Journal of Memory and Language 95 (2017) 78–88

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The list strength effect in cued recall

Jack H. Wilson *, Amy H. Criss

Department of Psychology, Syracuse University, United States

ARTICLE INFO

Article history:
Received 29 February 2016
revision received 26 August 2016

Keywords:
List strength effect
Cued recall
Episodic memory
Models of memory

ABSTRACT

Episodic memory is the process by which information about experienced events is encoded and retrieved. Successful retrieval of episodic memories is dependent on the way in which memory is tested and as a result many effects and theories of episodic memory are task dependent. One such finding is the list strength effect. In free recall, a positive list strength effect is observed; memory for a given item is harmed by the presence of other strongly encoded items and helped by the presence of other weakly encoded items. In recognition, a null list strength is observed; memory for a given item is unaffected by the strength of other items. Such differential empirical findings are crucial to understanding memory, but it is undesirable to have multiple task-specific theories rather than a unified theory of memory. Here we use cued recall, a task that shares properties of both free recall and recognition, to move toward that goal. In a series of 5 experiments, we observed a null list strength effect in cued recall. We suggest that a successful theory would entail the use of both item and context information during retrieval, consistent with the approach of the Search of Associative Memory model.

Introduction

Episodic memory is the process or processes by which experienced events are encoded into and retrieved from memory. Memory is often studied using one of two broad classes of memory tests: recognition, identifying whether or not an item was experienced in some specific context, or recall, the generation of items experienced within some context. These tasks often result in different patterns of empirical data and therefore are often explained through different proposed mechanisms (Criss & Howard, 2015). This is reflected by the models used, some dedicated to recognition tasks (e.g. Bind-Cue-Decide Model of Episodic Memory, Dennis & Humphreys, 2001), some to recall tasks (e.g. Temporal Context Model, Howard & Kahana, 2002).

The global matching models 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 (see Humphreys, Pike, Bain, & Tehan, 1989) These models, which include the Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1981) model and the Theory of Distributed Associative Memory model (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. In general these models predicted that items other than the test item affect memory for the test item. One specific prediction was that as the strength of other study-list items increased, accuracy decreased. This prediction was tested in the list strength paradigm.

In the list strength paradigm, participants study and are tested on a "pure strong" list where their memory for all the items is strengthened, a "mixed" list where their memory for only half of the items is strengthened, and a "pure weak" list where memory was not strengthened. Memory is tested after the presentation of each list. Memory is typically strengthened by repetition, study time, or encoding task. In contrast, the list strength effect measures the difference in performance for items of a given encoding strength (strong or weak) as a function of the strength of the other items on the list (pure list or mixed list). 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 – the prediction of global matching models. 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. Note that the general question of whether the strength of encoded items affects memory for a given item can be measured in many different paradigms, including the

* Corresponding author at: Department of Psychology, 430 Huntington Hall, Syracuse University, Syracuse, NY 13244, United States.
E-mail addresses: jhwilson@syr.edu (J.H. Wilson), amy.criss@gmail.com (A.H. Criss).

http://dx.doi.org/10.1016/j.jml.2017.01.006
0749-596X/

retrieval induced forgetting paradigm (Anderson, Bjork, & Bjork, 1994; Raaijmakers & Jakab, 2012; Verde, 2009, 2013). Here we focus on the simplest paradigm – the list strength paradigm just described. In the general discussion, we return to the broader question of the role of encoding strength on episodic memory.

Critically, Ratcliff, Clark, and Shiffrin (1990) observed that the effect of list strength in a list strength paradigm, when memory is strengthened via spaced repetitions, depends on how memory is tested. In free recall tasks, they observed a positive list strength effect (Malmberg & Shiffrin, 2005; Tulving & Hastie, 1972). 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. Simultaneously accounting for the positive list strength effect in recall and null effect in recognition and also list length effects proved difficult for the global matching models (see Shiffrin, Ratcliff, & Clark, 1990; Shiffrin, Ratcliff, Murnane, & Nobel, 1993; cf. Murdock & Kahana, 1993a, 1993b). This effect, along with the observation of a word frequency mirror effect and other empirical findings, led to new classes of models. While these "second generation" models have been incredibly successful, it is not optimal to have task-dependent explanations of memory absent a unified theory (see Criss & Howard, 2015).

Cued recall may be one means by which to bridge these lines of research (e.g. Criss, Aue, & Smith, 2011). Cued recall is a hybrid of recall and recognition tasks. Like recognition, a cue is provided during cued recall. Like free recall, a target item must be generated from memory and outputted. Given the list strength effect's history as a critical component in evaluating multi-task models of memory, it is important to have a robust estimate of the size and direction of the list strength effect in cued recall.

With that in mind, the goal of this paper is to conduct a careful evaluation of the list strength paradigm in cued recall. A number of studies have replicated the size and direction of the list strength effect in single item recognition (e.g. Murnane & Shiffrin, 1991a, 1991b; Ratcliff, McKoon, & Tindall, 1994) and free recall (e.g. Malmberg & Shiffrin, 2005; Ratcliff et al., 1990; Rose & Sutton, 1996; Sahakyan, Abushanab, Smith, & Gray, 2014). However, there are only two reports of the list strength paradigm in cued recall, both tangential to the primary aims of those papers. The first is the original report of a single experiment in Ratcliff et al. (1990). Kahana, Rizzuto, and Schneider (2005) also tested memory for mixed and pure lists with cued recall, in order to draw correlations between recognition and recall. However, the measurement of list strength was secondary to their goals and they thus allowed weak words in mixed lists to have a greater study-test lag than those in pure lists. In the following set of experiments, we evaluate the list strength effect paradigm in cued recall.

Overview of experimental design

We present five experiments that differ in methodological details but all implement the list strength effect paradigm. Each experiment progressively approximates the precise procedures used in Experiment 6 of Ratcliff et al. (1990; hereafter referred to as RCS(1990)e6). Table 1 outlines the differences between methods of each of the five experiments and RCS(1990)e6. 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. In Experiments 4 and 5, participants were additionally given free recall and single item recognition tests for some blocks, with memory task post-cued. Any participant who completed the experiment completed all study-test blocks, as such all comparisons are fully within-subjects. The order of study-test blocks was fully randomized. In all experiments, each study-test block used a unique and randomized set of words, such that no words repeated across block. Details of the timing and nature of repetition and mixing vary for each experiment and will be described for each individual experiment.

On cued recall trials, participants were instructed to type out the word they had studied alongside the test cue and were given the option to type out the phrase "idk" short for "I don't know" if they could not recall the target. Responses were scored as a correct response (matching the target word, with errors in spelling, tense, and pluralization allowed),[1] an intrusion (a response that is incorrect), or a response failure if they responded with an "idk" or left the question blank.

