks, again in a random order.

Results

90 people completed both sessions of the experiment. Due to technical problems complete data sets exist for only 51 participants and analyses were conducted using only data from those participants.

Cued Recall

Cued recall data (Figure 1, Table 1) was analyzed using multiple repeated measures ANOVAs. A 3 (weak first vs weak last vs weak shuffled) x 2 (weak vs strong) repeated measures ANOVA revealed no list strength interaction for the three response measures, sphericity assumed (correct responses: $F(2,100) = 2.93$, $p = .058$, intrusions: $F(2,100) = 1.91$, $p = .154$; response failures: $F(2,100) = 1.02$, $p = .364$, sphericity assumed in each case). No significant main effects of list type were observed, $F < 1$. We therefore collapse across the three mixed list types and report the calculations from a 2 (mixed vs pure) x 2 (weak vs strong) repeated measures ANOVA.

In the 2 x 2 repeated measures ANOVA, no significant list strength interaction was observed for correct responses, $F(1,50) = 2.72$, $p = .106$, intrusions, $F(1,50) < 1$, or response failures $F(1,50) = 3.10$, $p = .085$. The correct response data gives anecdotal evidence for the null hypothesis of a null list strength effect, $BF_{01} = 2.50$. Strong pairs elicited more correct responses and fewer response failures than weak pairs (corrects: $F(1,50) = 66.09$, $p < .001$; response failures: $F(1,50) = 51.4$, $p < .001$). No significant main effect of strength on intrusions, $F(1,50) = 1.91$, $p = .173$, or main effect of list (corrects and response failures: $F(1,50) > 1$, intrusions: $F(1,50) = 1.46$, $p = .233$) was observed.[4]

[4] To ensure that the reduction in sample size did not harm statistical power in a manner that changes the qualitative outcome of the cued recall statistics, we reran the 2x2 ANOVA using the entire set of participants who had data for all of the five relevant cued recall blocks (pure strong, pure weak, weak first, weak last, weak shuffled). No

Single Item Recognition

We collapsed the *d'* and hit rate data (Figure 1 for *d'*, Table 3 for hits and false alarms as well) to 2 (mixed vs pure) x 2 (weak vs strong) repeated measures ANOVAs and the false alarm rate data to a 1-way (pure weak vs pure strong vs mixed) repeated measures ANOVA, as a 3 (weak first vs weak last vs weak shuffled) x 2 (weak vs strong) repeated measures ANOVA revealed no main effect of mixed list type or list strength interaction for discriminability (main effect of list: $F(2,100) = 2.13$, $p = .124$, sphericity assumed; main effect of strength: $F(1,50) = 121.87$, $p < 0.001$; interaction: $F(2,100) < 1$, sphericity assumed). No significant list strength interaction was observed, $F(1,50) < 1$. This is substantial evidence for the null hypothesis of a null list strength effect, $BF_{01} = 8.09$. Strong words elicited a higher *d'* than weak words, $F(1,50) = 153.7$, $p < .001$. No main effect of list type was observed, $F(1,50) < 1$.

Strong words elicited higher hit rates than weak words, $F(1,50) = 131.72$, $p < .001$. Words on pure strong lists were more likely to elicit hits than those on pure weak lists, $F(1,50) = 20.42$, $p = .001$. A significant list strength interaction was observed for hit rates, $F(1,50) = 34.99$, $p < .001$. List type had a significant effect on false alarm rates, $F(2,196) = 40.46$, $p < .001$, sphericity assumed, with pure strong lists eliciting the most false alarms and pure weak lists eliciting the least.

Free Recall

A 3 (weak first vs weak last vs weak shuffled) x 2 (weak vs strong) repeated measures ANOVA revealed no significant main effect of list, $F(2,100) < 0$, sphericity assumed, or list strength interaction, $F(1,50) < 0$. Free recall correct response rates (Figure 1, Table 2 for intrusions also) were therefore collapsed to a 2 (mixed vs pure) x 2 (weak vs strong) repeated

---

significant list strength interaction was found ($F(1,80) = 2.395$, $p = .126$). This is substantial evidence for the null hypothesis, $BF_{01} = 3.55$.

measures ANOVA and intrusions to a 1-way (pure strong vs pure weak vs mixed) repeated measures ANOVA. A significant list effect interaction was observed for the proportion of correctly recalled words, $F(1,50) = 13.19$, $p = .001$, $BF_{01} = 0.0291$, strong evidence against the null hypothesis of a null list strength effect. Strong words were more likely to be recalled than weak words, $F(1,50) = 106.70$, $p < .001$. No effect of list type was observed was observed either for corrects, $F(1,50) < 1$, or intrusions, $F(2,100) < 1$, sphericity assumed.

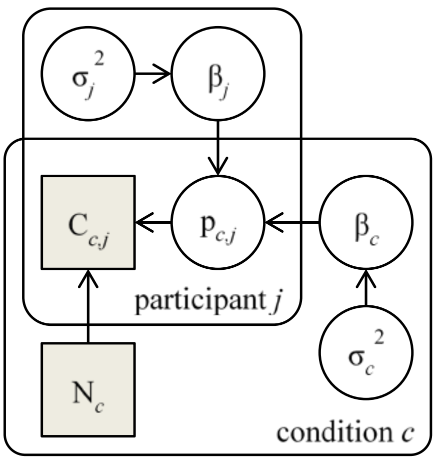
Discussion

The results of this experiment largely match those observed in the prior four experiments. While replicating the positive list strength effect in free recall and the null list strength effect observed in single item recognition, we failed to observe a positive list strength effect in cued recall.

It is perhaps worthy of note that this experiment, the closest replication to that of RCS(1990)e6, contains the most evidence favoring a positive list strength effect over a null effect and therefore comes closest to replicating the outcome of RCS(1990)e6. One possible explanation for this is that both RCS(1990)e6 and this experiment were conducted across multiple days and included a larger number of study-test blocks. In order to retrieve the items studied within a given block, participants must be able to discriminate between those words studied on the relevant block from those from prior, irrelevant blocks. The presence of other blocks and their spacing across multiple sessions may make this more difficult, requiring more attention paid to that information used to discriminate between lists—context—and therefore leading to a more positive list strength effect than otherwise.

