Journal of Memory and Language 63 (2010) 18–34

Contents lists available at ScienceDirect

Journal of Memory and Language

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

A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory

Jeffrey J. Starns *, Corey N. White, Roger Ratcliff

Department of Psychology, Ohio State University, United States

ARTICLE INFO

Article history:
Received 2 July 2009
revision received 3 March 2010
Available online 3 April 2010

Keywords:
Strength-based mirror effect
Differentiation
Retrieving Effectively from Memory
Bind-Cue-Decide Model of Episodic Memory

ABSTRACT

We explore competing explanations for the reduction in false alarm rate observed when studied items are strengthened. Some models, such as Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997), attribute the false alarm rate reduction to differentiation, a process in which strengthening memory traces at study directly reduces the memory evidence for lure items. Models with no differentiation mechanism, such as the Bind-Cue-Decide Model of Episodic Memory (BCDMEM; Dennis & Humphreys, 2001), explain the false alarm rate reduction in terms of the strength of items expected at retrieval. To contrast these explanations, we separately manipulated item strength at encoding and retrieval. Participants studied mixed lists of weak and strong items. Weak items were always presented once. On separate lists, strong items were either presented twice (strong-2X) or five times (strong-5X). Within each strength condition, participants completed separate tests with mixed (strong and weak) targets, pure weak targets, or pure strong targets. They were correctly informed of the type of target on each test. Results showed that false alarm rates decreased from the strong-2X condition to the strong-5X condition for the mixed and pure-strong tests, but not for the pure-weak tests. That is, false alarm rates were determined by the strength of targets appearing on the test, not by the content of the study list. The results support BCDMEM's expectation-based explanation and not REM's differentiation-based explanation.

Introduction

In a typical recognition memory task, participants study a list of words and are later asked to discriminate these studied words (targets) from non-studied words (lures). When memory is strengthened by increasing the duration or number of learning trials, performance usually improves in two ways: participants are better able to identify the studied words themselves (i.e., the hit rate increases), and they are also better able to reject the non-studied words (i.e., the false alarm rate decreases). This empirical phenomenon is referred to as the strength-based mirror effect (Criss, 2006; Ratcliff, Clark, & Shiffrin, 1990; Stretch & Wixted, 1998).

Recently, some theorists have interpreted the strength-based mirror effect as evidence for differentiation, a mechanism that produces false alarm rate reductions based on the inherent properties of strong memory traces (Criss, 2006, 2009; McClelland & Chappell, 1998). One popular differentiation model is Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997). In this model, items are represented by vectors of feature values that are probabilistically copied into episodic traces at encoding. At retrieval, the features of the test word are matched to the features stored in memory. Lures produce only a random match to each memory trace, whereas targets strongly match the one trace that was formed when the target was studied. Storing more features for an item (i.e.,

* Corresponding author at: Ohio State University, 225 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210, United States.
E-mail address: starns.4@osu.edu (J.J. Starns).

0749-596X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2010.03.004

strengthening the item) increases the match between the memory trace and the encoded item itself, which produces an increase in the hit rate. Critically, storing more features also *decreases* the match between the memory trace and the features of any other item. That is, with more feature values stored for studied items, there are more opportunities to observe mismatches when a non-studied item is tested. This differentiation mechanism produces a lower false alarm rate for stronger lists. Other models have different methods to implement differentiation (McClelland & Chappell, 1998; Shiffrin, Ratcliff, & Clark, 1990), but the result is always that strong memory traces directly decrease memory evidence for lure words. Therefore, all differentiation models predict false alarm rate differences based on the degree of learning at study.

Alternatives to differentiation explain the false alarm rate reduction in terms of changing expectations at retrieval. The standard expectation-based account posits a change in the response criterion in a standard signal-detection model (Hirshman, 1995; McCabe & Balota, 2007; Ratcliff et al., 1990; Stretch & Wixted, 1998). By this account, the false alarm rate decreases because participants adopt a more conservative criterion when they expect to have strong memories for target items (Brown, Lewis, & Monk, 1977). Proponents of differentiation have stressed that signal detection models have no mechanisms to explain how memory evidence is produced and usually fail to specify the processes involved in setting the response criterion (Criss, 2006). In contrast, differentiation models specify the encoding and retrieval mechanisms that underlie memory evidence and the decision processes that translate evidence values into overt decisions.

One goal of the current study is to show that the expectation-based account can be implemented in a computational model that specifies memory and decision processes to the same level of detail as the differentiation models. As we discuss shortly, the Bind-Cue-Decide Model of Episodic Memory (BCDMEM; Dennis & Humphreys, 2001) evaluates memory evidence based on the expected degree of learning for targets, with lower false alarm rates predicted when stronger targets are expected. Given that computational models can produce the strength-based mirror effect based on either differentiation or changes in retrieval expectations, we sought to distinguish these accounts by independently manipulating the strength of items on the study list and the expected strength of targets on the test.

In all of our experiments, participants completed 12 study/test cycles. In each cycle, they studied mixed-strength lists with half strong and half weak items. Weak items were always presented once. Strong items were presented twice on half of the lists (strong-2X condition) and five times on the remaining lists (strong-5X condition). Test lists contained lures mixed with either weak targets, strong targets, or both, and this variable was crossed with the number of study repetitions for the strong items. Before each test, participants were informed of the type of targets that would appear on the test (see McCabe & Balota, 2007, for a similar instructional manipulation). A differentiation-based account predicts a lower false alarm rate following strong-5X lists than strong-2X lists regardless of the test content. For example, in REM the memory traces for items presented five times produce more mismatches to lure items than the traces for twice-presented items. In other words, the items presented five times lower the false alarm rate simply by virtue of being studied and stored in memory – the items do not need to appear on the test. Alternatively, if false alarm rate reductions are based on the expected strength of memories at retrieval, then false alarm differences between strong-2X and strong-5X lists should arise only when the strong items are expected at test. Although our primary intent is to test the differentiation and expectation-based hypotheses in general, we also explored the specific predictions of two quantitative models from each category.

BCDMEM

In this section, we provide a brief description of BCDMEM, explain how this model implements an expectation-based account for the strength-based mirror effect, and report simulation results for our paradigm. In BCDMEM, items are represented as local nodes. Context is represented as a distributed set of features that are either "active" or "inactive" (i.e., that take on a value of 1 or 0, respectively). At encoding, a context vector is established to represent all contextual details of the study list, including the environmental context (such as the room in which the words were learned) and the processing context (such as the encoding strategy used by participants). The sparsity parameter (s) defines the proportion of active nodes in the context vector. Links are established between an item node and the context features active during the item's presentation with a probability given by the learning parameter, r. In addition, items may be pre-experimentally linked to features of the studied context because they have been encountered in similar contexts. Such pre-experimental links occur with probability p.

At test, a reinstated context is formed to represent the study list context. A forgetting parameter, d, gives the probability that a feature that was active in the studied context will be inactive in the reinstated context. For each test item, the model retrieves the set of context features that have been linked to the item, both pre-experimentally and in the learning phase. The retrieved context vector is matched to the reinstated context vector, and the model produces a decision based on the observed pattern of matches and mismatches (henceforth referred to as "the data"). Specifically, the model computes a likelihood ratio by dividing the probability that the data would arise if the test word was a target by the probability that the data would arise if the test word was a lure:

$$\begin{aligned} &P(\text{data}|\text{target})/P(\text{data}|\text{lure}) \\ &\quad = \{[1-s+d(1-r_{\text{EST}})s]/[1-s+ds]\}^{n00}(1-r_{\text{EST}})^{n10} \\ &\qquad \{[p(1-s)+d(r_{\text{EST}}+p-r_{\text{EST}}p)s]/[p(1-s)+dps]\}^{n01} \\ &\qquad [(r_{\text{EST}}+p-r_{\text{EST}}p)/p]^{n11} \end{aligned} \quad (1)$$

where $n00$ is the number of features with a value of zero in both the reinstated and retrieved context vectors, $n10$ is the number of features with a value of 1 in the reinstated

vector and 0 in the retrieved vector, and so forth. As is evident, the likelihood calculation depends on all of the parameter values, so the model must have an estimate of each of these values at retrieval. Instead of treating these estimates as free parameters, we constrained them to match the actual parameter values (as did Dennis & Humphreys, 2001). We use different notation for the actual learning parameter used to encode items (r) and the estimated learning parameter at test (r_{EST}) because will we frequently discuss them individually. The true r value affects responding by determining the number of matching features for target items (i.e., targets encoded with a higher r will retrieve more features of the study context), and the r_{EST} parameter influences how the observed matches and mismatches affect the likelihood calculation (more on this shortly).

To implement a Bayesian decision rule, the model computes an odds ratio by multiplying the likelihood ratio by the prior probability ratio. All of our experiments used tests with an equal number of targets and lures, so the prior probability ratio was 1 and the odds ratio equaled the likelihood ratio. In our simulations, the model produced a "yes" response when the test word was more likely to be a target (i.e., the odds ratio was greater than one), and a "no" response when the test word was more likely to be a lure (i.e., the odds ratio was less than one).

The critical issue for our purposes is how the model produces decreases in the false alarm rate when studied items are strengthened. The pattern of matches and mismatches observed for lure items in no way depends on the extent of learning for the studied items – the context features retrieved for lures depend only on pre-experimental links. In other words, there is no differentiation mechanism whereby lures produce more mismatches when targets are strengthened, as in REM (Shiffrin & Steyvers, 1997). Instead, false alarm rate differences are predicted due to changes in the estimate of the learning parameter used to compute the likelihood ratio at retrieval (r_{EST}).