On free recall tests, participants were instructed to type out as many words as they could recall. The number of correct responses was the number of unique correct outputs. Intrusions in free recall tasks were computed by counting the number of outputted items that were not on the studied list. In free recall, the recall of strong before weak words may contribute to the appearance of a positive list strength effect (e.g., Wixted, Ghadisha, & Vera, 1997). Whether this is a critical driver of the effect in free recall is an empirical question that is not considered in these experiments. Critically, this is not a concern in cued recall where the order of output is controlled by the experimenter.

In tests of single item recognition, a yes-no decision was cued for a series of targets and foils, randomly intermixed, and performance was measured primarily by d' with the loglinear correction (Hautus, 1995; see also Stanislaw & Todorov, 1999; Schooler & Shiffrin, 2005 for discussions of the correction):

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

where H and FA are the number of hits and false alarms, respectively, in a condition, H_{max} and FA_{max} are the maximum possible hits or false alarms, respectively. We report d' for comparison to Ratcliff et al. (1990). The criterion and d' are not independent when the variance of the target and foil distributions is unequal (as is often reported to be the case in memory, e.g., Egan, 1958; Ratcliff, Sheu, & Gronlund, 1992). Therefore a difference in d' cannot be interpreted when it is accompanied by a change in the criterion (e.g., Grider & Malmberg, 2008; Lockhart & Murdock, 1970). Some researchers believe that the criterion changes with list strength (e.g., Hirshman, 1995; Starns, Ratcliff, & White, 2012) despite several papers implicating changes in the distribution of memory strength for foil items rather than a criterion shift (Criss, 2009, 2010; Criss & Koop, 2015; Criss, Wheeler, & McClelland, 2013). To check the robustness of our d' analysis, we report a nonparametric measure of discriminability, A' (Grier, 1971; Pollack & Norman, 1964). Again, this is not critical to the thesis of this paper because we are primarily interested in cued recall not recognition.

Analysis plan

Repeated measures ANOVAs were used to assess differences between conditions. Bayes factors were then used to quantify the evidence for or against the null hypothesis of no-interaction (using default priors in JASP, JASP Team, 2016). This is done by comparing the evidence for a null-hypothesis model that includes just the main effects against an alternative-hypothesis model that also includes the list strength interaction. The Bayes factor BF_{01} may

[1] Using a strict scoring scheme—only words that were written precisely as presented at study were marked as correct—leads to the same qualitative patterns of data and conclusions as those presented in this paper.

Table 1
Manipulations in the present experiments and Ratcliff et al. (1990) Expt. 6.

Manipulations	Experiment					
	1	2	3	4	5	RCS(1990)e6
Study list						
Pure strong	x	x	x	x	x	x
Pure weak	x	x	x	x	x	x
Mixed: weak first			x		x	x
Mixed: weak last			x		x	x
Mixed: weak shuffled			x		x	x
Mixed: lag-controlled	x	x		x		
Pure strong & short					x	x
Pure weak & long					x	x
Training (pure weak)					x	x
Test of memory						
CR	x	x	x	x	x	x
FR				x	x	x
SIR				x	x	x
Strong pair presentations	2x	4x	4x	4x	4x	4x
List length	24	16	16	16	16, 10 (short), 40 (long)	16, 10 (short), 40 (long)
Test trials	24	16	16	16	16, 10 (short), 20 (long)	16, 10 (short), 20 (long)
Study time per presentation (s)	3	1.25	1.25	1.25	1.25	1.25
Rating task during study	Yes	Yes	No	No	No	No
Word frequency	Mixed	Mixed	HF	HF	HF	HF
Number of sessions	1	1	1	1	2	3

be interpreted in the context of these five experiments as the ratio of evidence in favor of a "null" model with only main effects of list and/or of strength versus that in favor of a model that also includes the list strength interaction. For example, a $BF_{01} = 1$ indicates equivalent evidence for the hypotheses under consideration and a Bayes factor $BF_{01} = 10$ indicates that the data are 10 times more likely to be the outcome of a model without an interaction than one with the interaction included. The farther from 1 in either direction, the more evidence favoring the respective hypothesis. Bayes factors incorporate the concept of power in the sense that a Bayes factor reflects the relative evidence and larger Bayes factors imply greater power. We scheduled sessions to meet a sample size of approximately 40 which is consistent with many studies and publications from our lab with the exception of Experiment 5 where we tried to approximate the sample size (N = 84) reported in RCS(1990)e6.

Experiment 1

Methods

Participants

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

Stimuli

The word pool consisted of 1643 words of letter length 4–8 (KF, Kucera & Francis, 1967: range = 1–500, mean = 44.8; logHAL, Lund & Burgess, 1996: range = 2.89–13.7, mean = 8.69).

Procedure

In the pure weak block, the study phase consisted of 24 word pairs presented on the screen for 3 s each with a 0.1 s blank screen between each presentation. Participants were instructed to generate a scene involving the words. Immediately following presentation of each pair, participants were cued to rate the difficulty of completing this scene generation task on a scale of 1–9 (1 = "very easy" and 9 = "very hard"). 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 the mean lag between the test and final study presentation of a word pair was controlled by presenting all strong pairs once followed by randomly intermixing the weak and strong pairs. After each study list, participants engaged in a 60 s distractor math task followed by a cued recall test where every pair was tested by randomly selecting one item from the pair to serve as the cue and the other as the to-be-retrieved-target. Between study-test blocks, participants were permitted to take a break.

Results

Means and standard errors are provided in Fig. 1 for correct responses and in Table 2 for all three response measures. Data were analyzed in a 2 (pure list vs mixed list) $\times$ 2 (strong pair vs weak pair) repeated measures ANOVA. Critically, there is evidence of a null list strength effect. That is, no list type by item strength (henceforth shortened to list strength) interaction was observed for correct responses, $F(1, 39) < 1$, $BF_{01} = 2.9$. A list strength interaction was present for both intrusions, $F(1, 39) = 4.51$, $p = 0.040$, $BF_{01} = 0.81$, and response failures, $F(1, 39) = 5.89$, $p = 0.020$, $BF_{01} = 0.52$. Strong pairs were more likely to be correctly recalled than weak pairs, $F(1, 39) = 74.8$, $p < 0.001$, $BF_{01} = 7.7 * 10^{-9}$ and less likely to be associated with an intrusion, $F(1, 39) = 4.48$, $p = 0.041$, $BF_{01} = 0.49$, or a response failure, $F(1, 39) = 43.7$, $p < 0.001$, $BF_{01} = 1.1 * 10^{-5}$. No main effect of list type (mixed vs pure) was observed for correct responses $F(1, 39) < 1$, $BF_{01} = 6.0$, intrusions, $F(1, 39) = 2.50$, $p = 0.124$, $BF_{01} = 2.4$, or response failures $F(1, 39) < 1$, $BF_{01} = 4.3$.