Cross-Experiment Analysis

Across five experiments, we have thus far failed to find evidence for a positive list strength effect when participants are tested for cued recall, and in fact found evidence in favor of a null list strength effect beyond anecdotal levels in 4 of the 5 experiments. We next evaluated the magnitude of the list strength effect by a combined analysis of all five experiments. To that effect, we constructed a Bayesian model that is neutral as to the relations between conditions and experiments and that accounts for variation by participant (Figure 2). The model states, simply, that each level of strength on each list of each experiment is its own independent condition denoted *c* in Figure 2. There are 28 conditions in total. Experiments 1,2, and 4 each contribute one mixed strong; mixed weak; pure strong; and pure weak condition. Experiments 3, and 5 each contribute three mixed strong and three mixed weak conditions, reflecting the three mixed list types on those experiments, and one pure strong and pure weak condition each. As such each participant, denoted by *j* in Figure 2, completed at least 4 of these conditions, corresponding to the list by strength conditions in the participant's experiment. A participant in Experiment 1, 2, or 4 completed 4 conditions, while a participant in Experiment 3 or 5 completed 8 conditions. Each participant *j*'s outcome (the total number of correct recalls) on each condition *c* is considered to be binomially-distributed with N_c = number of trials on that condition and a underlying recall probability $p_{c,j}$. Each $p_{c,j}$ is a logistic function



$$
\begin{aligned}
C_{c,j} &\sim \text{Binomial}(N_c, p_{c,j}) \\
p_{c,j} &= (1 + \exp(\beta_c + \beta_j))^{-1} \\
\beta_c &\sim \text{Gaussian}(0, \sigma_c^2) \\
\beta_j &\sim \text{Gaussian}(0, \sigma_j^2) \\
\sigma_j^2 &\sim \text{Gamma}(11, 0.06)^{-1} \\
\sigma_e^2 &\sim \text{Gamma}(11, 0.06)^{-1}
\end{aligned}
$$

Figure 2: Graphical model of the Bayesian cross-experiment analysis for the 5 experiments.

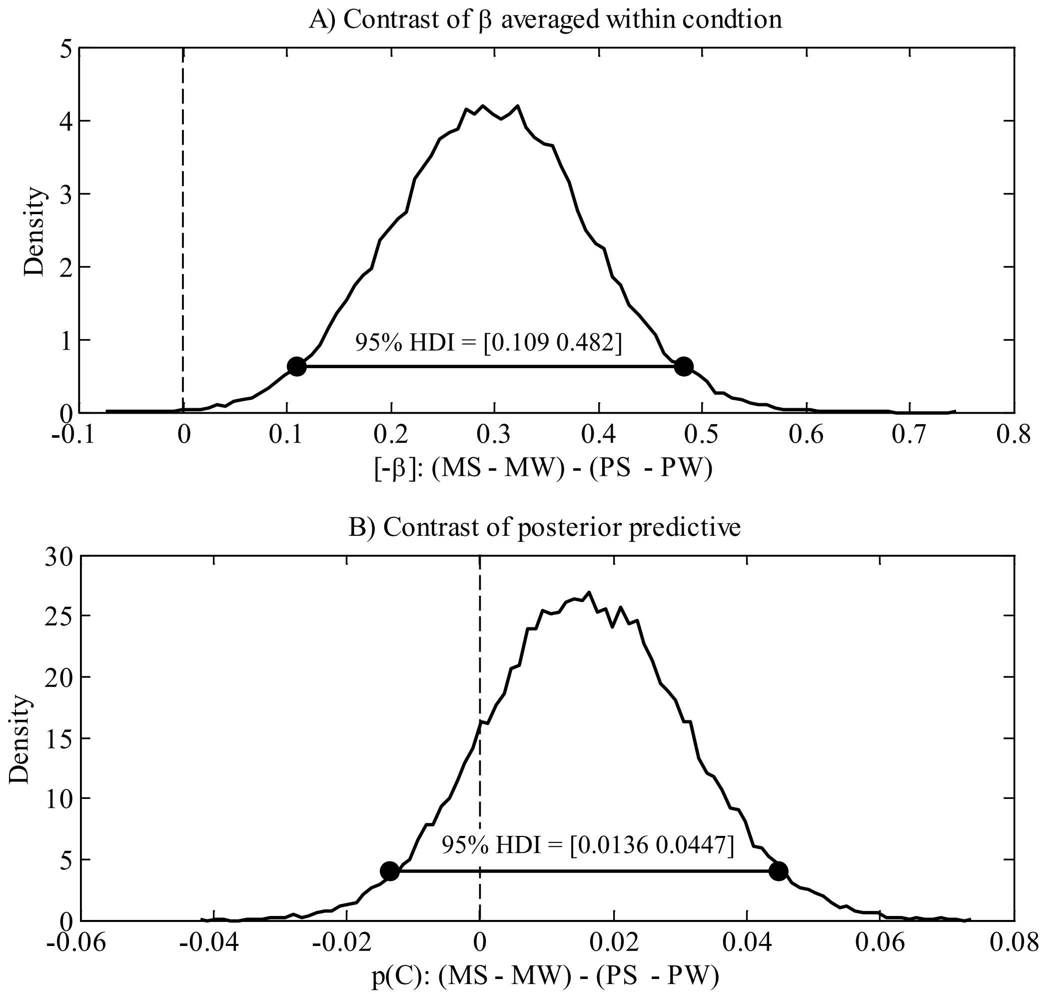


Figure 3: Results from the Bayesian statistical model. Solid curve gives density, dashed line highlights a contrast of 0. (A) Posterior distribution of the contrast by condition and (B) the posterior predictive contrast for correct responses. Contrasts are computed by first averaging the posteriors for all parameters β or predictives *p* within a given list strength condition then finding the difference of differences *DoD* = (*MS* - *MW*) - (*PS* - *PW*).

of $\beta_c + \beta_j$, where β_c is the overall performance across participants on condition *c* and β_j is the overall performance across conditions of participant *j*. To find the list strength effect under this model, simply average the β_c within each of the mixed weak (*MW*), mixed strong (*MS*), pure weak (*PW*), and pure strong (*PS*) conditions and find the contrast *DoD* = (*MS* - *MW*) - (*PS* - *PW*). This contrast is presented in Figure 3A, based off of Runjags simulations computed in R with an effective sample size greater than 1000 per each of the 28 conditions. The 95% highest density interval (HDI) of this contrast excludes zero, suggesting that a positive list strength effect is, in fact, occurring in cued recall tasks.

