Mem Cogn (2012) 40:1189–1199
DOI 10.3758/s13421-012-0225-1

The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation

Jeffrey J. Starns · Corey N. White · Roger Ratcliff

Published online: 27 June 2012

Abstract Criss (Cognitive Psychology 59:297–319, 2009) reported that subjective ratings of memory strength showed a mirror effect pattern in which strengthening the studied words increased ratings for targets and decreased ratings for lures. She interpreted the effect on lure items as evidence for differentiation, a process whereby lures produce a poorer match to strong than to weak memory traces. However, she also noted that participants might use different mappings between memory evidence and levels of the rating scale when they expected strong versus weak targets; that is, the effect might be produced by decision processes rather than differentiation. We report two experiments designed to distinguish these accounts. Some participants studied pure lists of weak or strong items (presented once or five times, respectively), while others studied mixed lists of half weak and half strong items. The participants from both groups had pure-strength tests: Only strong or only weak items were tested, and the participants were informed of which it would be before the test. The results showed that strength ratings for lures were lower when strong versus weak targets were tested, regardless of whether the study list was pure or mixed. In the mixed-study condition, the effect was produced even after identical study lists, and thus the same degree of differentiation in the studied traces. Therefore, our results suggest that the strength-rating mirror effect is produced by changes in decision processes.

Keywords Memory models · Recognition · Mirror effects

J. J. Starns (✉)
Department of Psychology, University of Massachusetts Amherst,
441 Tobin Hall,
Amherst, MA 01003, USA
e-mail: jstarns@psych.umass.edu

C. N. White
University of Texas,
Austin, TX, USA

R. Ratcliff
Ohio State University,
Columbus, OH, USA

In recognition memory experiments, participants study a list of items and then are asked to discriminate items from that list (targets, or old items) from items appearing for the first time at test (lures, or new items). The proportion of "old" responses to targets defines the hit rate (HR), and the proportion of "old" responses to lures defines the false-alarm rate (FAR). When memory for the studied items is strengthened, the HR increases and the FAR decreases, a pattern described as the *strength-based mirror effect* (SBME). This effect has been replicated with a variety of strengthening manipulations, such as repeating items on the study list (Stretch & Wixted, 1998), increasing encoding time (Ratcliff, Clark, & Shiffrin, 1990), decreasing the retention interval (Singer & Wixted, 2006), and presenting pictures versus words (Israel & Schacter, 1997).

Competing accounts have explained the SBME in terms of either decision processes or the inherent properties of memory traces. The most common decision-based account is a criterion shift in signal detection theory. Signal detection models assume that recognition decisions are based on continuously varying evidence (Egan, 1958). Targets tend to have higher evidence values than lures, but both show considerable variability from one trial to the next. A criterion is established at some point along the evidence continuum, and any strength value that exceeds this criterion leads to an "old" response. When target items are strengthened, the target distribution shifts up to a higher average evidence

value. Many theorists have assumed that participants respond to this shift by also moving their response criterion to a higher value; that is, when participants know that targets will have high memory strength, they require stronger evidence to say "old" (Brown, Lewis, & Monk, 1977; Hirshman, 1995; Stretch & Wixted, 1998). When memory is strengthened, the target distribution shift produces a higher HR, and the criterion shift produces a lower FAR, accounting for the SBME.

The decision-based account has also been extended to the bind–cue–decide model of episodic memory (BCDMEM; Dennis & Humphreys, 2001; Starns, White, & Ratcliff, 2010) and the drift criterion in the diffusion model (Starns, Ratcliff, & White, in press). Detailed discussions of the drift criterion can be found elsewhere (Ratcliff, 1978, 1985; Ratcliff, Van Zandt, & McKoon, 1999; Ratcliff, & McKoon, 2008; Starns et al., in press). Here, we simply note that shifts in the drift criterion are conceptually very similar to shifts in the signal detection criterion, and both occur for the same proposed reason (i.e., participants change the amount of memory evidence needed to support an "old" response). BCDMEM evaluates the matches and mismatches between retrieved context information and the learning context. Lures retrieve features of preexperimental contexts, and the preexperimental features have some random overlap with features of the study list context. Unlike the differentiation model discussed below, the number of matches and mismatches for lure items does not change from weak to strong study lists. However, when a higher degree of learning is expected at test, each mismatch provides stronger evidence that the test item is a lure. That is, if one expects to retrieve almost all features of the study context for a target item (strong learning), then a lure that fails to retrieve, say, two studied context features might receive a "new" response. In contrast, if one expects that targets will retrieve only a few features of the studied context (weak learning), then a lure that fails to retrieve two studied context features might still be considered "old." Thus, expecting strong targets at test produces a lower FAR. As such, BCDMEM can accommodate the SBME with decision processes that are conceptually similar to the criterion shift account.