The role of the r_{EST} parameter in moderating false alarm rates is most easily seen by considering a feature that is part of the reinstated context but missing in the retrieved context, a 1–0 mismatch in Dennis and Humphreys' terminology. The expected probability of this pattern for a target item is

$$s(1-d)(1-r_{EST})(1-p)$$

that is, the estimated joint probability that the feature was in the study context vector (s), was not forgotten in the reinstated context vector ($1-d$), and was neither learned at encoding ($1-r_{EST}$) nor linked to the item pre-experimentally ($1-p$). The probability for a lure item is

$$s(1-d)(1-p)$$

which is, of course, the same as the target probability without the added condition of failed learning. Dividing the equations above reveals that each 1–0 mismatch has a likelihood ratio of $1-r_{EST}$ (corresponding to the term raised to the $n10$ power in Eq. (1)). Clearly, as the learning estimate increases, 1–0 mismatches increasingly favor the hypothesis that the test item is a lure, i.e., the likelihood ratio drops farther below 1. Although strengthening the studied items does not lead to more 1–0 mismatches for lures, the learning estimate will be higher if the strong items are expected at test, and the false alarm rate will decrease as a consequence.

In our simulations of BCDMEM, the structure of the study and test lists matched our empirical paradigm. Study lists in the strong-2X condition contained 12 items studied once (with learning parameter r_1) and 12 item studied twice (with learning parameter r_2). Study lists in the strong-5X condition contained 12 items studied once (with learning parameter r_1) and 12 items studied five times (with learning parameter r_5). We ran separate simulations with mixed-strength tests, pure-weak tests, and pure-strong tests. The parameters used in the simulations are reported in Table 1. Parameter values were selected to be close to those used by Dennis and Humphreys (2001) and to produce performance levels near those observed in our experiments.

The results of our simulations depend on how the strength estimate used at retrieval (r_{EST}) varies based on participants' expectations. Three major sources of information can influence these expectations: the content of the study list (see Hirshman, 1995), the instructions provided at test (see McCabe & Balota, 2007), and the actual content of the test (see Verde & Rotello, 2007). Dennis and Humphreys (2001) neither prioritize nor dismiss any of these information sources, although they generally posit a

Table 1
Parameter values used for model simulations and produced in model fits for Experiment 1.

REM parameters				BCDMEM parameters			
Par.	Sim.	Fit	Description	Par.	Sim.	Fit	Description
w	20	20	Vector length	v	20	20	Vector length
g	.475	.519	Feature frequency base rate	s	.2	.229	Sparsity – proportion of active features in study context
g_r	.4	.424	Base rate used in retrieval calculations	p	.32	.477	Probability of pre-experimental links to context features
c	.7	.642	Probability of correctly copying stored feature	d	.6	.607	Probability of forgetting a feature in the reinstated context
μ_1	.115	.195	Probability of storing a feature across all learning attempts for weak items	r_1	.37	.356	Probability of linking item and context features (learning) for weak items
μ_2	.185	.250	Learning probability for strong-2X items	r_2	.50	.621	Learning probability for strong-2X items
μ_5	.362	.384	Learning probability for strong-5X items	r_5	.85	.884	Learning probability for strong-5X items

Note: Par. = parameter; Sim. = simulation.

prominent role for instructions in their simulations. For instance, to model process dissociation data they assume that the test instructions determine which list context is reinstated at retrieval.[1] Just as test instructions can focus participants on a particular list context, we used instructions to focus participants on a particular strength class in defining their expectations for the test. In the mixed test condition, initial instructions informed participants that they would always be tested on both strong and weak items, and they were reminded of this before each individual test. In the pure test condition, initial instructions informed participants that each test would have only strong or only weak targets. Before each test, they were (correctly) informed of whether the strong targets or the weak targets would be tested. To make sure they attended to this message, participants had to type in the number of times they had studied the targets before the test began. These procedures ensured that participants' expectations matched the actual test content. Accordingly, in our applications of BCDMEM, we set the r_{EST} parameter equal to the actual learning parameter (r) of the tested items (we used the average learning parameter for strong and weak items on mixed tests).[2]

Fig. 1 shows the simulation results. For the mixed and pure-strong tests, the model produces the standard strength-based mirror effect. Increased learning for the strong items leads to a higher hit rate in the strong-5X than the strong-2X condition. For the weak targets and lures, the pattern of matches does not change, but the proportion of "studied" responses decreases because the r_{EST} parameter used in the retrieval calculation is higher (because all or some of the targets on the test are stronger). For the pure-weak tests, the expected degree of learning is based on only the weak targets, so it does not change from the strong-2X to the strong-5X condition. Therefore, BCDMEM predicts no change in the hit rate or false alarm rate on the pure-weak tests. To be clear, BCDMEM can produce predictions other than the ones reported here under different assumptions regarding participants' expectations. Again, we see no reason why full knowledge of the test content would fail to guide their expectations, so our simulations are the most direct method for applying BCDMEM to our paradigm.

As noted by Dennis and Humphreys (2001), BCDMEM predicts null list strength effects; that is, strengthening a subset of items on a list does not impair memory for the non-strengthened items. Our simulations also show this property, in that memory performance for weak items does not change based on whether the strong items were studied twice or five times. This is most apparent with pure-weak tests, in which the model predicts no changes in hit rate or false alarm rate. In the mixed tests, both the weak hit rate and the false alarm rate decrease from the strong-2X to the strong-5X condition, but the model predicts no change in discriminability as measured by d' (.70 for strong-2X and .69 for strong-5X).

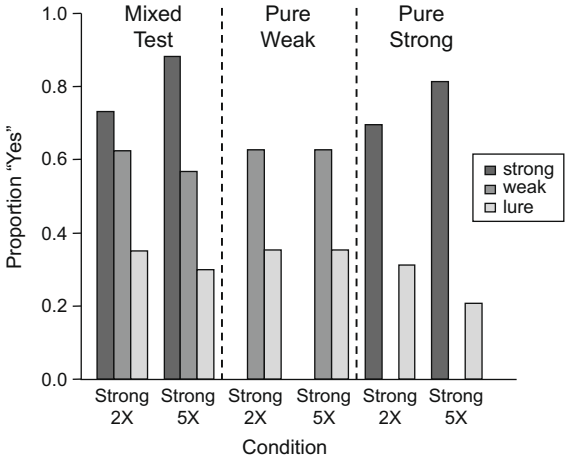

Fig. 1. Simulation results for BCDMEM. The three panels represent the different test conditions, with mixed tests including both weak and strong targets and pure tests containing only weak or only strong. Within each test condition, the left set of bars gives results from the strong-2X condition and the right set of bars gives results for the strong-5X condition. The darkest bars show results for strong targets (studied either 2 or 5 times), the light gray bars show results for weak targets (always studied once), and the white bars show results for lures. Naturally, results for strong targets are absent in the pure-weak test panel and results for weak targets are absent in the pure-strong test panel.

REM simulations

We ran simulations of REM to highlight how the predictions of a differentiation model differ from those of BCDMEM. We consider the basic recognition model proposed by Shiffrin and Steyvers (1997). Details on how to apply this model to recognition experiments can be found in Shiffrin and Steyvers (also see Criss, 2006, for a clear discussion of the model mechanisms underlying differentiation). The model parameters used in our simulations are reported in Table 1. We used parameter values that were close to those used by Shiffrin and Steyvers with a few slight adjustments to produce performance levels similar to those observed in our experiments. Some parameters define properties of the vectors used to represent items, such as the number of features for each item (w) and the probability parameter of the geometric distribution used to generate the feature values (g). Other parameters govern the encoding process. Namely, μ^* is the probability of storing a feature on each learning attempt, c is the probability of correctly copying a stored feature value from the lexical trace, and t is the number of learning attempts. We selected three t values for items presented once, twice, or five times to produce levels of performance near those observed in our experiments. As discussed by Criss (2006,

[1] In a paragraph in which they liken the processes needed to set BCDMEM's participant-controlled parameters (such as r_{EST}) to the processes needed to set response criteria in other models, Dennis and Humphreys also acknowledge the common assumption that "participants are capable of altering criteria as a consequence of instructions" (p. 456, note that they do not notationally distinguish the actual and estimated learning parameters as we have done).

[2] The qualitative predictions are not dependent on the assumption that the estimated learning parameter exactly matches the actual learning parameter, but this assumption is useful because it powerfully constrains the model. Even if error is introduced into the estimates, the model makes the same qualitative predictions under the reasonable assumption that participants expect stronger items when they know the test will contain items with more leaning attempts.

footnote 1), μ^* and t can be combined into the total probability of storing a feature across all learning attempts, $\mu = 1 - (1 - \mu^*)^t$. The μ values in our simulations correspond to $\mu^* = .04$ with three storage attempts for items studied once, five for items studied twice, and eleven for items studied five times.

At retrieval, the model computes a ratio contrasting the likelihood that the observed pattern of matches and mismatches would occur if the episodic trace was from the tested item vs. from a different item. To calculate these likelihoods, the model must have an estimate of both the c parameter and the g parameter (see Shiffrin & Steyvers, 1997, for details). As is customary, we simply used the correct c value in the retrieval calculations, but we used the default base-rate estimate of .4 (g_r) instead of the exact value used to construct our stimuli ($g = .475$).[3] The test decision is based on the average of the likelihood ratios produced by matching the test candidate to each trace in the memory set (usually all of the traces from the study list). Averaging yields the total odds that the test item was studied vs. not studied, with values above 1 indicating that the test candidate is most likely a studied item and values below 1 indicating the opposite. Like BCDMEM, REM's default is to produce a "studied" response if the odds value is above 1 and a "not studied" response otherwise. REM can adjust the criterion odds value needed to produce a "studied" response to accommodate special circumstances, such as tests with unequal proportions of targets and lures or with differential payoffs for "studied" vs. "not studied" responses. Our tests had equal numbers of targets and lures and no payoff structure, so we left the response criterion at its natural value of 1 instead of treating it as a free parameter.