Additionally, one may transform the parameters β_c into probabilities, $p_c = (1 + \exp(\beta_c))^{-1}$, then average and compute the contrast as done with the underlying parameters. The outcome of this is presented in Figure 3B. The null effect falls with the 95% HDI. That the contrast of the underlying parameters yields a positive list strength effect, but the contrast of the transformation does not, highlights the importance of computing the underlying parameters of the conditions rather than simply taking a naïve contrast of the correct response rates. The list strength effect in cued recall is small enough that it appears almost null when observing along the dimension of correct response rates or, in this case, its posterior predictive distribution.

Discussion

Across five experiments, we observe a null or very small positive list strength effect in cued recall. This is in contrast with both the null-to-negative list strength effect that has been previously observed in recognition and the large, positive list strength effect observed in free recall.

The list strength effect, has, historically, been a diagnostic effect for models of memory. Two prior experiments have demonstrated a positive list strength effect in cued recall (Ratcliff, Clark, & Shiffrin, 1990; Kahana, Rizzuto, & Schneider, 2005), however, our data more clearly place the list strength in cued recall as a very small positive effect that is indistinguishable from a null effect. Now that the size and direction of the effect in cued recall has been more firmly characterized, we turn towards modeling this effect within REM.

The List Strength Effect in Recognition

The null list strength in recognition, according to REM and similar models (e.g. SLiM) is due differentiation during encoding. Additional information in a trace provides more evidence for or against that trace being a representation of some presented stimulus. If two items are distinct, more information in the representation of one provides evidence that it is not a representation of the other. Strongly encoded mismatching traces are therefore less likely to be confused as a matching trace than weakly encoded ones. This is a fundamental property of how information is represented and assessed in the model: so long as two stimuli are distinct and already represented in memory, learning more about one does not add more noise to decisions about the other.

The List Strength Effect in Recall

The positive list strength effect in free recall is driven by competition between how well memory traces match the reinstated context. On pure lists, all the context information, being encoded with approximately equal strength, competes on an equal footing to be retrieved. On a mixed list with spaced repetitions, repeated items have more strongly encoded contexts than weakly encoded pairs. Strong contexts in this case dominate weak contexts in the competition to be retrieved, leading to a greater probability of outputting a strong item and a lesser probability of outputting a weak item on a mixed list, in comparison to a pure list.

One prior implementation of cued recall in REM was outlined by Diller, Nobel and Shiffrin (2001), which prioritized accounting for response time and accuracy in cued recall. The key element of our model, in contrast, is accounting for context as a retrieval cue: we do not consider response times in this paper. Details pertaining to response times, superfluous for the stated purpose, are therefore omitted from the implementation presented here.

Retrieving Effectively Memory

In Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) stimuli are represented as vectors consisting of w features. Those features are distributed geometrically, such that, if g_e is the base rate parameter, the probability of feature v taking value i is

$$\Pr(v = i) = g_e(1 - g_e)^{i-1} \tag{3}$$

Experimental manipulations of word frequency are represented in the model as changes in g_e, where lower frequency words are distributed using a smaller g_e, and higher frequency words are distributed using a larger value of g_e. Each item is represented by a randomly generated vector of features. Context information within a study-test block is assumed for the sake of simplicity to be the same: a single study-test context represents the context at both study and test. Deeper instantiations of context allow for some degree of contextual mismatch between study and test and a drift of context across the course of the experiment (e.g.: Mensink & Raaijmakers, 1989; Criss & Shiffrin 2004b; Lehman & Malmberg, 2013).

Exposure to a stimulus prompts storage of features from that stimulus and the encoding context in episodic memory, a process that is assumed to be both noisy and incomplete. For any given stimulus presented, the odds of storing any given feature into its representative trace is given by u. If a feature is to be stored, it is directly copied from the stimulus with probability c. If it is not copied, then the stored feature takes a value from the geometric distribution with parameter g_s, where g_s is an internal estimate of the shape of the geometric distribution for stimuli, based upon a lifetime of experience. Although it need not be the case, here it is assumed for the purposes of simplicity that $g_e = g_s$. In the event that multiple items are presented simultaneously, such as in paired associate learning, the episodic association between the items is represented as a concatenation of the encoded vectors, although more realistic implementations of the model allow the storage of emergent associative features (e.g., Criss, 2004; Criss & Shiffrin, 2004a; 2005). As such, a list of paired associates would be represented in memory as a list of traces where each trace is a concatenation of three vectors: the two items plus an experimental context.

If a paired associate is presented multiple times within the same context, additional features are added to the traces. For the purposes of simplicity, we assume that the model always updates the correct image (the image that was first encoded from the stimulus) if it has already been encoded. This can produce different results from a study-time manipulation in that repetition strengthens contextual information, while increased study time need not (Malmberg & Shiffrin, 2005). An elaborated model of encoding would allow the decision of whether to add a new trace or update an existing trace (and, in that case, which trace to update) to depend upon a comparison of the currently presented stimuli to the long-term store (Shiffrin and Steyvers (1997) and Criss (2006) Criss, Malmberg, & Shiffrin (2010).

During test, a stimulus upon presentation is compared to traces. For each trace-stimulus comparison, a likelihood ratio λ is computed, denoting the likelihood the trace was encoded from the test stimulus versus the likelihood that the trace was encoded from any other stimulus. For some image j being compared to stimulus k, the likelihood ratio λ_{jk} for j being a representation of k is:

$$\lambda_{jk} = \frac{\Pr(\text{Old}|\text{Data})}{\Pr(\text{New}|\text{Data})} = (1-c)^{n_{jq}} \prod_{\forall i} \left(\frac{c + (1-c)g_s(1-g_s)^{i-1}}{g_s(1-g_s)^{i-1}} \right)^{n_{ijm}} \tag{4}$$

where n_{jq} counts the number of encoded features that do not match the stimulus and n_{ijm} counts the number of features with value $v = i$ that do match the stimulus. See Shiffrin and Steyvers (1997) for a derivation of this equation. This occurs both for the stimulus presented and the study-test context in which it was presented, yielding λ_{Ijk} for item information and λ_{Cjk} for context information.