Discussion

We found a null list strength effect for cued recall, which does not match the qualitative outcome of RCS(1990)e6 who found a small positive list strength effect. As noted in Table 1, there are a number of methodological differences between this experiment and the original study. Specifically, there are differences in the way the lists were mixed, the number of repetitions, the strength of the items, the list length, and encoding strategy that each may have resulted in the null rather than positive list strength effect observed here. These are addressed in the experiments that follow.

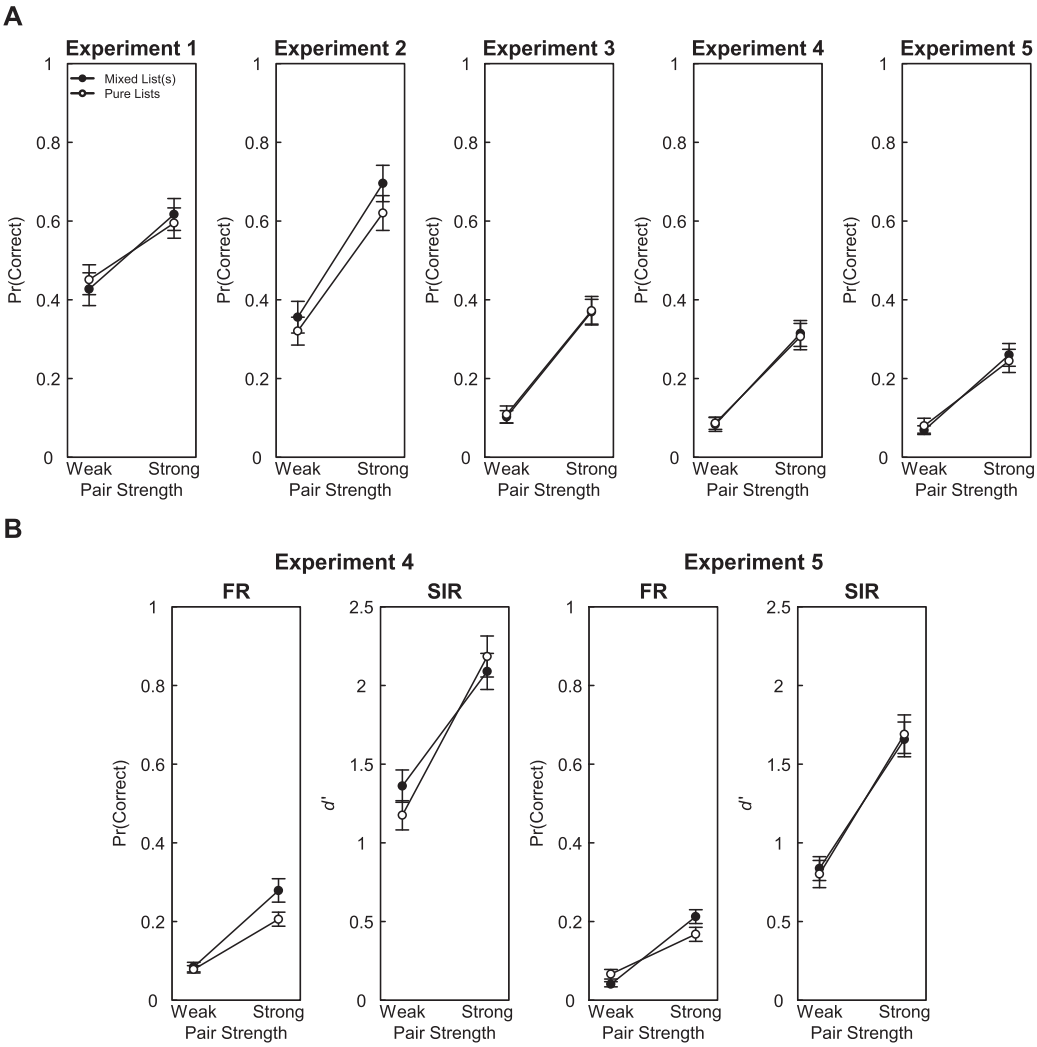

Fig. 1. List strength effect means ± 1 standard error of the mean for (A) cued recall in all five experiments and (B) single item recognition (SIR) and free recall (FR) in Experiments 4 and 5.

Table 2
Means and standard errors of the mean (in parentheses) for cued recall performance by Experiment and condition. For Experiments 3 and 5, this table reports mixed list performance as an average of three forms of mixed list (see Table 3).

	Mixed		Pure	
	Weak	Strong	Weak	Strong
Experiment 1				
Corrects	0.427 (0.042)	0.617 (0.041)	0.451 (0.038)	0.595 (0.039)
Intrusions	0.127 (0.027)	0.121 (0.025)	0.177 (0.031)	0.114 (0.022)
Response failures	0.446 (0.043)	0.263 (0.034)	0.372 (0.031)	0.292 (0.033)
Experiment 2				
Corrects	0.356 (0.040)	0.696 (0.046)	0.321 (0.035)	0.620 (0.044)
Intrusions	0.196 (0.039)	0.154 (0.035)	0.196 (0.032)	0.159 (0.031)
Response failures	0.449 (0.042)	0.151 (0.033)	0.484 (0.037)	0.221 (0.028)
Experiment 3				
Corrects	0.103 (0.016)	0.370 (0.032)	0.109 (0.021)	0.372 (0.036)
Intrusions	0.223 (0.037)	0.227 (0.033)	0.249 (0.038)	0.226 (0.035)
Response failures	0.675 (0.038)	0.403 (0.033)	0.642 (0.039)	0.402 (0.036)
Experiment 4				
Corrects	0.083 (0.017)	0.315 (0.033)	0.087 (0.016)	0.307 (0.033)
Intrusions	0.211 (0.032)	0.257 (0.034)	0.171 (0.029)	0.240 (0.032)
Response failures	0.706 (0.035)	0.428 (0.034)	0.742 (0.031)	0.454 (0.035)
Experiment 5				
Corrects	0.069 (0.011)	0.260 (0.029)	0.080 (0.019)	0.245 (0.029)
Intrusions	0.161 (0.027)	0.207 (0.029)	0.205 (0.035)	0.182 (0.033)
Response failures	0.770 (0.028)	0.532 (0.035)	0.714 (0.034)	0.573 (0.039)

Experiment 2: repetition, study time, and list length

One possible explanation for the null list strength effect observed in Experiment 1 is that the difference between strong and weak pairs in Experiment 1 was not sufficiently large to elicit an effect. Here we replicate Experiment 1 with the exception of list length, study time per trial, and number of repetitions.