We structured the study and test lists in the same way as in the BCDMEM simulations. The simulation results for REM appear in Fig. 2. The first panel shows the result of mixed-strength tests. The strong hit rate increases from the strong-2X to the strong-5X condition. The additional learning attempts fill in more features in the episodic traces, leading to a better match between each trace and its corresponding target. In contrast, both the weak hit rate and the false alarm rate decrease from the strong-2X to the strong-5X condition. The additional learning for the strong items increases the number of mismatches between the strong memory traces and all other items. When weak targets and lures are matched to the entire memory set, they produce a low likelihood ratio for each strong memory trace, especially in the strong-5X condition.

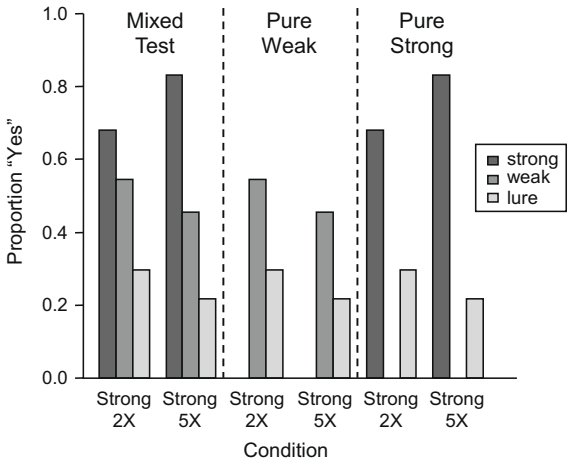

Fig. 2. Simulation results for the REM model. The three panels represent the different test conditions, with mixed tests including both weak and strong targets and pure tests containing only weak or only strong. Within each test condition, the left set of bars gives results from the strong-2X condition and the right set of bars gives results for the strong-5X condition. The darkest bars show results for strong targets (studied either 2 or 5 times), the light gray bars show results for weak targets (always studied once), and the white bars show results for lures. Naturally, results for strong targets are absent in the pure-weak test panel and results for weak targets are absent in the pure-strong test panel.

The model predictions for the pure-weak and pure-strong tests are consistent with those for the mixed-strength test. Critically, on the pure-weak test, the weak hit rate and the false alarm rate change based on the number of repetitions for the strong items. Even if participants have full knowledge that the strong items will not be tested, they must still assess memory evidence by matching the test probe to all items in the memory set, and the differentiation produced by the strong items still affects performance.[4] Like BCDMEM, REM predicts no substantial change in discriminability for weak items from the strong-2X to the strong-5X condition ($d' = .65$ for strong-2X, $d' = .67$ for strong-5X).

Summary of the simulation results

The simulation results validate the reasoning underlying our experiment design. BCDMEM predicts false alarm rate differences based on the degree of learning expected for targets on a test. REM predicts false alarm rate differences based on differentiation without changes in retrieval expectations. By dissociating the content of the study list

[3] The g parameter is useful for modeling word frequency effects. For example, Shiffrin and Steyvers (1997) constructed stimuli based on a low value of g (.325) for low frequency words and a high value of g (.45) for high-frequency words. The result is that the features that represent low frequency words tend to have rarer values than the features that represent high-frequency words. Shiffrin and Steyvers assume that the retrieval system does not try to estimate the frequency of the each word and tailor the retrieval calculations accordingly; instead, the system uses one long-run base-rate estimate for all words, which they set to .4 (this is what we call g_r). All of our stimuli were high-frequency words, so we used a high g value to generate stimuli (.475) and a more moderate g_r value at retrieval (.4).

[4] In our experiments, participants could not establish different match sets for the strong and weak items based on qualitative context differences because all items were mixed randomly in the study list. Furthermore, participants had no motivation for establishing different match sets, as they were never asked to discriminate strong from weak targets. In pilot work, we had participants complete cycles in which they were accurately told whether strong or weak targets would be tested (as in all of our experiments), and then they completed an ostensibly weak test that actually had a mixture of strong and weak targets. Participants recognized a higher proportion of strong (.77) than weak (.46) targets even though only weak targets were expected, which establishes that strong targets were still in the match set.

from the strength of targets expected at test, we will be able to discriminate differentiation from expectation-based mechanisms. With well defined predictions from the models, we now turn to our empirical results. Our first experiment matches the exact design used in the model simulations. We also report follow-up experiments to replicate our results when the number of study trials for strong items was manipulated between- vs. within-subjects (Experiment 2) and when performance feedback was not available (Experiment 3). A fourth experiment explored whether our procedures allowed participants to make within-test criterion shifts based on item strength (Verde & Rotello, 2007).

Experiment 1

Method

Participants

Students in undergraduate psychology courses participated to earn course credit. Twenty-three participants completed the pure test condition and 25 completed the mixed test condition.

Materials and procedure

The stimuli for each participant were randomly selected from a pool of 812 high-frequency words (78–10,595 occurrences per million; Kucera & Francis, 1967). Participants completed 12 study/test cycles, and no words were re-used across cycles. Each study list contained 24 words, half weak and half strong. Weak words were always presented once. Strong words were presented twice for half of the lists and five times for the remaining lists, and at least one item intervened before any word was repeated. The order of the strength conditions was randomized across study/test cycles. Study words remained on the screen for 800 ms with 150 ms of blank screen between words.

Immediately following the last word on every study list, participants began a digit memory task. For this task, participants saw a continuous random sequence of digits 1–9. Each digit remained on the screen for 900 ms followed by 100 ms of blank screen. Participants were asked to press a response key every time the current digit was the same as the digit that was two-items back in the sequence. Participants saw 63 digits following strong-2X lists and 46 digits following strong-5X lists. We varied the length of the distracter task to equate the average retention interval for weak items across the strength conditions. After the full sequence of digits, participants saw a feedback screen that reported the number of targets in the sequence and the number of times they successfully responded to the targets.

The memory test began immediately after the digit task feedback. The test instructions informed participants of how many times the targets on the upcoming test had been studied. Namely, they were told that the targets were studied once for pure-weak tests, twice pure-strong tests in the strong-2X condition, five times for pure-strong tests in the strong-5X conditions, once *or* twice for mixed tests in the strong-2X condition, and once or five times for mixed tests in the strong-5X condition. To verify that they attended to this information, participants had to type in the number of times the targets were studied to begin the test. Throughout each test, an ERROR message followed all incorrect responses.

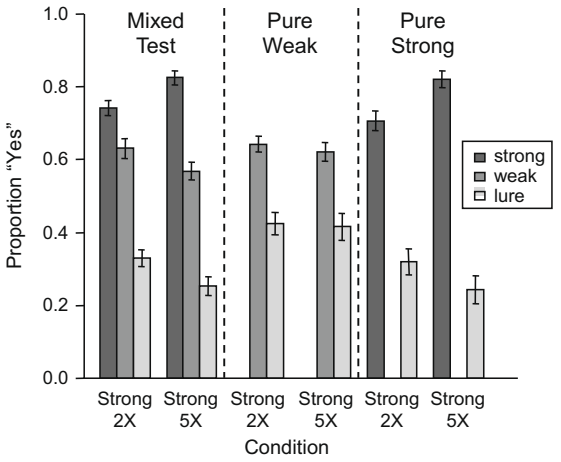

Fig. 3. Results from Experiment 1. The three panels represent the different test conditions, with mixed tests including both weak and strong targets and pure tests containing only weak or only strong. Within each test condition, the left set of bars gives results from the strong-2X condition and the right set of bars gives results for the strong-5X condition. The darkest bars show results for strong targets (studied either 2 or 5 times), the light gray bars show results for weak targets (always studied once), and the white bars show results for lures. The error bars show one standard error below and above each mean.

Test composition depended on condition. The pure tests consisted of either the 12 strong items or the 12 weak items from the study list with 12 lures. The mixed tests consisted of 6 strong targets, 6 weak targets, and 12 lures. The order of the test lists was random.

Results and discussion

Fig. 3 displays the proportion of "yes" responses across the conditions of Experiment 1. In the mixed-test group, increasing the number of presentations for strong items increased the hit rate for these items (.74 vs. 83), $t(24) = 3.44$, $p < .01$. The weak target hit rate decreased from the strong-2X condition (.63) to the strong-5X condition (.57), $t(24) = 2.42$, $p < .05$, as did the false alarm rate (.33 vs. .25), $t(24) = 3.93$, $p < .001$. Combining the weak target hit rates and false alarm rates into d' scores showed no evidence for a list strength effect. The difference in discriminability between strong-2X lists ($d' = .82$) and strong-5X lists ($d' = .90$) did not approach significance, $t(24) = .86$, $p = .40$. With pure-weak tests, the false alarm rate did not change when strong items were presented twice (.42) vs. five times (.42), $t(22) = 0.22$, *ns*. The slight decrease in the weak target hit rate (.64 vs. .62) did not approach significance, $t(22) = 1.10$, $p = .28$, nor did it create a significant difference in d' (.59 vs. .56), $t(22) = 0.24$, *ns*. On the pure-strong tests, the hit rate for strong targets was lower in the strong-2X condition (.71) than in the strong-5X

condition (.82), $t(22) = 4.19$, $p < .001$. Additional presentations for the strong items also significantly decreased the false alarm rate (.32 vs. .24), $t(22) = 2.66$, $p < .05$.

The results show a very clear pattern: increasing the number of presentations for the strong items led to a reduced false alarm rate only when the strong items appeared at test, i.e., in the pure-strong and mixed test conditions. This pattern is inconsistent with a differentiation-based explanation. Differentiation should reduce the false alarm rate any time words are encoded more effectively, regardless of whether or not the strengthened words appear on the test. The pattern is consistent with an expectation-based explanation and with the predictions of BCDMEM, which accommodates the reduced false alarm rate in terms of the degree of learning expected for the targets that will appear on the test.