REM can account for effects from multiple test types by utilizing different retrieval strategies to account for task demands. For single item recognition, these likelihood ratios are averaged to give a global familiarity Φ, which is then compared to some criterion ε. If the global familiarity exceeds this criterion, then the model states that the presented stimulus is "old" or has been studied in the cued context. Otherwise it states that the stimulus is unstudied, or "new." The recognition decision can be written as

$$\text{Response}_k = \begin{cases} \text{"old"}, & \Phi_k > \varepsilon \\ \text{"new"}, & \Phi_k \leq \varepsilon \end{cases} \tag{5}$$

REM completes recall tasks by repeatedly sampling for to-be-recovered information and attempting to recover that information (Figure 4). In brief terms, sampling occurs via the Luce choice rule and is a function of the degree of match between the test cue and the traces in memory. Upon sampling of some trace, any sampled item information is compared to a threshold ε. If the item passes this threshold, then the model attempts to output the unsampled item information, with the probability of output equal to the proportion of correctly encoded traces, raised to the power τ. The number of times this sample-recovery loop may attempt this process is limited by K_{max} failures.

For free recall, we assume for purposes of simplicity that context serves as the sole test cue. As such, the probability of selecting an item j from the long term store is a function of the degree of match between that item's concatenated context information and the test context k:

$$\Pr(j|k) = \frac{{\lambda_{Cjk}}^{\gamma}}{\sum_{n=1}^{N} {\lambda_{Cnk}}^{\gamma}} \tag{6}$$

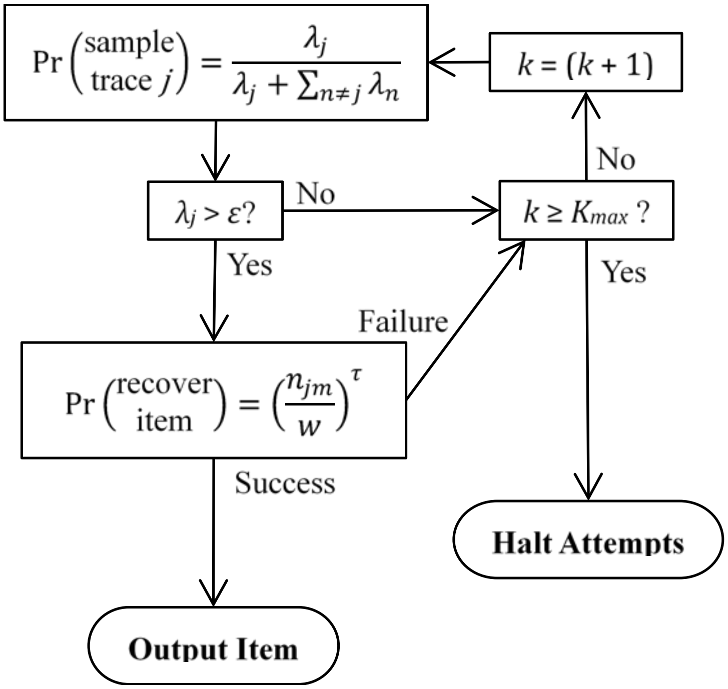


Figure 4: The sample-recovery process of REM. The model samples for a portion of a trace using some test cue, tests the sampled item information against threshold ε, then attempts to output the unsampled item features from that trace. λ represents the comparison of the test cue to the trace, in the case of cued recall $\lambda = \lambda_C{}^{\gamma}\lambda_I$. Other parameters and values: k = current attempt; K_{max} = maximum number of retrieval attempts; n_{jm} = number of features in the to-be-recovered trace that match the lexical-semantic features of the word the features represent; w= the length of the vector that contains the proposed target, $\boldsymbol{n_{jm}/w}$ gives the proportion of features in the proposed target that match the lexical-semantic trace it represents; τ = factor weighting probability of retrieval.

where N gives the number of relevant items in the long-term store. The "squishing" of λ_{Cjk} with parameter γ is consistent with prior free recall instantiations of REM (Malmberg & Shiffrin, 2005) and is used to attenuate the heavily skewed distribution of λ (Shiffirn & Steyvers, 1998, Diller, Nobel, & Shiffrin, 2001, Malmberg & Shiffrin, 2005). Essentially, the sampling of items from memory is demonstrably noisier than predicted by an unweighted Luce choice of the contexts, as such γ accounts for this additional noise. Potential sources of this noise include, but are by no means limited to, item information in the test cue, mismatches between the study and test contexts, generally weaker encoding of context information relative to item information, and/or limited activation of contextual features from the test cue during test, due perhaps to attentional weighting. This question, although worthy of study, is not more deeply considered here.

In the case of multiple unsampled items in a trace, as in free recall testing of studied word pairs, one item is selected at random for a recovery attempt upon sampling of its context. An elaboration of the model would allow the other member of the pair to then be recovered if the first recovery was successful (Lehman & Malmberg, 2013), which would account for a zigzag effect (so termed by Davelaar et al. 2006) in the serial position curve during some free recall tasks. This modification would not qualitatively alter the outcome of the list strength effect in free recall. Furthermore, because in cued recall an item and context are used to sample and therefore only one unsampled item would exist in the sampled trace, this elaboration would have no effect on our instantiation of cued recall.

During the sampling of cued recall, in contrast to Diller, Nobel, and Shiffrin (2001), both item and context information contribute to the sampling probabilities. Specifically, the probability of the selecting item *j* from a set of *N* items, given some test cue *k*, is

$$\Pr(j|k) = \frac{{\lambda_{Cjk}}^{\gamma}\lambda_{Ijk}}{\sum_{n=1}^{N}{\lambda_{Cnk}}^{\gamma}\lambda_{Ink}} \tag{7}$$

where *I* denotes the likelihood comparison for the item information, and *C* denotes that for the concatenated context. Setting $\gamma = 0$ sets the model to an items-only version similar to that of Diller, Nobel, and Shiffrin (2001).