Methods

Participants

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

Stimuli

The word pool was the same as Experiment 1.

Procedure

The following modifications were made from Experiment 1: Study time in all lists was reduced from 3 s to 1.25 s (the 0.1 s interstimulus blank screen was preserved), and the word pair was removed from the participants' monitor during the rating task to reduce the potential for residual study while rating. Strong pairs were presented four times rather than twice. The study-test lag control was maintained by presenting the set of strong words three times, in a random order each time, followed by a fourth presentation intermixed with the weak items. Finally, the list length was reduced from 24 unique pairs to 16 unique pairs. All other methodological details were the same as Experiment 1.

Results & discussion

The statistical analyses used in Experiment 1 were used again in Experiment 2 and the same critical set of effects for correct responses (Fig. 1, Table 2) was observed, suggesting that the absence of the effect in Experiment 1 cannot be attributed to insufficient repetitions or differences in list length or presentation time. No list strength interactions were present, $F(1, 38) < 1$ for all three response measures. The data provide evidence in favor of a null list strength effect (correct responses: $BF_{01} = 3.5$), evidence for no effect of list strength on intrusions, $BF_{01} = 4.4$, and evidence for a null effect of list strength on response failures: $BF_{01} = 3.5$. Strong word pairs were associated with more correct responses, $F(1, 38) = 124$, $p < 0.001$, $BF_{01} = 2.4 * 10^{-18}$, fewer intrusions, $F(1, 38) = 4.32$, $p = 0.044$, $BF_{01} = 0.71$, and fewer response failures $F(1, 38) = 114$, $p < 0.001$, $BF_{01} = 3.9 * 10^{-16}$. No main effects of list type were observed (corrects: $F(1, 38) = 3.24$, $p = 0.080$, $BF_{01} = 2.5$; intrusions: $F(1, 38) < 1$, $BF_{01} = 5.7$; response failures: $F(1, 38) = 2.65$, $p = 0.112$, $BF_{01} = 2.3$).

Experiment 3: mixing, word frequency, and encoding task

In this experiment, we removed three more methodological differences between our procedure and RCS(1990)e6, none of which have any theoretical bearing on the list strength effect, but nonetheless differed: normative frequency of the word pool, rating task during encoding, and the type of mixed list. Multiple types of mixed lists were used in RCS(1990)e6, all of them different from the mixing procedure we used in Experiments 1 and 2. The three mixing conditions in RCS(1990)e6 included placing all the weak pairs at the beginning of the list, all the weak pairs at the end of the list, and randomly mixing in the weak pairs throughout the 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. In addition to the changes made in Experiment 2 (list length, study time, and number of repetitions), we matched the word frequency and procedure for building mixed lists and eliminated the rating task in the following experiment.

Methods

Participants

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

Stimuli

The word pool consisted of 800 high-frequency words (KF frequency $\geqslant 50$, logHAL $\geqslant 9$, 4–11 letters long).

Procedure

Within-subjects, participants now ran through five study-test blocks including one strong, one weak, and three mixed lists. Rather than provide 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 that the previous blocks were no longer relevant.[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 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 & discussion

In a 3 (mixed list types) $\times$ 2 (weak vs strong) repeated measures ANOVA, no main effect of list (corrects: $F(2, 92) = 1.56$, $p = 0.216$; intrusions, $F(2, 92) < 1$; response failures: $F(2, 92) = 2.72$, $p = 0.07$) or list strength interaction ($F(2, 92) < 1$ for all three measures) was observed therefore the three mixed lists were averaged (Fig. 1, Table 2). Again, no list strength interaction was observed, $F(1, 46) < 1$, for all three response measures. There is evidence in favor of a null list strength effect for correct responses, $BF_{01} = 5.0$, evidence for the null for intrusions, $BF_{01} = 3.9$, and evidence for the null for response failures: $BF_{01} = 3.7$. Strong items were more likely to elicit correct responses, $F(1, 46) = 126$, $p < 0.001$, $BF_{01} = 6.1 * 10^{-25}$, and less likely to elicit response failures, $F(1, 46) = 98.4$, $p < 0.001$, $BF_{01} = 3.0 * 10^{-20}$, than weak items, but no main effect of strength was observed on intrusions, $F(1, 46) < 1$, $BF_{01} = 5.6$. No main effect of list was observed, $F(1, 46) < 1$ for all three response metrics (corrects: $BF_{01} = 6.4$; intrusions: $BF_{01} = 5.3$; response failures: $BF_{01} = 5.7$). We have now observed a null list strength effect in cued recall in three experiments across variations in list length, presentation time, study task, and word frequency, suggesting a robust finding.

Experiment 4: test expectancy

The remaining difference between the prior three experiments and the original positive list strength effect experiment for cued recall is that cued recall blocks from RCS(1990)e6 were intermixed with blocks of single item recognition and free recall. People tend

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

to alter their encoding strategies in response to an anticipated test (Tversky, 1973). For example, Neely and Balota (1981) found that participants who expected a test of free recall outperformed those who expected a test of recognition. Similarly, Hockley and Cristi (1996) demonstrated that focusing on forming associative bindings at study improves performance on associative recognition tests compared to a focus on item encoding. The anticipation of a free recall or recognition test in RCS(1990)e6 may have influenced participants' study strategy during cued recall blocks. For example, participants may have focused on binding studied items more closely to contexts than to the other studied items which might have led to a positive list strength effect. This experiment specifically tests that hypothesis.

Methods

Participants

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

Stimuli

The word pool was the high-frequency word pool from 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 each task there was a pure weak, a pure strong, and a mixed list. All nine study-test blocks were randomly intermixed and the test was post-cued. The mixed list block was arranged in the same manner as in Experiment 2, controlling for study-test lag. For item recognition, one member of each studied pair, randomly chosen, served as a target. Targets were randomly intermixed with an equal number of unstudied foils. For free recall, participants were prompted to recall as many words as they could remember from the studied lists. Participants had 4 min to complete this task and could terminate the recall test at any time. All other details were identical to Experiment 3.

Results

Cued recall

No list strength interaction was observed for correct responses, intrusions, or response failures, $F(1, 53) < 1$ in each case (Fig. 1, Table 2). The data provide evidence for a null list strength effect on correct responses, $BF_{01} = 4.8$, intrusions, $BF_{01} = 4.5$, and response failures, $BF_{01} = 4.8$. A main effect of item strength was observed, with repeated word pairs associated with more correct responses, $F(1, 53) = 86.9$, $p < 0.001$, $BF_{01} = 8.3 * 10^{-19}$, more intrusions, $F(1, 53) = 10.9$, $p = 0.002$, $BF_{01} = 0.15$, and fewer response failures, $F(1, 53) = 129$, $p < 0.001$, $BF_{01} = 1.6 * 10^{-21}$. No main effect of list type was observed for corrects, $F(1, 53) < 1$, $BF_{01} = 6.8$, intrusions, $F(1, 53) = 1.42$, $p = 0.239$, $BF_{01} = 2.7$, or response failures, $F(1, 53) = 1.64$, $p = 0.206$, $BF_{01} = 4.3$.