Quantitative model fits

Our primary goal is to compare the qualitative predictions of the expectation-based and differentiation accounts; we are less concerned with the quantitative fit of specific models to our experiments. However, given that we did use specific models to derive qualitative predictions, it is important to establish that these predictions are not limited to the parameter values that we chose for our simulations. To achieve this, we performed fits of both models allowing a SIMPLEX search routine (Nelder & Meade, 1965) to freely adjust the parameters to minimize G^2 (Bishop, Fienberg, & Holland, 1975, chap. 14). We fit grouped data from the participants in the pure test condition, because the models make the same qualitative predictions for the mixed test condition and Experiments 2 and 3 only include the pure test condition.

We constrained parameters in the same way for both models. First, parameters that do not depend on the number of learning trials were constrained to be equal across all conditions. For BCDMEM, these parameters include the sparsity of the context vector (s), the proportion of pre-experimental links between item and context features (p), and the probability of omitting a feature of the study context in the reinstated context (d). For REM, these parameters include the frequency base rate of the studied items (g), the frequency base rate used in the retrieval calculation (g_r), and the probability of copying the correct value for stored features (c). Second, parameters governing changes in strength as a result of study repetition were constrained to be equal for pure-weak and pure-strong tests, which reflects the fact that test type was not signaled until after identical study phases. In BCDMEM, this means that the learning parameter for items presented once (r_1), twice (r_2), or five times (r_5) could not vary across test conditions. In REM, the same constraint was applied to the total probability of feature storage for items presented once (μ_1), twice (μ_2), or five times (μ_5). Third, both models were constrained to always select the most likely response; that is, the criterion was constrained to be 1. Overall, both models had six parameters to fit eight independent response frequencies (the frequencies of hits and false alarms in each of the four conditions created by crossing test type with the number of repetitions for strong items); therefore, each G^2 is associated with two degrees of freedom. Because the models have the same number of parameters and are subject to the same constraints, differences in fit should reflect differences in the veracity of the models and not differences in flexibility (although number of parameters is only a rough proxy for model flexibility).

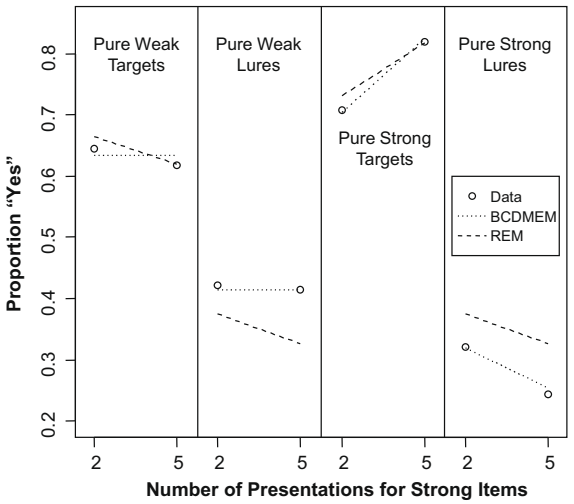

Fig. 4. Fitting results for BCDMEM and REM for the pure-test participants in Experiment 1. Each panel shows the change in performance produced by increasing the presentations for strong items from two to five. The points show the empirical results, the dotted lines show the results from the best-fitting BCDMEM model, and the dashed lines show the results from the best-fitting REM model.

The best fitting parameter values for both models are reported in Table 1, and fitting results are displayed in Fig. 4. The figure highlights each model's predictions for the difference between the strong-2X and strong-5X conditions across the item and test types. Each pair of points shows the data for the strong-2X and strong-5X conditions, the dotted lines show BCDMEM's fit, and the dashed lines show REM's fit. The fitting results clearly indicated a better fit for BCDMEM [$G^2(2) = 2.23$] than REM [$G^2(2) = 77.22$]. One problem for REM is that differentiation operates any time there are strong memory traces even if the strong items are not tested. Thus, REM must predict a differentiation-based decrease in the false alarm rate on both the pure-weak and pure-strong tests, whereas the data show a large decrease for pure-strong tests and no change on pure-weak tests. REM could only split the difference, producing a decrease that is too small for the pure-strong tests and too large for the pure-weak tests. The hit rate on the pure-weak tests shows a hint of the decrease predicted by REM, but the effect in the data was non-significant and smaller than the decrease in REM's fit. Without expectation-based mechanisms, REM also misses the overall differences in false alarm rate between the pure-weak and pure-strong tests.

The most important thing to note from the fitting results is that both models show the same qualitative predictions as in our simulations in the Introduction. On the pure-strong tests, both models show an increase in the hit rate and a decrease in the false alarm rate when strong items receive extra presentations. On the pure-weak tests, BCDMEM shows no difference from strong-2X to

strong-5X, and REM shows decreases in the hit rate and false alarm rate. This provides further validation for our use of these qualitative patterns as a general test of the expectation-based and differentiation explanations of the strength-based mirror effect.[5]

Experiment 2

The REM predictions shown in Fig. 2 are based on the assumption that participants only match test probes to the items in the last study list. In the experiments, participants were correctly informed that only items from the last list would appear on the test. However, they may not have succeeded in limiting retrieval assessments to these memory traces given the high degree of contextual overlap between lists. Differentiation effects would be diluted with multiple lists in the match set, because nominally weak cycles would be affected by strong memory traces from previous lists, and vice versa.[6] In Experiment 2, separate groups of participants studied strong-2X and strong-5X lists. All participants completed half of the study test cycles with pure-strong tests and the remaining cycles with pure-weak tests. Strong items should have more differentiated memory traces on every list seen by the strong-5X participants compared to every list seen by the strong-2X participants, so REM predicts false alarm rate differences for both the pure-strong and pure-weak tests irrespective of the number of previous lists in the match set.

Method

Participants

Seventy-one participants from the same pool as Experiment 1 contributed data, with 36 in the strong-2X condition and 35 in the strong-5X condition.

Materials and procedure

The methodological details matched those of Experiment 1, except that participants completed only the pure test conditions with the number of repetitions for strong items manipulated between participants.

Results and discussion

The results of Experiment 2 appear in Fig. 5. On the pure-weak tests, the false alarm rate did not differ between strong-2X and strong-5X participants (.37 vs. .37), $t(69) = .78$, *ns*, and the slight decrease in the hit rate (.65 vs. .63) did not approach significance, $t(69) = .78$, $p = .44$. As in Experiment 1, the d' scores for weak items showed no evidence of a list strength effect, as the difference between the strong-2X (.79) and strong-5X (.73) conditions did not approach significance, $t(69) = 0.64$, $p = .53$. On the pure-strong tests, the hit rate increased from strong-2X (.74) to strong-5X (.81), $t(69) = 3.39$, $p < .001$. Mirroring the hit rate, the false alarm rate significantly decreased (.31 vs. .24), $t(69) = 2.02$, $p < .05$. As in Experiment 1, increasing the number of learning trials for the strong items lowered the false alarm rate only when these strong items appeared at test. The pure-weak tests showed no evidence of differentiation.

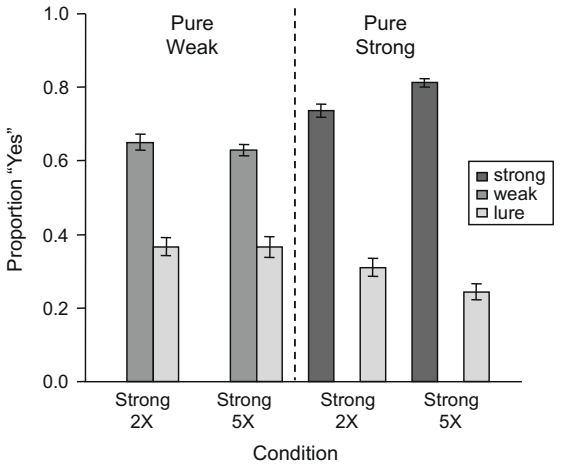

Fig. 5. Results from Experiment 2, with separate groups of participants in the strong-2X and strong-5X conditions. The left panel shows results for the pure-weak tests and the right shows results from the pure-strong tests. Within each test condition, the left set of bars gives results from the strong-2X participants and the right set of bars gives results for the strong-5X participants. The darkest bars show results for strong targets (studied either 2 or 5 times), the light gray bars show results for weak targets (always studied once), and the white bars show results for lures. The error bars show one standard error below and above each mean.

Experiment 3

Along with previous experiments (Marsh et al., 2009; McCabe & Balota, 2007), Experiments 1 and 2 strongly support the contention that retrieval expectations influence the false alarm rate. After studying a mixed list of words studied either once or five times, participants made many more false alarms when they were told that only the weak items would be tested than when they were told that only the strong items would be tested. Differentiation can play no role in this difference, because the study lists were identical. Our design also included a pair of conditions in which false alarm rates could be affected by differentiation with test expectations held constant; that is, the strong-2X and strong-5X conditions with pure-weak tests. We found no evidence for the change in false alarm rates predicted by differentiation. Clearly, a model that proposes that item strength affects false alarm rates only via a differentiation mechanism is not viable.

In Experiments 3 and 4, we explore the possibility that differentiation may work in conjunction with another retrieval-based mechanism. In REM, the most direct way to implement a retrieval-based strategy would be to allow

[5] For completeness, we also evaluated both models' ability to fit the combined data from the mixed-test and pure-test participants. This fit required no new parameter values for either model, so both still have 6 free parameters. The combined dataset contains 14 independent response frequencies, resulting in 8 degrees of freedom for the G^2 statistics. Results were the same as the pure test fits reported in the paper in all important respects. BCDMEM [$G^2(8) = 37.57$] fit the data much more closely than REM [$G^2(8) = 111.99$], and both models produced the same qualitative patterns that we reported based on the simulations in the Introduction.