The addition of contextual information is critical to allow the cued recall model to predict a positive list strength effect. This is made clear when the Luce choice rule for cued recall is algebraically extended thusly:

$$\Pr(j|k) = \frac{{\lambda_{Cjk}}^{\gamma}\lambda_{Ijk}}{{\lambda_{Cjk}}^{\gamma}\lambda_{Ijk} + \sum_{n\neq j}{\lambda_{Cnk}}^{\gamma}\lambda_{Ink}} \tag{8}$$

The probability of sampling the correct image is a function of (1) the degree of match for the correct trace and (2) the summed degree of match for all the mismatching traces. The effect of

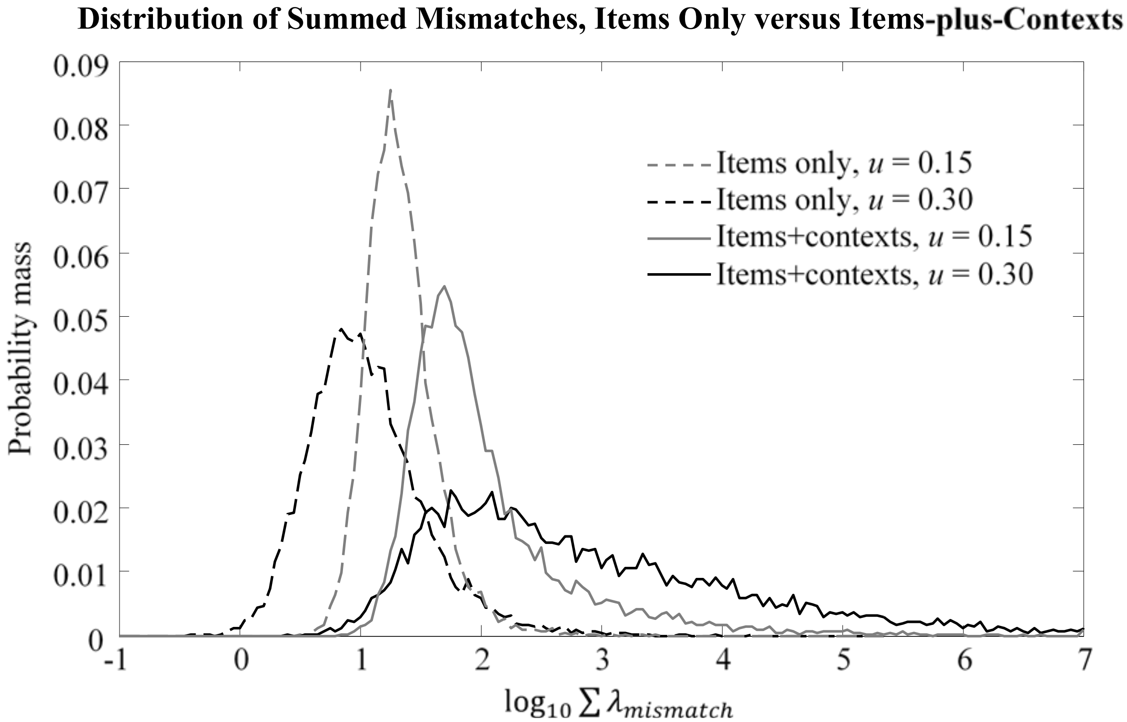


Figure 5: Distribution of mismatching lambdas, computed using just items or with items plus contexts. Parameters: u_{strong} = 0.30, u_{weak} = 0.15, number of mismatching items = 31, γ = 0.2.

strengthening the correct trace impacts only the term ${\lambda_{Cjk}}^{\gamma}\lambda_{Ijk}$. The effect of strengthening other traces impacts only the summand $\sum_{n\neq j}{\lambda_{Cnk}}^{\gamma}\lambda_{Ink}$. Manipulations of memory strength for other items on the study list are felt in the model by altering the distribution of this summand. Consider the items-only case where $\gamma = 0$. By property of differentiation, strengthening other items on the list will tend to shift the distribution of λ_{Ink} and therefore the distribution of $\sum_{n\neq j}\lambda_{Ink}$ towards smaller values, and therefore increase the probability of sampling the correct trace. This leads, then, to a negative list strength effect: strong items on a pure strong list are more likely to be correctly sampled (and therefore recalled) than those a mixed list, and weak items on a pure weak list are less likely to be correctly sampled (and therefore recalled) than those on a mixed list. Weighting item information, as per Diller, Nobel, and Shiffrin (2001), would reduce the size of this effect, but cannot reverse the pattern.