Single item recognition

Discriminability (Fig. 1 for d', Table 4) was analyzed with a 2 (mixed vs pure lists) $\times$ 2 (strong vs weak items) repeated measures ANOVA. No list strength interaction was observed, $F(1, 53) = 3.41$, $p = 0.07$, $BF_{01} = 1.5$, replicating the null-to-negative list strength effect observed throughout recognition tests in Ratcliff et al. (1990). Strong items were more discriminable than weak items, $F(1, 53) = 104$, $p < 0.001$, $BF_{01} = 6.7 * 10^{-17}$. No main effect of list type on discriminability was observed, $F(1, 53) < 1$, $BF_{01} = 6.1$. For converging evidence, note that the A' measure of discriminability also showed no list strength interaction $F(1, 53) < 1$, $BF_{01} = 3.9$.

For archival purposes, hits and false alarms (Table 4) 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) $\times$ 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. Strong items had more hits than weak items, $F(1, 53) = 124$, $p < 0.001$, $BF_{01} = 9.2 - * 10^{-17}$. An interaction between list type and item strength was observed for hit rates, $F(1, 53) = 4.96$, $p = 0.030$, $BF_{01} = 1.7$. False alarms were highest for the pure weak lists and lowest for pure strong lists with the mixed list false alarm rate falling in between, $F(2, 106) = 9.83$, $p < 0.001$, sphericity assumed, $BF_{01} = 0.004$.

Free recall

The proportion of correct responses (Fig. 1, Table 5 also has intrusions) was analyzed with a 2 (mixed vs pure lists) $\times$ 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 expected, a positive list strength was observed for correct responses, $F(1, 53) = 4.5$, $p = 0.039$, $BF_{01} = 0.88$. Correct responses were greater for strong items than weak items, $F(1, 53) = 81.0$, $p < 0.001$, $BF_{01} = 5.5 * 10^{-15}$. Mixed lists had more correct responses than pure lists, $F(1, 53) = 5.16$, $p = 0.027$, $BF_{01} = 1.2$. No effect of list type was observed on intrusions, $F(2, 106) < 1$, sphericity assumed, $BF_{01} = 16$.

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 with that observed in prior experiments: a null list strength effect in recognition and a positive list strength effect in recall (Hirshman, 1995; Malmberg & Shiffrin, 2005; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1900, 1994; Rose & Sutton, 1996; Sahakyan et al., 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 methodological differences between this experiment and that of RCS(1990)e6 were the following, so far as we can tell: the experiment was completed over two sessions separated by 4–10 days, rather than on three consecutive days, and participants were compensated with class credit rather than financially.

Methods

Participants

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

Stimuli

A pool of 1779 high frequency words (KF frequency > 50, logHAL > 9, 4–11 letters long) was used.

Procedure

Across two sessions (4–10 days apart, separated by a weekend), participants received a total of 3 practice and 21 experimental study-test blocks, 8 each for cued recall, single item recognition, and free recall with the type of test post-cued. For each of the 8

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

	Weak first		Weak last		Weak shuffled	
	Weak	Strong	Weak	Strong	Weak	Strong
Experiment 3						
Corrects	0.074 (0.019)	0.351 (0.044)	0.109 (0.024)	0.346 (0.042)	0.125 (0.026)	0.412 (0.044)
Intrusions	0.213 (0.040)	0.218 (0.037)	0.221 (0.041)	0.239 (0.037)	0.234 (0.041)	0.223 (0.039)
Response failures	0.713 (0.040)	0.431 (0.043)	0.670 (0.044)	0.415 (0.042)	0.641 (0.043)	0.364 (0.040)
Experiment 5						
Corrects	0.079 (0.013)	0.240 (0.034)	0.043 (0.013)	0.281 (0.041)	0.084 (0.019)	0.260 (0.035)
Intrusions	0.156 (0.033)	0.237 (0.035)	0.179 (0.033)	0.204 (0.036)	0.148 (0.028)	0.181 (0.031)
Response failures	0.765 (0.032)	0.523 (0.042)	0.778 (0.034)	0.515 (0.044)	0.767 (0.035)	0.559 (0.042)

Table 4
Means and standard errors (in parentheses) for single item recognition performance, by condition.

	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)
A'	0.812 (0.017)	0.905 (0.013)	0.788 (0.016)	0.899 (0.013)
Hits	0.611 (0.029)	0.852 (0.023)	0.647 (0.021)	0.802 (0.024)
False alarms	0.148 (0.022)		0.225 (0.022)	0.115 (0.017)
Experiment 5				
d'	0.837 (0.076)	1.657 (0.111)	0.801 (0.086)	1.690 (0.123)
A'	0.715 (0.019)	0.844 (0.017)	0.716 (0.021)	0.844 (0.020)
Hits	0.440 (0.023)	0.740 (0.022)	0.573 (0.023)	0.714 (0.026)
False alarms	0.183 (0.020)		0.284 (0.028)	0.165 (0.025)

	Mixed lists					
	Weak first		Weak last		Weak shuffled	
	Weak	Strong	Weak	Strong	Weak	Strong
Experiment 5						
d'	0.825 (0.084)	1.692 (0.133)	0.757 (0.122)	1.458 (0.144)	0.930 (0.105)	1.822 (0.144)
A'	0.714 (0.021)	0.841 (0.024)	0.675 (0.030)	0.807 (0.024)	0.725 (0.024)	0.874 (0.021)
Hits	0.378 (0.033)	0.707 (0.030)	0.482 (0.032)	0.732 (0.032)	0.459 (0.032)	0.781 (0.030)
False alarms	0.149 (0.024)		0.233 (0.027)		0.167 (0.023)	

Table 5
Means and standard errors (in parentheses) for free recall performance, by experiment and condition.

	Mixed		Pure	
	Weak	Strong	Weak	Strong
Experiment 4				
Corrects	0.084 (0.012)	0.279 (0.030)	0.079 (0.009)	0.206 (0.018)
Intrusions	0.055 (0.013)		0.050 (0.010)	0.054 (0.010)
Experiment 5				
Corrects	0.041 (0.007)	0.213 (0.018)	0.066 (0.012)	0.168 (0.018)
Intrusions	0.084 (0.015)		0.082 (0.020)	0.115 (0.033)

	Mixed lists					
	Weak first		Weak last		Weak shuffled	
	Weak	Strong	Weak	Strong	Weak	Strong
Experiment 5						
Corrects	0.028 (0.007)	0.199 (0.026)	0.060 (0.014)	0.186 (0.023)	0.034 (0.008)	0.253 (0.028)
Intrusions	0.079 (0.011)		0.103 (0.030)		0.070 (0.013)	

blocks per test type, 5 of them were arranged as they were in Experiment 3 (pure strong, pure weak, weak first, weak last, and weak shuffled). 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. Inter-item spacing was restricted to between 5 and 15 intervening presentations.