[6] We thank Ken Malmberg for this suggestion.

the criterion to vary across test conditions. One possibility is to implement an expectation-based criterion setting strategy in REM, creating a hybrid model. For example, the model could use a more liberal criterion on pure-weak tests than on pure-strong tests, which would accommodate the higher false alarm rate for the pure-weak tests. However, a differentiation-and-expectation hybrid would still miss key aspects of the results. Most notably, differentiation should still influence performance when retrieval expectations are the same. Therefore, a hybrid model would still incorrectly predict that the number of learning trials for strong items would affect the false alarm rate on the pure-weak tests.

Instead of adjusting REM's criterion based on expected memory strength, we considered a less constrained approach in which the criterion could take on different values even on tests with the same content and same instructions. Such shifts in criterion could mask differentiation effects in the pure-weak tests. Specifically, if participants set a more liberal criterion in the strong-5X than the strong-2X condition, the resulting increase in the strong-5X false alarm rate may have canceled out the decrease produced by differentiation. For example, participants in the strong-5X condition may have responded "studied" for any odds ratio over .8, meaning that they would have claimed to remember some items that were actually more likely to be lures than targets. This explanation seems problematic because it requires participants to become more liberal following more effective learning of the study items. Furthermore, REM does not need to posit a criterion shift to correctly predict results from the mixed and pure-strong tests. It seems unlikely that participants would shift their criterion based on the degree of learning for the strong words *only* when those words do *not* appear on the test. Although such shifts seem implausible, in our last two experiments we explored the possibility that these shifts explain the lack of evidence for differentiation in our data.

In Experiment 3, we replicate our design both with and without accuracy feedback to explore the possibility that feedback induces differentiation–masking criterion shifts. On the pure-weak tests, if the amount of evidence for both weak targets and lures decreased in the strong-5X condition, then participants would see more ERROR messages after "no" responses and fewer ERROR messages after "yes" responses compared to the strong-2X condition. This differential feedback could prompt participants to become more liberal, thus canceling out differentiation effects. If this is the case, then we should see evidence for differentiation in the pure-weak tests when no feedback is provided.

Method

Participants

Fifty-five participants from the same pool as the previous experiments contributed data, with 28 in the feedback condition and 27 in the no-feedback condition.

Materials and procedure

The methodological details matched those of Experiment 1, except that participants completed only the pure test conditions and separate groups of participants received either feedback or no feedback.

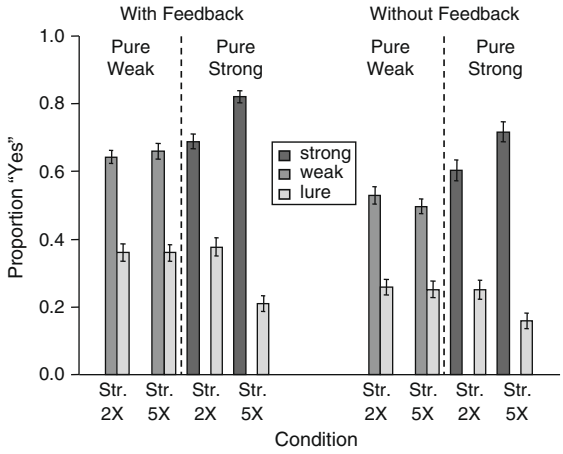

Fig. 6. Results from Experiment 3 both with (left side of plot) and without (right side of plot) accuracy feedback. For each feedback group, the left panel shows results for the pure-weak tests and the right shows results from the pure-strong tests. The strong-2X and strong 5X conditions are abbreviated Str. 2X and Str. 5X, respectively. The darkest bars show results for strong targets (studied either 2 or 5 times), the light gray bars show results for weak targets (always studied once), and the white bars show results for lures. The error bars show one standard error below and above each mean.

Results and discussion

The results of Experiment 3 appear in Fig. 6. On the pure-strong tests, the hit rate increased from the strong-2X to the strong-5X condition both with feedback (.69 vs. .82), $t(27) = 5.09$, $p < .001$, and without feedback (.60 vs. .72), $t(26) = 3.97$, $p < .001$. Moreover, these hit rate increases were mirrored by false alarm rate decreases both with feedback (.38 vs. .21), $t(27) = 7.73$, $p < .001$, and without (.25 vs. .16), $t(26) = 5.37$, $p < .001$. None of these differences were evident on the pure-weak tests. The weak hit rate showed a small, non-significant increase from strong-2X to strong-5X with feedback (.64 vs. .66), $t(27) = .83$, *ns*, and a small, non-significant decrease without feedback (.53 vs. .50), $t(26) = 1.40$, *ns*. Most importantly, false alarm rates showed no differences both with feedback (.36 vs. .36), $t(27) = .05$, *ns,* and without (.26 vs. .25), $t(26) = .35$, *ns*.

In the no-feedback condition, participants were overly conservative, i.e., they made many more errors by missing targets (.21) than by false alarming to lures (.12). For comparison, the feedback participants made a similar proportion of misses (.15) and false alarms (.16). In either REM or BCDMEM, this global shift in conservativeness can be accommodated by changing the response criterion on the likelihood dimension. In our simulations and our fits to the first experiment, we set the criterion to its optimal value in both models (i.e., both models always produced the response that was most likely to be correct given the evidence from memory). BCDMEM's successful fit in the first experiment suggests that using the optimal

criterion value can be reasonable when feedback is provided, but Experiment 3 suggests that an optimal criterion may not always work for conditions without feedback. This conclusion is reinforced by prior research demonstrating that feedback can help participants approach optimal criteria settings (e.g., Maddox & Bohil, 2005; Starns & Ratcliff, in press). Critically, a global criterion shift does not change either model's predictions for the patterns within the feedback and no-feedback conditions.

The pattern of results from the previous experiments replicated both with and without feedback. False alarm rates did not change from the strong-2X to the strong-5X condition on pure-weak tests, matching the predictions of the expectation-based account but not the differentiation account. The results do not support the hypothesis that the use of feedback prompts a specific criterion shift that masks the influence of differentiation, although feedback did induce a global criterion shift across all conditions.

As in the previous experiments, results showed no evidence for a list strength effect. Feedback participants showed a small, non-significant increase in weak d' from the strong-2X (.76) to the strong-5X (.82) condition, $t(27) = 0.64$, $p = .53$. No-feedback participants showed a small, non-significant decrease in d' (.77 vs. .74), $t(26) = 0.27$, $p = .79$.

Experiment 4

Experiment 4 also addressed the possibility of a criterion shift on the pure-weak tests. Again, to offset the effects of differentiation, participants would have to adopt a different criterion between tests with the same content and the same instructions. They may have made such shifts in the previous experiments because they noticed the difference in memory strength that was produced by differentiation. Experiment 4 was designed to determine whether our design allows participants to adjust their retrieval criterion based on memory strength independent of the test instructions. We used only mixed tests, with some tests beginning with a block of all strong targets followed by a block of all weak targets, some tests in the reverse order, and some with strong and weak targets dispersed evenly throughout the test (see Verde & Rotello, 2007, for a similar manipulation). Participants were not informed of the blocked structure of the lists. As in previous experiments, they were informed of the overall test composition before each list, and the information provided to them was always accurate.

If participants use a constant criterion based on the overall test composition reported in the test instructions, then false alarm rates should remain constant across the test blocks regardless of the strength of the items within the block. If participants make criterion adjustments even when they are given identical test instructions, then they should become more conservative on strong blocks and more liberal on weak blocks compared to mixed blocks. As in Experiment 3, we ran separate groups of participants with and without accuracy feedback to explore this variable's potential role in facilitating within-test criterion shifts.

Method

Participants

Fifty-three participants from the same pool as the previous experiments contributed data, with 27 in the feedback condition and 26 in the no-feedback condition.

Materials and procedure

The methodological details matched the first experiment with the following exceptions. Participants completed only the mixed test condition, and each test contained all of the studied words (12 strong and 12 weak) along with 24 lures. Test order was not completely random; instead, the tests were structured to create blocks of items organized by strength. Strong blocks contained 12 strong targets and 12 lures; weak blocks contained 12 weak targets and 12 lures; and mixed blocks contained 6 strong targets, 6 weak targets, and 12 lures. Each participant completed four tests with a strong block followed by a weak block, four tests in the reverse order, and four tests with two mixed blocks. The order of these test types was randomized across study/test cycles.

Results and discussion

The full results are reported in Table 2, and the false alarm rate results averaged across the feedback variable are displayed in Fig. 7 (feedback did not interact with any other variable). False alarm rates remained relatively constant regardless of whether lures were in a test section with strong targets (.32), weak targets (.33), or a mixture of the two (.32), $F(2, 102) = .52$, *ns*, $MSE = .011$. In the first half of the tests, the false alarm rate was slightly higher when weak targets were tested, but the difference was well within the range of error and did not appear in the second half of the tests. Replicating the mixed test results from Experiment 1, false alarm rates were higher when strong items were studied two times (.36) vs. five times (.29), $F(1, 51) = 41.59$, $p < .001$, $MSE = .018$. Thus, participants did lower their false alarm rate when the test instructions informed them that stronger items would be tested, but they did not respond to unsignaled changes in target strength within the tests. Participants made more false alarms on the second half of the test (.36) than on the first half (.29), $F(1, 51) = 52.84$, $p < .001$, $MSE = .014$, and they were more conservative overall without feedback (.28) than with feedback (.37), $F(1, 51) = 5.63$, $p < .05$, $MSE = .266$. As in Experiment 3, feedback prompted an overall criterion shift but did not induce specific criterion shifts within conditions. None of the interactions were significant.

We also evaluated the hit rates for evidence that participants were adjusting the response criterion across strength blocks. To avoid cells with missing data, we ran separate ANOVAs on the weak and strong target hit rates. If participants made strength-based shifts, then the weak hit rate should increase from mixed to weak blocks and the strong hit rate should decrease from mixed to strong

Table 2
Recognition data (proportion of "yes" responses) from Experiment 4.