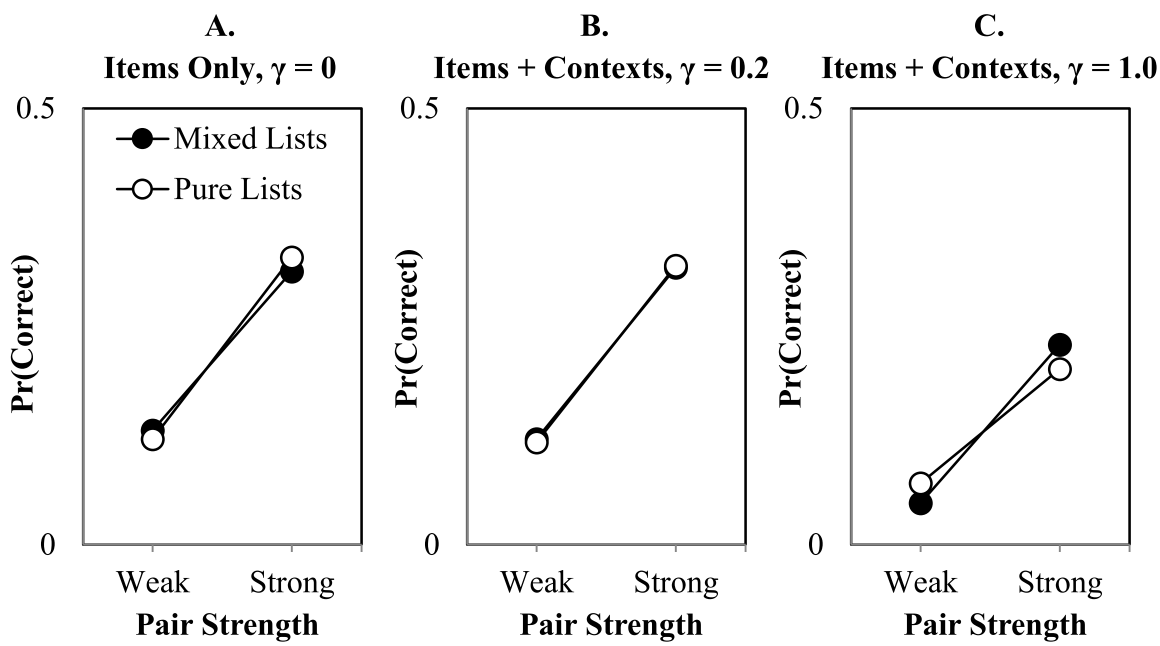


Figure 6: Cued recall predictions for the (A) items only model, $\gamma = 0$, (B) the items-plus-contexts model with context weighted to $\gamma = 0.2$, and (C) the items-plus-contexts model with context weighted to $\gamma = 1.0$. Parameters $u_{strong} = 0.30$ and $u_{weak} = 0.15$ approximate levels of performance for item recognition on Experiment 5. Parameters $\tau = 0.5$, $\varepsilon = 1.0$ are standard (Diller, Nobel, & Shiffrin, 2001; Malmberg & Shiffrin, 2005). $K_{max} = 1$ approximates overall levels of cued recall performance on Experiment 5 and does not alter the direction of the list strength effect.

Context, on the other hand, can reverse the effect from negative to positive (Figure 5). Because context information is largely the same within a study-test block (in other words, all study contexts are similar to each other), this happens in the same way that similarity between items can reverse the effect of differentiation (Criss, 2006). All items encoded within a study-test block have matching contexts, thus strongly encoded mismatching items also have strongly encoded matching contexts. The increase in familiarity from more strongly encoded contexts outweighs the decrease in familiarity for more strongly encoded items, such that strongly encoded mismatches become more familiar than weakly encoded ones. This reverses the direction of the list strength effect. Figure 6 is demonstrative of the impact of context on cued

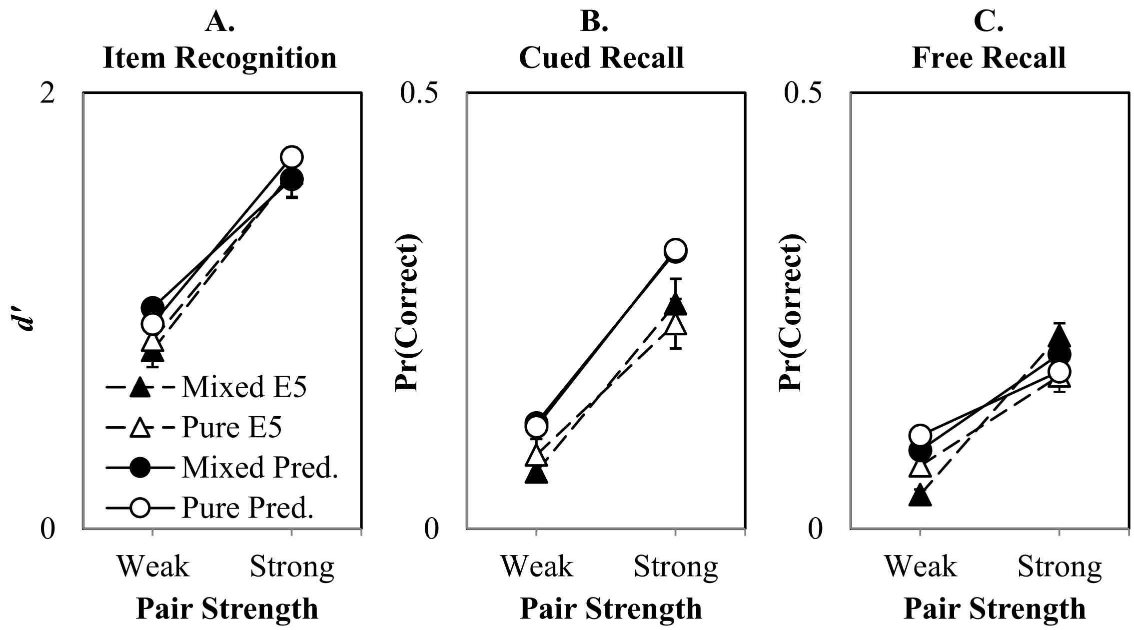


Figure 7: Data (triangles and dotted lines) from and model predictions (circles and solid lines) to Experiment 5 for (A) single item recognition, (B) cued recall (C) free recall. Parameters: u_{strong} = 0.30 and u_{weak} = 0.15 are set to the levels of performance for item recognition. Paremeters γ = 0.2, τ = 0.5, ε = 1.0 are standard (Diller, Nobel, & Shiffrin, 2001; Malmberg & Shiffrin, 2005) and fixed a priori for demonstrative purposes. The only free parameters are K_{max} = 1 for cued recall and K_{max} = 8 for free recall. K_{max} does not alter the direction of the list strength effect.

recall in this regard: increasing the contribution of context to sampling in cued recall increases the list strength effect.

Figure 7 provides qualitative predictions for the simple and constrained implementations of free recall, cued recall, and single item recognition described above. These constrained parameters also provide decent quantitative fits to the actual data from Experiment 5. After adjusting *u* to approximate item recognition performance on Experiment 5, the only "free" parameters were K_{max} for cued and free recall. The simple inclusion of context information, weighted to the same degree as in free recall, accounts for list strength patterns observed in the data: the list strength effect in cued recall falls about halfway between single item recognition and free recall.

As such, cued recall may be thought of as a middle ground between recognition and recall tasks. Its position as a middle ground is important: it means that cued recall can serve as a point of contact between effects and theories more closely associated with recall, and those more closely associated with recognition. One limitation to the separation of memory models into free recall and recognition is that findings from one sub-discipline can have limited impact on the theories in the other. However, both item and context information are critical in cued recall in the REM model.