During session 1, participants completed the 3 training blocks (one per retrieval task, order randomized) followed by 9 other blocks selected at random. Participants were informed after the 3 training blocks that they would be tested using those three methods throughout the remainder of the experiment. During session 2, participants completed the remaining 12 blocks in random order. Data from the training blocks were collected, but not analyzed or reported.

Results & discussion

90 people completed both sessions of the experiment. Due to technical problems complete data sets exist for only 49 participants. Analyses were conducted using only data from those 49 participants. We do not analyze the list length manipulation because it is not relevant here, but report descriptive statistics for archival purposes (see Table 6).

Cued recall

A 3 (weak first vs weak last vs weak shuffled) × 2 (weak vs strong) repeated measures ANOVA on mixed list performance (Table 3) revealed no list strength interactions, sphericity assumed (correct responses: $F(2, 96) = 2.13$, $p = 0.124$, intrusions: $F(2, 96) = 1.40$, $p = 0.252$; response failures: $F(2, 96) < 1$). No main effects of list type were observed, corrects and response failures: $F(2, 96) < 1$, intrusions: $F(2, 96) = 1.15$, $p = 0.323$ therefore we collapsed over mixed lists. Cued recall data (Fig. 1, Table 2) were analyzed using 2 (mixed vs pure) × 2 (weak vs strong) repeated measures ANOVAs. Again, we find evidence for a null list strength effect for correct responses, $F(1, 48) < 1$, $BF_{01} = 3.6$. Strong pairs elicited more correct responses and fewer response failures than weak pairs (corrects: $F(1, 48) = 66.4$, $p < 0.001$, $BF_{01} = 5.9 * 10^{-17}$; response failures: $F(1, 48) = 44.6$, $p < 0.001$, $BF_{01} = 1.6 * 10^{-11}$). We observed no main effect of strength on intrusions, $F(1, 48) < 1$, $BF_{01} = 5.3$, and no main effect of list type, $F(1, 48) < 1$ (corrects: $BF_{01} = 6.4$; intrusions: $BF_{01} = 5.6$; response failures: $BF_{01} = 6.3$). A list strength interaction was observed for response failures $F(1, 48) = 5.32$, $p = 0.025$, $BF_{01} = 0.67$, but not intrusions $F(1, 48) = 3.47$, $p = 0.069$, $BF_{01} = 0.76$.

Single item recognition

A 3 (weak first vs weak last vs weak shuffled) × 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: $F(2, 96) = 2.25$, $p = 0.090$; interaction: $F(2, 96) = 1.26$, $p = 0.307$), therefore we collapsed the d' data for mixed lists (Fig. 1). We conducted a 2 (mixed vs pure) × 2 (weak vs strong) repeated measures ANOVA. There is evidence for a null list strength effect in d', $F(1, 48) < 1$, $BF_{01} = 4.3$. Strong words elicited a higher d' than weak words, $F(1, 48) = 135$, $p < 0.001$, $BF_{01} = 5.9 - * 10^{-21}$. No main effect of list type was observed, $F(1, 48) < 1$, $BF_{01} = 6.5$. We also find evidence for a null list strength effect when discriminability is measured via A', $F(1, 48) < 1$, $BF_{01} = 4.9$, see Table 4.

The hit rates are analyzed in a 4 (weak first vs weak last vs weak shuffled vs pure) × 2 (weak vs strong) repeated measures ANOVA, and the false alarm rates in a 1-way (weak first vs weak last vs weak shuffled vs pure weak vs pure strong) repeated measures ANOVA Strong words elicited higher hit rates than weak words, $F(1, 48) = 163$, $p < 0.001$, $BF_{01} = 2.0 * 10^{-30}$. A main effect of list was observed, $F(3, 144) = 4.73$, sphericity assumed, $p = 0.004$, $BF_{01} = 1.1$: pure lists elicited the most hits, followed by lists where the weak words were shuffled, followed by weak last, followed by weak first. A strength by list type interaction was observed for hit rates, $F(3, 144) = 6.19$, sphericity assumed, $p < 0.001$, $BF_{01} = 0.053$. False alarms differed across list type, $F(4, 192) = 10.7$, $p < 0.001$, sphericity assumed, $BF_{01} = 5.6 * 10^{-6}$, see Table 4.

Free recall

Free recall correct response rates (Fig. 1, Table 5) were analyzed in a 4 (weak first vs weak last vs weak shuffled vs pure) × 2 (weak vs strong) repeated measures ANOVA. A list type by strength interaction was observed for the proportion of correctly recalled words, $F(3, 144) = 4.92$, sphericity assumed, $p = 0.003$, $BF_{01} = 0.13$, indicating evidence for a positive list strength effect. The difference in mean performance between strong and weak words was largest for mixed lists in this order: weak shuffled, weak first, weak last. A positive list strength effect was observed, a conclusion supported by collapsing across mixed list conditions and re-analyzing the data in a 2 (mixed vs pure) × 2 (weak vs strong) repeated measures ANOVA. A list strength interaction for corrects, $F(1, 48) = 9.30$, $p = 0.004$, $BF_{01} = 0.11$, indicated a positive list strength effect in this task. Strong words were more likely to be recalled than weak words, $F(1, 48) = 113$, $p < 0.001$, $BF_{01} = 3.8 * 10^{-28}$. No main effect of list was observed, $F(3, 144) = 1.34$, $p = 0.263$, sphericity assumed, $BF_{01} = 30$. The results of this experiment are consistent with those 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 found a null list strength effect in cued recall.

Cross experiment meta analyses of cued recall

Here we combine data across the experiments to evaluate whether there is evidence for a null list strength effect (as indicated for each individual experiment) or whether there is evidence for a very small effect that is not captured by any single experiment. To do so, we averaged across the mixed lists in each experiment for the cued recall data and analyzed the set in a 2 (mixed vs pure) × 2 (weak vs strong) × 5 (Experiment 1 vs 2 vs 3 vs 4 vs 5) repeated measures ANOVA. We compute the Bayes factor for this analysis by comparing the null-hypothesis models that exclude the list strength interaction effect to the various alternative-hypothesis models that include the list strength interaction using the default prior in JASP. The Bayes factor BF_{01} may be interpreted in this case as a model-averaged ratio of evidence in favor of the potential "null" models (the list strength interaction was excluded) to evidence in favor of the potential models that include the list strength interaction.