List type and block order	Test half and item type					
	Half 1			Half 2		
	Weak target	Strong target	Lure	Weak target	Strong target	Lure
			Feedback participants			
Strong-2X						
Mixed–mixed	0.67 (0.03)	0.77 (0.03)	0.34 (0.02)	0.68 (0.03)	0.80 (0.03)	0.44 (0.03)
Weak–strong	0.68 (0.02)	–	0.41 (0.03)	–	0.78 (0.03)	0.47 (0.03)
Strong–weak	–	0.75 (0.02)	0.37 (0.04)	0.70 (0.02)	–	0.46 (0.03)
Strong-5X						
Mixed–mixed	0.61 (0.03)	0.86 (0.03)	0.27 (0.03)	0.66 (0.03)	0.89 (0.02)	0.38 (0.03)
Weak–strong	0.63 (0.03)	–	0.31 (0.03)	–	0.86 (0.02)	0.38 (0.03)
Strong–weak	–	0.84 (0.02)	0.29 (0.03)	0.63 (0.03)	–	0.38 (0.03)
			No-feedback participants			
Strong-2X						
Mixed–mixed	0.51 (0.04)	0.64 (0.04)	0.29 (0.04)	0.56 (0.05)	0.63 (0.03)	0.33 (0.04)
Weak–strong	0.51 (0.03)	–	0.28 (0.04)	–	0.58 (0.04)	0.33 (0.04)
Strong–weak	–	0.58 (0.03)	0.27 (0.04)	0.47 (0.03)	–	0.33 (0.04)
Strong-5X						
Mixed–mixed	0.50 (0.04)	0.70 (0.04)	0.23 (0.04)	0.44 (0.05)	0.69 (0.03)	0.29 (0.04)
Weak–strong	0.49 (0.03)	–	0.24 (0.03)	–	0.74 (0.03)	0.28 (0.04)
Strong–weak	–	0.77 (0.03)	0.21 (0.03)	0.51 (0.03)	–	0.25 (0.04)

Note: "Weak–Strong" indicates a test with weak targets on the first half and strong targets on the second half, etc.; standard errors are in parentheses.

blocks. The strong target ANOVA showed no difference between mixed (.75) and strong (.74) blocks, $F(1, 51) = 0.55$, *ns*, $MSE = .018$. Participants were more conservative without feedback (.67) than with feedback (.82), $F(1, 51) = 28.54$, $p < .001$, $MSE = .087$, and hit rates were higher when the strong items were studied five times (.79) vs. twice (.69), $F(1, 51) = 82.12$, $p < .001$, $MSE = .013$. There was no change from the first (.74) to the second (.74) half of the tests, $F(1, 51) = 0.16$, *ns*, $MSE = .017$. The weak hit rate ANOVA showed no difference between mixed (.58) and weak (.58) blocks, $F(1, 51) = 0.04$, *ns*, $MSE = .024$. The weak hit rate did drop from strong-2X (.60) to strong-5X (.56) lists, $F(1, 51) = 8.16$, $p < .01$, $MSE = .018$, which again shows that participants changed their expectations based on the overall test content reported in the instructions. Again, responding was more conservative without feedback (.50) than with feedback (.66), $F(1, 51) = 21.76$, $p < .001$, $MSE = .124$, and hit rates did not change from the first (.58) to the second (.58) half of the tests, $F(1, 51) = 0.21$, *ns*, $MSE = .016$.

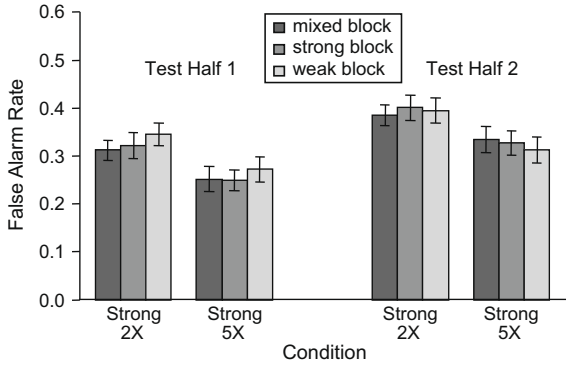

Fig. 7. False alarm rate results from Experiment 4 on the first half (left side of plot) and second half (right side of plot) of the tests. The bars indicate the type of target on the given test half. The mixed-block bar shows the false alarm rate for test halves with both strong and weak targets, the strong-block bar shows results for test halves with only strong targets, and the weak-block bar shows results for test halves with only weak targets. The strong block results from half 1 are from the same tests as the weak block results from half 2, and vice versa. Within each test half, the left group of bars is from the strong-2X condition and the right group of bars is from the strong-5X condition.

The strong hit rate ANOVA showed a 3-way interaction between the type of list (strong-2X vs. strong-5X), feedback, and the strength of the test block, $F(1, 51) = 5.13$, $p < .05$, $MSE = .019$, and the weak hit rate ANOVA showed a 4-way interaction also involving test half, $F(1, 51) = 5.55$, $p < .05$, $MSE = .015$. To ensure that these interactions did not mitigate our conclusion of no criterion shifts across test blocks, we performed *t*-tests on every comparison between mixed-strength and pure-strength blocks across the other variables. Even with uncorrected *p*-values, only one of these 16 tests reached significance, and the difference was in the opposite direction predicted by strength-based shifts [the weak hit rate dropped from mixed (.56) to weak (.47) blocks on the second half of the tests in the strong-2X condition for no-feedback participants]. Overall, neither the hit rate nor false alarm rate data show any evidence of shifts across strength blocks.

Discriminability decreased from the first to the second half of the test. Averaged across all other variables, d' fell from .85 to .67 for items presented once, 1.11 to .92 for items presented twice, and 1.73 to 1.45 for items presented five times. False alarm rates were also higher on the second half of the test, with almost no change in hit rates. The performance drop may reflect the longer retention interval for later items or interference from previous test items (Ratcliff & Hockley, 1980; Ratcliff & Murdock, 1976). Neither the expectation-based nor the differentiation account makes a specific prediction for the test-half variable, but

the increase in false alarm rate could be reconciled with an expectation-based approach. Participants may have been aware of the performance drop across the test, and as a result, they may have dynamically shifted their expectations regarding memory strength. If this were true, it would indicate that participants can make certain types of dynamic adjustments to retrieval expectations (i.e., adjusting to retention interval throughout a test) while failing to make other types (i.e., adjusting to the strength of items within a block). This account is speculative, and targeted empirical work will be needed to uncover the true basis of the test half effects. On the theoretical side, BCDMEM currently has no mechanism to moderate r_{EST} based on the retention interval. Our results suggest that this may be an important future development for the model.

Experiment 4 suggests that a freely varying response criterion cannot save REM from its mispredictions on the pure-weak tests. When we dramatically manipulated the strength of items across test blocks following identical instructions, participants used a consistent criterion. If participants do not shift their criterion based on getting entirely strong or entirely weak targets to begin a test, one cannot posit that they shift their criterion based on a differentiation-induced decrease in memory strength. Although participants did not make criterion shifts across strength blocks in our experiment, we do not claim that participants will always fail to make unsignaled criterion shifts. In the General Discussion, we review research suggesting that dynamic criterion shifts are sensitive to experimental procedures, and we discuss ways in which our procedures may have hindered unsignaled criterion adjustments. Critically, these procedural details also characterize our first three experiments. Thus, Experiment 4 suggests that the pure test results from Experiments 1–3 reflect a lack of differentiation, not differentiation offset by a criterion shift.

General discussion

We sought to contrast two explanations for the strength-based mirror effect, a pattern in which false alarm rates decrease when studied items are strengthened. The differentiation account holds that strong memory traces directly reduce the memory evidence for lures, and the expectation-based account holds that false alarm rates decrease when participants expect stronger targets at test. Our experiments manipulated the number of presentations for repeated (strong) items that were mixed with singly-presented (weak) items at encoding. Additional learning for strong items led to false alarm rate decreases only when participants knew that the strong items would be tested, with no corresponding effect when participants knew that only weak items would be tested. Thus, our results indicate that strength-based changes in false alarm rate are produced via changes in participants' expectations at test. If differentiation were at work, false alarm rates should have been based on the content of the study list; that is, the additional learning in the strong-5X condition should have lowered the false alarm rate even when the strong items were not tested.

In the process of testing the differentiation and expectation-based explanations, we explored the predictions of two computational models. Comparing the predictions from the simulations to the empirical results revealed convincing support for BCDMEM over the REM model. The critical result distinguishing the models was performance on the pure-weak tests. In all of our experiments, the weak target hit rate and the false alarm rate did not change from the strong-2X to the strong-5X condition, as predicted by BCDMEM. Although we capitalized on each models' ability to define specific predictions for our experiments, we do not wish to limit our conclusions to these particular models. REM is only one of a class of differentiation models (McClelland & Chappell, 1998; Shiffrin et al., 1990), each of which predicts that strengthening encoded traces should reduce the false alarm rate even if the strengthened items are not tested. Similarly, a criterion-shift account (Hirshman, 1995; Ratcliff et al., 1990; Stretch & Wixted, 1998) is as consistent with the expectation-based explanation as BCDMEM. We focused on BCDMEM because it provides a complete specification of memory and decision processes, just like REM. To match this standard, the criterion-shift explanation would need a "background" model specifying the mechanisms involved in setting and adjusting the response criterion. Several researchers have already taken steps toward this goal (Hirshman, 1995; Maddox, 2002; Treisman & Williams, 1984), and it remains an important direction for future research.