Other Models of Episodic Memory

TODAM (Murdock, 1982) is an interesting comparison to REM in this instance for two reasons. First, the pattern of the list strength effect in TODAM is driven by memory for items prior to the studied list and interference from recalled items in unstructured retrieval tasks (Murdock & Kahana, 1993), rather than utilization of context. Second, it utilizes associative information in a unique fashion that is not fully accounted for in the REM model, yet might explain some observed dissociations between item recognition and cued recall tasks (e.g. Aue, Criss, & Fischetti, 2012) and item and associative recognition tasks (e.g. Hockley & Cristi, 1996, Criss, 2004). The focus of this review of TODAM is the first point. The second, and beyond that whether and to what degree explicit associative features are necessary within the REM model, is a topic of future study.

TODAM is a single-store model where information is added to and slowly decays from memory. In TODAM2 (Murdock, 1993, Murdock, 1997), stimuli are represented by vectors of normally-distributed feature values, and the association between two stimuli is represented by their sum[5]. Each item is additionally flanked by information pertaining to context, which allows

[5] In prior versions of TODAM and (Murdock, 1982), the association between two items was represented a convolution of their vectors.

for a more rapid forgetting of item information than associative information (Murdock, 1997). Items and associations are encoded by auto-convolving the vectors and adding them directly to memory. This encoding is probabilistic: a limited set of features are activated before autoconvolution. Additionally, the model subscribes to a limited attention hypothesis, wherein the contributions of individual items and their association are weighted. Because item information is present in the association, attending to associative information does not seriously harm item recognition performance, but failing to attend to associative information will harm associative recognition performance. This is the half-seesaw effect observed by Hockley & Cristi (1996). Furthermore, upon each encoding event, memory decays along with the information contained within. Retrieval tasks begin by computing the cross-correlation between the test cue and memory. The resulting output is an approximation of the item studied alongside the test cue, or if unassociated an approximation of the test cue itself.

How this approximation is used during retrieval is task-dependent. In recognition, the dot product of the output and the test cue is compared to some criterion, with values above the criterion being called "old" and sub-criterion values being called "new." Recall can be modeled by feeding the output into a Brain-State-in-a-Box deblurring algorithm as per Lewandowsky (1994). The deblurred image is then compared to the possible output options via dot-product. The best match is outputted if that dot-product is greater than some output criterion. In either task, the output is fed back into the model which can interfere with future retrievals (Lewandowsky & Murdock, 1989).

TODAM is able to predict a null list strength effect in recognition and cued recall and a positive list strength effect in free recall, then, because the memory system contains information leading into a study phase and continues to learn during test. The information that was already within memory before the study phase buffers the memory system against the additional variance incurred by strengthening items. This prevents the positive list strength effect that might otherwise occur if no prior memories existed (Murdock & Kahana, 1993). In free recall, a positive list strength effect is driven by re-encoding of recalled items and the order in which these items tend to be retrieved, namely that strong items tend to be retrieved and be re-encoded first. On a mixed list, this means that strong items are more likely to be retrieved than weak items in free recall simply because they are less likely to have been interfered with by prior recalls. In cued recall and recognition, because the order of testing is fixed, re-encoding the retrieved items cannot drive a positive list strength effect (unless of course the experimenter tests strong items before weak items), therefore a null list strength effect occurs.

The experimental results of this paper are something of a mixed bag for TODAM when it comes to its account of list strength. On the one hand, that our estimate the list strength effect in cued recall is smaller than previously thought (modal DoD from the posterior predictive distribution of the cross-experiment analysis = 0.015 (Figure 3), DoD from summary means of RCS(1990)e6 = 0.058) is a good thing for the model: TODAM predicts a null list strength effect for cued recall tasks, and the effect now appears be closer to null than before. However, TODAM also predicts no clear dissociation between item recognition and cued recall when it comes to the list strength effect. This is in conflict of our finding of a small but positive effect for cued recall that contrasts with the null-to-negative effect in item recognition. An accounting of this contrast in TODAM will likely require an adjustment to the item recognition process such that strengthening memory removes noise from the decision.

Summary

Across five experiments, a very small positive list strength effect was found, in contrast to the canonical null-to-negative list strength effect in recognition and large, positive list strength effect in free recall. This finding can be accounted for in the REM model if context information is used alongside items as part of the test cue during the memory search. This establishes cued recall as a middle ground between free recall, where the effects of context dominate, and recognition, where the effects of items dominate, because both of these components are critical in accounting for the cued recall data presented here. If, as memory researchers, we endeavor to understand the basic processes of memory, this means developing models that encompass a wide variety of retrieval tasks. Cued recall may, in this regard, be an important link in that endeavor.

References

Balakrishnan, J.D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22*(3), 615-633. http://dx.doi.org/10.1037/0096-1523.22.3.615

Clark, S.E., & Gronlund, S.D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*(1), 37-60.

Criss, A.H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*, 461-478.

Criss, A.H., & Howard, M.W. (2015) Models of episodic memory. In Busmeyer, J.R., Wang, Z., Townsend, J.T., & Eidels, A. (Eds.). *The Oxford handbook of computational and mathematical psychology* (pp. 165-183). New York, NYOxford University Press.

Criss, A.H., & McClelland, J.L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55*, 447-460.

Criss, A.H. & Shiffrin, R.M. (2004a). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review, 111*(3), 800-807.

Criss, A.H., & Shiffrin, R.M. (2004b). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition 32*(8), 1284-1297.

Criss, A.H. & Shiffrin, R.M. (2005). List discrimination and representation in associative recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*(6), 1199-1212.

Davelaar, E.J., Haarmann, H.J., Goshen-Gottstein, Y., & Usher, M. (2006). Semantic similarity dissociates short- from long-term recency effects: Testing a neurocomputational model of list memory. *Memory & Cognition, 34*(2), 323-334.

Dennis, S., & Humphreys, M.S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452-478.

Hautus, M.J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods, Instruments, & Computers, 27*(1), 46-51.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 302-313. http://dx.doi.org/10.1037/0278-7393.21.2.302

Hockley, W.E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. Memory & Cognition, 24(2), 202-216.

Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269-299.

Humphreys, M.S., Pike, R., Bain, J.D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33*, 36-67.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Clarendon Press, 3ed.

Kahana, M.J., Rizzuto, D.S., & Schneider, A.R. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 933-953.

Lehman, M. & Malmberg, K.J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review 120*(1), 155-89. http://dx.doi.org/10.1037/a0030851

Lewandowsky, S. (1994). Memory for serial order revisited. *Psychological Review, 101*(3), 539-543. http://dx.doi.org/10.1037/0033-295X.101.3.539.