Critically, the finding of a null list strength effect in cued recall is strong and robust. A null list strength effect was observed for correct responses, $F(1, 224) = 2.39$, $p = 0.124$, $BF_{01} = 19$. This null effect does not change with experiment, $F(4, 224) < 1$, sphericity assumed, $BF_{01} = 5500$. For intrusions, no list strength interaction was observed, $F(1, 224) = 2.59$, $p = 0.109$, $BF_{01} = 90$, and this does not vary with experiment, $F(4, 224) = 1.30$, sphericity assumed, $p = 0.271$, $BF_{01} = 6100$. A list strength interaction was present for response failures, $F(1, 224) = 7.414$, $p = 0.007$, but the difference is so small that it is evidence for a null list strength effect, $BF_{01} = 3.9$. This effect did not change with experiment, $F(4, 224) = 1.40$, sphericity assumed, $p = 0.237$, $BF_{01} = 100$.

General discussion

Across five experiments, we observe a null list strength effect in cued recall alongside a positive list strength effect in recall and a null list strength effect in recognition. Because the list strength effect in free recall and recognition have been widely replicated, modeled, and discussed, we focus solely on cued recall.

The motivation for this series of studies was that a better understanding of cued recall may help aid in the development of a unified theory of memory across multiple memory tasks. As we noted, many "second generation" models account for either free recall or recognition, but not both, and the global matching models were unable to account for the Ratcliff et al. (1990) findings. This leaves only Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997), its predecessor SAM, and the dual process models as viable accounts for these data. We consider implications for these two classes of models and then consider implications for model development.

Table 6
Performance on long weak and short strong lists in Experiment 5, by task. SEM is the standard error of the mean.

	Short strong		Long weak	
	Mean	SEM	Mean	SEM
Cued recall				
Corrects	0.269	(0.037)	0.049	(0.009)
Intrusions	0.190	(0.029)	0.143	(0.030)
Response Failures	0.541	(0.042)	0.808	(0.030)
Single item recognition				
d′	1.686	(0.157)	0.670	(0.087)
HR	0.732	(0.030)	0.557	(0.027)
FAR	0.163	(0.028)	0.307	(0.023)
Free recall				
Corrects	0.268	(0.030)	0.042	(0.007)
Intrusions	0.097	(0.019)	0.047	(0.001)

REM and SAM, like many of their counterparts, consider two primary sources of information in episodic memory—item and context. Item information reflects the content of the item including meaning whereas context refers to information in the external and internal environment that is not related to the item (e.g., temperature of the room, color of the walls, mood of the participant). Further, they assume that each type of information is stored, updated, and retrieved with some degree of independence. That is, each type of information can be used to cue memory and the evidence generated by item and context cues can be differentially weighted (Clark & Shiffrin, 1987, 1992; Criss & Shiffrin, 2004; Raaijmakers & Shiffrin, 1980). More importantly, context information is nearly identical (even if it is slowly drifting across trial ala Estes, 1955 or Mensink & Raaijmakers, 1989) for each memory trace in a given experiment and item information is quite a bit different for each trace. Of course, similarity can be manipulated by encoding in different contexts (Lehman & Malmberg, 2009; Murnane & Phelps, 1993, 1995; Murnane, Phelps, & Malmberg, 1999; Park, Arndt, & Reder, 2006; Smith, 1979) or including words from the same category (Criss, 2006; Roediger & McDermott, 1995; Shiffrin, Huber, & Marinelli, 1995), but in a standard experiment like the ones presented here, context information is shared across memory traces and item information is not. The fact that free recall and recognition primarily rely on different cues—context for free recall and item (plus context) for recognition—and that differences in the similarity of these cues and the contents of memory cause different levels of interference—is what produces qualitatively different list strength effects in these paradigms.

For free recall, REM samples a single trace based on how well the context cue matches the context stored in each trace. Sampling is based on a Luce choice rule which results in a competition between traces. On a mixed list, repeated items have more strongly encoded contexts than weakly encoded items. 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 pure lists. The result is a positive list strength (Malmberg & Shiffrin, 2005; Shiffrin et al., 1990) in free recall.

Cued recall, by definition, includes item information in the cue. The contribution of context will be as just described, inducing a positive list strength effect, but what will the item cue contribute? The centerpiece of REM is the assumption of differentiation during encoding (Criss, 2006; Criss & Koop, 2015; Criss & McClelland, 2006; Murnane & Shiffrin, 1991a, 1991b; Shiffrin et al., 1990). Every individual event is encoded as a trace in memory and more information is added to that trace when it is remembered as having appeared in the same context. The result of this updating is that a when a target item is presented, it better matches the memory trace that represents its previous occurrence compared to a weakly encoded memory trace. At the same time, the target item matches *other* strongly encoded memory traces less well than weakly encoded memory traces. Strongly encoded mismatching traces are therefore less likely to be confused as a matching trace than weakly encoded ones. A pure strong list contains all strongly differentiated items whereas a mixed list brings more noise (in the form of incorrectly matching weakly encoded traces) to the decision. The result is a null to negative list strength effect in recognition.

In principle, cued recall requires cuing from both context (which leads to a positive list strength effect) and item (which leads to a negative list strength effect). However, as we describe next the two implementations of cued recall in REM do not meet that requirement. Diller, Nobel, and Shifrrin (2001) published a response time model of cued recall that has various quirks (e.g., response time is unrelated to accuracy) which were relevant for the specific data under consideration but does not make for a general model. With respect to the list strength paradigm, the model implementation uses an item-only cue which predicts a negative list strength effect, inconsistent with the data.

Some have drawn parallels between the retrieval induced forgetting paradigm and the list strength effect considered here (e.g., Bauml, 1997) and have modeled the former within the REM framework (Verde, 2009, 2013). In a typical experiment, individuals study words alongside a category label (FRUIT-grape, FRUIT-apple, VEHICLE-car, VEHICLE-motorcycle) and then practice through cued recall the cue-target pairings for half of the pairs of one category (FRUIT-grape). At test the category label and a letter of the target are presented as the retrieval cue (VEHICLE-c___). One of the key findings is that non-practiced words associated with practiced cues (apple) have worse memory than non-practiced words associated with non-practiced cues (car; Anderson et al., 1994; Norman, Newman, & Detre, 2007; Raaijmakers & Jakab, 2012). To directly compare this paradigm with the list strength effect described here is an oversimplification that ignores factors critical to the list strength predictions described above including fan effects or category length effects (e.g., Anderson, 1974; Criss & Shiffrin, 2004; Shiffrin et al., 1995), the role of item similarity in interference and the fact that similar items lead to reversal of differentiation (Criss, 2006). Indeed in Verde's (2009, 2013) application of REM to the retrieval induced forgetting paradigm, he treated the category label as context and the category members as independent items without shared similarity. In other words, he implemented a free recall paradigm and found a positive list strength effect. This cannot be generalized to the experiments reported here because each item cue is unique and both item and context cues are necessary. Nevertheless, this is a potentially fruitful point of entry to study the relationship between retrieval induced forgetting and the standard list strength paradigm.