Some may not wish to grant that the BCDMEM model predicted our results *a priori*, and we acknowledge that the model can accommodate many alternative patterns of data if the r_{EST} parameter is allowed to vary freely across conditions. While this is a reasonable concern for complex models like BCDMEM, we believe that the flexibility of such models can be constrained through experimental design. We took great care to manipulate subjects' expectations by providing useful, accurate information about the content of the upcoming test. We also verified that they had attended to the instructions by having them type the number of learning trials for the targets on the upcoming test list. Therefore, our procedures ensured that participants' expectations were appropriate for the test content, and we constrained BCDMEM's r_{EST} values to match the actual r value for the tested items. Under this constraint, BCDMEM does make a clear prediction that was verified in our empirical work. Moreover, we note that the r_{EST} parameter can *only* influence responding at retrieval. Therefore, BCDMEM correctly predicts that the strength-based mirror effect can be produced with a purely test-based manipulation.

Research on differentiation and the strength-based mirror effect

In a standard strength manipulation, the actual learning driving the differentiation process and participants' expectations at test are confounded. A participant who studies only strong items will naturally expect to be tested on strong items, and vice versa. We deliberately designed our experiments to include comparisons in which changing expectations could create false alarm rate differences but

differentiation could not (e.g., the strong-5X condition with pure-weak vs. pure-strong tests) as well as conditions in which differentiation could create false alarm rate differences but changing expectations could not (e.g., the strong-2X vs. strong-5X conditions with pure-weak tests). We observed expectation effects that could not be explained by differentiation, but not the reverse. In this section, we review the previous literature on differentiation and the strength-based mirror effect with special attention to this question: do any results show an influence of differentiation in conditions that held test expectations constant?

Criss (2006) examined differentiation by comparing pure and mixed-strength lists with tests including lures that rhymed with the target words. When targets were strengthened via additional presentations, the false alarm rate for rhyming lures slightly increased in a mixed list design, but decreased in a pure list design. Criss shows how REM predicts this result based on differentiation. However, retrieval expectations also would have changed from tests with pure weak, pure strong, and mixed targets. BCDMEM cannot accommodate the results of this experiment, because it lacks a mechanism for representing related lures. However, an expectation-based criterion shift can explain the result in a signal detection model with separate distributions for targets, related lures, and unrelated lures (as Criss acknowledges in the General discussion). Repeating the targets associated with a related lure increases the memory evidence for the targets themselves and, to a lesser degree, for their related lures (Starns, Hicks, & Marsh, 2006). This results in a higher related-lure false alarm rate when participants must use the same criterion value for repeated and non-repeated items (i.e., on mixed tests). On pure tests, participants use a higher criterion when tests contain only repeated targets than when tests contain only weak targets. This criterion shift overcompensates for the boost in memory evidence for rhyming lures, leading to a lower rhyming lure false alarm rate. Thus, the result does not show an influence of differentiation independent of changes in retrieval expectations.

In a third experiment with rhyming lures, Criss (2006) evaluated memory for words that were presented in a sentence context at encoding. She compared pure-weak lists to two types of pure-strong lists, one in which words were presented three times in the same sentence and one in which words were presented three times in three different sentences. The purpose of changing the sentence contexts was to disrupt differentiation by establishing separate memory traces for each repetition of the same word (as opposed to storing all repetitions in the same trace, REM's default). Results showed that the rhyming lure false alarm rate significantly decreased from weak lists to strong lists in the same-sentence condition. In contrast, results showed only a slight, non-significant decrease in the different-sentence condition. Results from the different-sentence condition fell in between REM's predictions with complete differentiation and REM's predictions with no differentiation, which is cited as support for the role of differentiation in reducing the false alarm rate in the same-sentence condition.

Criss' (2006) Experiment 3 result is impressive, especially because the manipulation was directly inspired by the differentiation mechanism. However, nothing rules out the possibility that participants expected to have better memory for targets in the same-sentence condition than in the different-sentence condition. Indeed, such an expectation would not be unreasonable given that discriminability was higher in the same-sentence condition. These conditions were tested on separate lists, providing an opportunity for list-wide criterion adjustments based on expected strength. The differentiation account seems more elegant in this circumstance; nevertheless, the manipulation does not completely disentangle differentiation from competing hypotheses.

Criss (2009) attempted to discriminate differentiation from an expectation-based criterion shift in a study using 20-point confidence ratings. Criss showed that participants rejected lure items with greater confidence following pure-strong than pure-weak lists. In a separate experiment, Criss manipulated the proportion of targets on the test and showed less evidence for a change in confidence ratings for lures. Specifically, by a Kolmogorov–Smirnov test on the distributions of confidence ratings, 62% of participants showed evidence that the lure distribution changed in response to the strength manipulation compared to 38% for the target proportion manipulation. The target proportion results are cited as evidence that 20-point ratings are not subject to decision biases, thereby eliminating a criterion shift as an explanation for the strength-based changes in lure ratings.

Criss (2009) presents a compelling case for differentiation, but test expectations certainly could have changed from the pure-strong to pure-weak tests. Demonstrating that the 20-point scale did not show differences with a bias (target proportion) manipulation militates against an expectation-based criterion-shift account. However, previous experiments demonstrate that bias manipulations do affect confidence judgments on a 6-point scale (Mueller & Weidemann, 2008; Van Zandt, 2000), so one may not want to dismiss the possibility of bias effects on a 20-point scale based on a null result from a single experiment.[7] We leave this issue for future work, but we do wish to note that the rating results create no difficulty for BCDMEM's implementation of the expectation-based account. In BCDMEM, actual memory evidence for lures changes based on expected memory strength via the r_{EST} parameter, so this model predicts differences in lure ratings even if one is willing to assume that 20-point rating scales are immune to bias. The target proportion manipulation may have affected another model component such as the criterion or the prior probability ratio; indeed, given that manipulating target proportion changes the actual prior, changing the prior in the model would be by far the most direct way to accommodate this variable. Therefore, BCDMEM does not necessarily predict

[7] The bias-related changes in Van Zandt (2000) and Mueller and Weidemann (2008) involved not only shifts in the rating scale proportions, but also changes in z-ROC slope. Signal-detection theory without variability in criteria cannot produce the slope changes, but Van Zandt accommodated them with changing criteria in a response time model and Mueller and Weidemann accommodated them by changing both the position and variability of criteria in a signal detection model.

that the target proportion manipulation must show the same results as the strength manipulation.

In another investigation of differentiation, Criss (2010) shows that the strength-based mirror effect is best implemented in Ratcliff's (1978) diffusion model by allowing the drift rate for both targets and lures to change from pure-weak to pure-strong lists, as opposed to allowing changes only in the starting point of the diffusion process. This pattern is consistent with differentiation, but it is also easily accommodated by an expectation-based shift in the drift criterion (Ratcliff, 1985; Ratcliff & McKoon, 2008), which is the most direct way to implement an expectation-based criterion shift in a diffusion model analysis (again, Criss recognizes this possibility). Furthermore, a change in drift rates for lures would also be predicted by BCDMEM's expectation-based mechanism, because the actual memory evidence for lures changes when a higher learning estimate is used in the retrieval calculations. Thus, fitting the diffusion model to a standard list strength experiment does not help to discriminate the competing hypotheses.

Perhaps the best conclusion to draw from the research reviewed in this section is that several empirical results provide support for differentiation's role in the strength-based mirror effect, but none of them completely rule out a role for expectation-based mechanisms. Similarly, most prior studies advocating expectation-based mechanisms do not rule out a role for differentiation (e.g., Hirshman, 1995). When we carefully held test expectations constant and manipulated the content of the study list, we found no evidence for differentiation. Furthermore, our results and previous studies (Marsh et al., 2009; McCabe & Balota, 2007) show that retrieval expectations can affect false alarm rates even when the level of differentiation is held constant. In light of this strong evidence in favor of the expectation-based account, studies that wish to garner support for differentiation must be careful to hold retrieval expectations constant.

Comparisons to previous research

Several previous studies have explored how false alarm rates vary based on participants' expectations at test (e.g., Hirshman, 1995; McCabe & Balota, 2007; Stretch & Wixted, 1998; Verde & Rotello, 2007). These false alarm rate differences are usually explained in terms of a criterion shift in signal detection theory, but in BCDMEM they can be implemented as a change in the r_{EST} parameter. These explanations are similar, and in our review of this literature we will simply refer to expectation effects without committing to a criterion shift or a change in the r_{EST} parameter as the underlying mechanism.

Hirshman (1995) reported an experiment similar to our own in that strength was independently manipulated at study with unchanged test content. Participants studied either a pure-weak list or a mixed list of half strong and half weak items. In both conditions, the test contained only weak items. Results showed a reduced false alarm rate following the mixed list, even though the strong items never appeared at test. Critically, participants were not informed of the test content, so they may have expected to see strengthened items at test. Indeed, in an experiment similar to Hirshman's mixed-strength condition, McCabe and Balota (2007) showed higher false alarm rates when participants were informed that only weak items would be tested than when participants were given no special test instructions.[8] In another experiment similar to our own, Marsh et al. (2009); Experiment 3) had participants study a mixed list of repeated and non-repeated items. They had separate tests for strong and weak targets with instructions to make participants aware of the target strength on the upcoming test. Participants made more false alarms on weak tests than on strong tests. Together with our own findings, these studies support the expectation-based account and suggest that test instructions can play an important role in defining expectations.

Whether or not test expectations change when target strength is manipulated within a single test has proven to be a complex issue. Participants often show no differences in false alarm rate for strong and weak items when strength is demarcated on an item-by-item basis within a test (Morrell, Gaitan, & Wixted, 2002; Singer & Wixted, 2006, Experiments 1 and 2; Stretch & Wixted, 1998). However, item-by-item differences have been observed under some conditions (Hockley & Niewiadomsky, 2007; Singer, 2009; Singer & Wixted, 2006, Experiments 3 and 4).