Lewandowsky, S., & Murdock, B.B. (1989). Memory for serial order. Psychological Review, 96(1), 25-57.

Malmberg, K.J., & Shiffrin, R.M. (2005). The “one-shot” hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 322-336.

Mensink, G-J. M., Raaijmakers, J.G. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology, 33*(2), 172-186.

McClelland, J.L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective likelihood approach to the effects of experience in recognition memory. *Psychological Review 105*(4), 724-760.

Murdock, B.B., (1982). A Theory for the Storage and Retrieval of Item and Context Information. *Psychological Review 89*(6), 609-626.

Murdock, B.B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*(2), 183-203.

Murdock, B.B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104*(4), 839-862.

Murdock, B.B., & Kahana, M.J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 689-697.

Murnane, K., & Shiffrin, R.M. (1991). Word repetitions in sentence recognition. *Memory & Cognition, 19*(2), 119-130.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134.

Ratcliff, R., Clark, S.E., & Shiffrin, R.M. (1990). List strength effect I: Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163-178.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*(4), 763-785.

Rose, R.J., & Sutton, L.T. (1996). Encoding conditions and the list strength effect. *Canadian Journal of Experimental Psychology, 50*(3), 261-269.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review 16*(2), 225-237.

Sahakyan, L., Abushanab, B., Smith, J.R., & Gray, K.J. (2014). Individual differences in contextual storage: Evidence from the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*(3), 873-881.

Shiffrin, R.M., Ratcliff, R.C., & Clark, S.E. (1990). List strength effect II: Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179-195.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers, 31*(1), 137-149.

Stretch, V., & Wixted, J.T. (1998). Decision rules for recognition memory confidence judgements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1397-1410. http://dx.doi.org/10.1037/0278-7393.24.6.1397

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive Psychology, 5*, 275-287.

Wetzels, R., Matzke, D., Lee, M.D., Rouder, J.N., et al. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science 6*(3), 291-298.

Curriculum Vitae

Education

Graduate Student Cognition, Brain, and Behavior Program, Syracuse University

Advisor: Amy H. Criss, Ph.D.

B.S. *Psychology*, College of Charleston, 2011

B.S. *Biology*, College of Charleston, 2011

Publications and Presentations

Wilson, J.H., & Criss, A.H. (May 2015). Release from output interference in cued recall: Disassociations between cues and targets. Poster presented at the 2015 Context and Episodic Memory Symposium, Philadelphia, PA.

Wilson, J.H, & Criss, A.H. (July 2015). Release from output interference in cued recall: Cue-target dissociations. Poster accepted to the 48th Meeting of the Society of Mathematical Psychology, Newport Beach, CA.

Wilson, J.H, & Criss, A.H. (July 2014). The Retrieving Effectively from Memory Model and the list strength effect in cued recall. Poster presented at the 2014 Summer School for the Computational Modeling of Cognition, Laufen, Bavaria, Germany.

Wilson, J.H., & Criss, A.H. (July 2014). The list strength effect in cued recall. Poster presented at the 47th Annual Meeting of the Society for Mathematical Psychology, Québec City, Canada.

Wilson, J.H., Aue, W.R., & Criss, A.H. (November 2013). The effects of cue and target strength in cued recall. Poster presented at the Annual Meeting of the Psychonomic Society, Toronto, Canada.

Wilson, J.H., & Oprisan, S.A. (2012). Resistance between channels may lead to increased action potential efficiency. *BMC Neuroscience, 13*(Suppl 1), P77.

Wilson J.H., & Oprisan, S.A. (October 2011). Effects of channel clustering on action potential threshold and propogation efficiency in a computational model. Poster presented at Neuropalooa – MUSC and CofC Neuroscience, Charleston, SC.

Wilson J.H., & Oprisan S.A. (July 2011). Non-equilibrium diffusive fluxes of ions and their impact on the refractory period of tightly packed neurons. Poster accepted for presentation at the Twentieth Annual Computational Neuroscience Meeting: CNS*2011, Stockholm. Published in: *BMC Neuroscience 2011*, 12(Suppl 1):P243.

Wilson J.H., & Oprisan S.A. (April 2011) Analysis of potassium-based conductance models. Poster presented at 23rd Annual Research Poster Session, School of Sciences and Mathematics, College of Charleston, Charleston, SC.

Wilson, J.H., & Oprisan, S.A. (April 2011). Computational investigation of the influence of extracellular ionic fluctuation on neural activity. Poster presented at the South Carolina Academy of Sciences 2011 Conference, SC State University, Orangeburg, SC.

Wilson J.H., & Oprisan S.A. (March 2011) Comparison of ion-based periodic spiking models using phase response curves. Poster presented at the Symposium for Young Neuroscientists and Professors of the South East (SYNAPSE), Wake Forest University, Winston-Salem, NC.

Research Experience

Graduate Student, Memory Modeling Lab

Fall 2012 – Present

PI: Amy H. Criss, Ph.D., Syracuse University

Research Assistant, “Alcohol Disruptions to Long-Term Memory Consolidation”

Fall 2011 – Spring 2012

PI: Judson Chandler, Ph.D., Medical University of South Carolina

Post-Baccalaureate Researcher, Computational Neuroscience Laboratory

Summer 2011 – Spring 2012

PI: Sorinel A. Oprisan, Ph.D. College of Charleston

Research Assistant, PHAGE

Spring 2011

PI: Anastasia Zimmerman, Ph.D., College of Charleston

Research Assistant, “Project Boost”

Spring 2010

PI: Cynthia May, Ph.D., College of Charleston

Research Assistant, “Priming for Impulsivity”

Fall 2009

PIs: Chad Galuska, Ph.D., Anthony Bishara, Ph.D., Lisa Thomson-Ross, Ph.D., College of Charleston

Awards and Honors

Department of Psychology Travel Awards	Fall 2013
	Spring 2015
Graduate Student Organization Travel Awards	Fall 2014
HHMI Undergraduate Science Education Grant	Summer 2011

Awarded for students “to engage in cutting-edge research at the College of Charleston.”

Teaching Experience

Teaching Assistant:

Introduction to Psychology

Biophysical Modeling of Excitable Cells