To successfully account for cued recall and the null list strength effect in particular, we suggest a model including both context and item information in the cue (cf. Criss & Shiffrin, 2004). A pure con-

text cue induces a positive list strength effect but also poor performance because cued recall requires the retrieval of an item tied to the item cue. A pure item cue induces a negative list strength effect and also the strong possibility of extra-list recall (e.g., recalling semantic associates). Thus it seems necessary to assume that some combined cue would lead to reasonable levels of accuracy and a null list strength effect, as observed in the data. An exact prediction would depend on the implementation but preliminary analysis supports the basic intuition (Wilson, 2015). This suggests that the SAM/REM framework is a viable theory that accounts for multiple memory tasks including free recall, cued recall, and recognition and has a coherent theoretical explanation for the different empirical patterns of the list strength effect across task.

As just described, REM and other models describe different processes for recall and recognition. Other models assume that recognition includes two independent processes of recollection and familiarity. The many models vary in their operational definition of recollection and familiarity (and how the two contribute to memory), but generally recollection is described as similar to recall and familiarity as a non-specific global match. Within this dual process framework, Norman (2002) predicted and observed a positive list strength effect for recollection based responses but not familiarity based responses (see also Diana & Reder, 2005). We find a null list strength effect for cued recall, which is certainly a recollection based task. This presents problems for, or limitations on, the general idea that positive list strength effects are found for recollection based tasks.

In all, we find that the list strength effect in cued recall tends towards the null. The variety of manipulations conducted in these experiments suggests that this finding is robust, and the amount of evidence obtained across the experiments overwhelms the prior set by the original finding. This poses issues for dual process theories of recognition, but may be accounted for in the SAM/REM framework by allowing context and items to both contribute to a compound cue. Along this dimension, at least, cued recall offers a theoretical middle-ground by which to simultaneously assess theories of recall the theories of recognition and, in turn, make progress towards a unified account of memory across a variety of test tasks.

Author note

This work was partially supported by the National Science Foundation (grant number 0951612) awarded to A.H. Criss. Data and analyses are available for download at <http://memolab.syr.edu/publications.html>.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6*, 451–474.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 1063–1087.

Bauml, K.-H. (1997). The list strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin & Review, 4*(2), 260–264.

Clark, S. E., & Shiffrin, R. M. (1987). Recognition of multiple-item probes. *Memory & Cognition, 15*(5), 367–378.

Clark, S. E., & Shiffrin, R. M. (1992). Cuing effects and associative information in recognition memory. *Memory & Cognition, 20*(5), 580–598.

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. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59*, 297–319.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(2), 484–499.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language, 64*, 119–132.

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

Criss, A. H., & Koop, G. J. (2015). Differentiation in episodic memory. In J. Raaijmakers, A. H. Criss, R. Goldstone, R. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin* (pp. 112–125). Psychology 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. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review, 111*(3), 800–807.

Criss, A. H., Wheeler, M. E., & McClelland, J. L. (2013). A differentiation account of recognition memory: Evidence from fMRI. *Journal of Cognitive Neuroscience, 25*(3), 421–435.

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

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory & Cognition, 33*(7), 1289–1302.

Diller, D. E., Nobel, P. A., & Shifrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 414–435.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* Report No. AFCRC-TN-58-51. Hearing and Communication Laboratory, Indiana University.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145–154.

Grider, R. C., & Malmberg, K. J. (2008). Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory & Cognition, 36*(5), 933–946.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*(6), 424–429.

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., & 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.

JASP Team (2016). *JASP (Version 0.7.5.5)* [Computer software].

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.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*(4), 970–988.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74*(2), 100–109.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*, 203–208.

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. W. (1989). A model of contextual fluctuation. *Journal of Mathematical Psychology, 33*, 172–186.

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

Murdock, B. B., & Kahana, M. J. (1993b). List-strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel (1993). *Journal of Experimental Psychology, Learning, Memory, and Cognition, 19*(6), 1450–1453.

Murnane, K., & Phelps, M. P. (1993). A global activation approach to the effect of changes in environmental context on recognition. *Journal of Experimental Psychology: Learning, Memory & Cognition, 19*(4), 882–894.

Murnane, K., & Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(1), 158–172.

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General, 128*(4), 403–415.

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

Murnane, K., & Shiffrin, R. M. (1991b). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*(5), 855–874.

Neely, J. H., & Balota, D. B. (1981). Test-expectancy and semantic organization effect in recall and recognition. *Memory & Cognition, 9*(3), 283–300.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory & Cognition, 28*(6), 1083–1094.

Norman, K. A., Newman, E. L., & Detre, G. (2007). A neural network model of retrieval- induced forgetting. *Psychological Review, 114*, 887–953.

Park, H., Arndt, J., & Reder, L. M. (2006). A contextual interference account of distinctiveness effects in recognition. *Memory & Cognition, 34*(4), 743–751.

Pollack, I., & Norman, D. A. (1964). A nonparametric analysis of recognition experiments. *Psychonomic Science, 1*, 125–126.

Raaijmakers, J. G. W., & Jakab, E. (2012). Rethinking inhibition theory: On the problematic status of the inhibition theory of forgetting. *Journal of Memory & Language, 68*, 98–122. http://dx.doi.org/10.1016/j.jml.2012.10.002.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. *Psychology of Learning and Motivation – Advances in Research and Theory, 14*(C), 207–262.

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.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518–535.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(4), 803–814.

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

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.

Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavioral Research Methods, 37*(1), 3–10.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(2), 267–287.

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.

Shiffrin, R. M., Ratcliff, R. C., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19*(6), 145–149.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 460–471.

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

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory & Cognition, 38*(5), 1137–1151.

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92*(3), 297–304.

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

Verde, M. F. (2009). The list-strength effect in recall: Relative-strength competition and retrieval inhibition may both contribute to forgetting. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*(1), 205–220.

Verde, M. F. (2013). Retrieval-induced forgetting in recall: Competitor interference revisited. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*(5), 1433–1448.

Wilson, J. H. (2015). *The list strength effect in cued recall: Estimation, implications, and models* Unpublished master's thesis. Syracuse, NY: Syracuse University.

Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure- and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 523–538.