Participants also show a limited ability to dynamically adjust retrieval expectations when entire blocks of strong and weak targets alternate but are not demarcated with instructions. Verde and Rotello (2007) had participants study mixed lists of strong and weak words, and then take a test in which all of the weak targets were followed by all of the strong targets, or vice versa. Results from four experiments without accuracy feedback showed no differences in false alarm rate for unmarked strong and weak blocks within a single test. Their Experiment 5 did show a within test shift when feedback was provided. Although participants showed limited flexibility in adjusting their expectations within a test, they did adjust their expectations based on the first test items they encountered. That is, they made more false alarms overall when tests started with the weaker items than when tests started with stronger items. The results of our Experiment 4 suggest that our participants were even less sensitive to unsignaled strength changes than those of Verde and Rotello. That is, our participants did not show different false alarm rates based on the strength of the first test block, and they showed no strength-based differences within a single test even when feedback was provided.

The discrepancies between our results and those of Verde and Rotello (2007) most likely reflect differences in the number of trials available for participants to make dynamic adjustments to their retrieval expectations. In Verde & Rotello, the separate strength blocks contained 80 items (40 targets and 40 lures). In our Experiment 4, the strength blocks contained only 24 items (12 targets and 12 lures). We chose this length because Experiment 4 was designed to test the possibility of unsignaled criterion adjustments

[8] McCabe and Balota (2007) proposed an expectation-based account that is similar to our own, but they focus on "remember" responses in the remember/know paradigm and do not base their predictions on a specific computational model like BCDMEM.

in Experiments 1–3, in which the tests were 24 items long. Just as participants are often unwilling or unable to quickly change retrieval expectations when strength varies across individual test trials (e.g., Stretch & Wixted, 1998), they may fail to make adjustments for relatively small blocks of trials. The exact number of trials that participants need to adjust their retrieval expectations is likely to differ based on procedural variables. For example, explicit models for the timing of criterion shifts in a signal detection framework suggest that participants require around 14 trials to make criterion adjustments in a lexical decision task (Brown & Steyvers, 2005) compared to only three trials in a picture-recognition task (Brown, Steyvers, & Hemmer, 2007). In a simple decision task, Ratcliff, Van Zandt, and McKoon (1999) also found evidence of criterion shifts within the first few trials of blocks containing either 20% or 80% targets.

Our results may also differ from those of Verde and Rotello (2007) because our test instructions alerted participants to changes in target strength between tests. This salient cue to memory strength may have overshadowed within-test influences on participants' expectations. Our instructions may have been particularly influential because they always reported accurate information.

Based on the current literature, we can make the following conclusions about how participants define their expectations for a test: participants probably base their expectations on the study list in normal situations. If the study list contains items of mixed-strength, participants will expect a mixed-strength test (Hirshman, 1995). When strength changes between study and test, results depend on whether or not this strength difference is highlighted with instructions (McCabe & Balota, 2007). When given instructions, participants can match their expectations to the test content (the current experiments; Marsh et al., 2009; McCabe & Balota, 2007). Participants are less likely to change expectations without relevant instructions, and the extent to which they adjust to unsignaled strength changes seems highly dependent on experimental procedures (our Experiment 4, Verde & Rotello, 2007). Results outside of the strength-effect literature reinforce the conclusion that participants can be more strongly influenced by test instructions than by actual experience with the test. For example, Starns, Hicks, Brown, and Martin (2008) found higher false alarm rates when instructions reported that 75% vs. 25% of the test words would be targets even though the tests were always 50% targets.

List strength effect

Our results replicate the null list strength effect (Ratcliff et al., 1990). This is most easily seen in the pure-weak tests, where the number of repetitions for the strong items affected neither the weak item hit rate nor the false alarm rate. The null list strength effect is problematic for the "first wave" of global matching models (Shiffrin et al., 1990), but several current models, including REM, accommodate this effect with a differentiation mechanism (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). However, BCDMEM also correctly predicts this phenomenon without a differentiation mechanism (Dennis & Humphreys, 2001). BCDMEM predicts null list strength effects because items do not interfere with one another; instead, each item is independently linked to context features. Recent evidence supports a lack of item interference in recognition memory for unrelated words by demonstrating null list length effects in carefully controlled experiments (Dennis, Lee, & Kinnell, 2008). Notably, the null list length effects are problematic for differentiation models, which predict decrements in performance for longer lists (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).

Although the lack of item interference in BCDMEM allows this model to correctly predict null list strength effects and perhaps correctly predict null list length effects for unrelated words, it leaves the model without a mechanism for producing high false alarm rates for lures that are related to the studied material. In most matching models, these effects are produced by introducing overlap in the features representing related items. BCDMEM's local item codes preclude such a strategy. The inability of BCDMEM to accommodate related lures is a major limitation of the model, and future work must explore whether this phenomenon can be accommodated with mechanisms other than feature overlap.

Final notes on criterion shifts

In this section, we focus on the expectation-based mechanism that is usually contrasted with differentiation: a standard signal detection model in which the criterion value but not the lure evidence changes based on strength (Criss, 2006, 2009, 2010). Critically, our work shows that REM would also need a free response criterion parameter in the pure-weak tests to accommodate the null effect from the strong-2X to the strong-5X condition; specifically, the response criterion would have to be more liberal in the strong-5X condition to offset the effects of differentiation. Experiments 3 and 4 suggest that participants do not make such a criterion shift. Nevertheless, some may find it compelling that REM can accommodate our results with a free response criterion. For this reason, we wish to clarify some aspects of testing REM with a criterion shift against the pure criterion-shift account.

If the criterion is treated as a completely free parameter, models are constrained only by the discriminability of target and lure evidence. Any pattern of hit and false alarm rates can be accommodated; thus, no evidence can be marshaled for or against the notion that differentiation plays a role in the strength-based mirror effect. Of course, differentiation models can be empirically tested given a fixed criterion; indeed, we found clear evidence against differentiation under this constraint.

Differentiation is also discriminable from a pure criterion shift in situations in which the criterion is neither completely fixed nor completely free, but can shift in ways constrained by principled considerations concerning when and how participants change their retrieval expectations (Brown et al., 1977). On these grounds, our data also support a pure criterion-shift explanation over a differentiation-plus-criterion-shift explanation. If lure evidence does not change, our results are consistent with a simple

principle for criterion placement: participants set the criterion based on the strength of targets they expect at test, becoming more conservative when they are tested on stronger items (see Hirshman, 1995, for a detailed discussion of strength-based criterion setting mechanisms). If lure evidence changes as prescribed by differentiation, then participants' criterion-setting policy becomes more enigmatic. The policy would have to explain why the degree of learning for strong items affects criterion placement when these items are *not* tested, but has little or no effect on criterion placement when they *are* tested. It would also need to explain why strengthening items that are not tested results in a more liberal criterion.[9]

Conclusion

We reported four experiments that challenge differentiation models by showing that strengthening items does not directly reduce the false alarm rate. Instead, changes in the false alarm rate occur based on changes in the expected strength of the test targets. We have also shown that the expectation-based explanation can be expressed in terms of a formal model that specifies memory processes to a degree of detail similar to that of popular differentiation models. The evidence for differentiation's role in the strength-based mirror effect has come from tasks in which changes in retrieval expectations are potentially confounded with changes in study list content (Criss, 2006, 2009, 2010). When we manipulated study list content in conditions designed to hold retrieval expectations constant, we saw no differentiation-based differences.

Acknowledgments

Preparation of this article was supported by NIMH Grant R37-MH44640 and NIA Grant RO1-AG17083.

References

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology, 29*, 461–473.

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 587–599.

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science, 18*, 40–45.

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.

[9] In fits of the diffusion model, Criss (2010) observed that the starting point changed from .062 to .067 from weak to strong lists, with higher values indicating more liberal responding. This result provides some empirical precedent for expecting more liberal responding on stronger lists. However, as we have noted, the changes in drift rates in this experiment are consistent with a conservative shift in the drift criterion, the model's other bias parameter. The drift rate changes far outweighed the small change in starting point, and participants had fewer false alarms for strong lists as expected.

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

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

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*, 361–376.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 302–313.

Hockley, W. E., & Niewiadomsky, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition, 35*, 679–688.

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

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior, 78*, 567–595.

Maddox, W. T., & Bohil, C. J. (2005). Optimal classifier feedback improves cost-benefit but not base-rate decision criterion learning in perceptual categorization. *Memory & Cognition, 33*, 303–319.

Marsh, R. L., Meeks, J. T., Cook, G. I., Clark-Foos, A., Hicks, J. L., & Brewer, G. A. (2009). Retrieval constraints on the front end create differences in recollection on a subsequent test. *Journal of Memory and Language, 61*, 470–479.

McCabe, D. P., & Balota, D. A. (2007). Context effects on remembering and knowing: The expectancy heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 536–549.

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

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 1095–1110.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*, 465–494.

Nelder, J. A., & Meade, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313.

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

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92*, 212–225.

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*, 163–178.

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: Nonmonotonic lag functions. In R. S. Nickerson (Ed.). *Attention and performance* (Vol. VIII). Hillsdale, NJ: Erlbaum.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.

Ratcliff, R., & Murdock, B. B. Jr., (1976). Retrieval processes in recognition memory. *Psychological Review, 83*, 190–214.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*, 261–300.

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

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

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*, 976–984.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34*, 125–137.

Starns, J. J., & Ratcliff, R. (in press). The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging*.

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition, 36*, 1–8.

Starns, J. J., Hicks, J. L., & Marsh, R. L. (2006). Repetition effects in associative false recognition: Theme-based criterion shifts are the exception, not the rule. *Memory, Special Issue: Memory Editing Mechanisms, 14*, 742–761.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1379–1396.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*, 68–111.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582–600.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*, 254–262.