An alternative account explains the SBME on the basis of the properties of strong memory traces, without assuming changes in decision standards at retrieval. Models in this class share a property called *differentiation*, whereby strengthening a memory trace makes it less confusable with the traces of other items (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997; see also Shiffrin, Ratcliff, & Clark, 1990). This account has been most thoroughly explored using the retrieving effectively from memory model (REM; Criss, 2006, 2009, 2010). In REM, items are represented by vectors of feature values (with length w). Each feature is a random draw from a geometric distribution with the probability parameter g. Episodic traces begin as vectors of zeros, and the zeros are replaced with a feature value with probability u on each of t learning attempts. The stored feature is accurately copied from the studied item with probability c, otherwise a random draw from the geometric distribution is stored. Strong items receive more learning attempts than do weak items; thus, a higher proportion of the item's features are stored in the memory trace. At retrieval, the features of the test word are compared to the features in each memory trace to yield the relative likelihood that the trace and the probe represent the same item versus different items. These likelihood ratios are averaged across all traces in the match set to produce the global odds ratio that the test item is a target versus a lure. Because strong traces have more stored features than do weak traces, they produce a better match when the target item is used as a test probe. Therefore, the strong-target HR is higher than the weak-target HR.

REM's account of why the FAR is lower on strong lists hinges on the effect of missing versus mismatching features on the likelihood ratio (with high likelihood ratios representing strong evidence that the trace and the probe represent the same item). Consider the match between a lure test probe and one of the traces from the study list. Some features will be missing in the test probe because they were not successfully encoded during learning. These missing features provide no information about whether or not the trace and the test probe represent the same item; thus, missing features do not affect the likelihood ratio. Some features of the test probe might match the studied trace due to random overlap in features between items, but these matches should be rare. Most of the features stored in the trace should mismatch the probe, and these mismatching features decrease the likelihood ratio. Strengthening target traces tends to replace missing feature values with feature values that mismatch lure items, thereby lowering the likelihood ratio for lures and decreasing the FAR. Critically, the drop in the FAR is produced by actual changes in evidence, not by changes in decision processes.

Supporting differentiation with direct strength ratings

Criss (2009) attempted to discriminate the criterion-shift and differentiation accounts of the SBME using a performance measure that is not affected (or at least minimally affected) by response bias. She adopted the direct-ratings method developed by Mickes, Wixted, and Wais (2007), whereby participants are asked to rate the strength of their memory for each test item on a scale with many levels (e.g., a 20-point scale). If the subjective strength ratings reflect the underlying distributions of memory strength, differentiation models predict that ratings should be higher for targets and lower for lures following strong lists versus weak lists. Criss (2009) found precisely this pattern in her first two experiments. To establish that strength ratings are minimally

affected by response biases, in a third experiment she evaluated the effect of target proportions on ratings. Specifically, following equal-strength lists, the participants completed a test composed of either 30 % or 70 % target items. Changing the proportion of targets produced biases in participants' old–new judgments, with a higher HR and FAR in the 70 % target condition as compared to the 30 % target condition. In contrast, the strength ratings were only slightly higher in the 70 % condition, and the effect did not approach significance for either targets or lures. Criss (2009) interpreted these results as evidence that strength ratings are relatively insensitive to bias, meaning that the decrease in lure strength ratings from weak to strong lists must have indicated a change in the lure distribution produced by differentiation.

Although suggestive, the target proportion results do not decisively rule out a role for decision processes in producing the SBME. The conclusion that biases do not affect strength ratings is based on a nonsignificant result for a null-hypothesis significance test, which of course does not establish the absence of an effect (but does suggest that the effect might not be very large, if it exists). More importantly, direct comparisons between the strength effect and target proportion experiments are difficult, given that a different variable is being manipulated. Even if strength ratings are not affected by biases introduced by target proportion, they might be influenced by changes in decision processes introduced by a strength manipulation.

Our goal was to more closely evaluate whether or not strength ratings can be affected by changes in decision processes. We followed the general design used by Starns et al. (2010; see also Starns et al., in press), with the addition of collecting subjective strength ratings at test. Specifically, we ran a pure-study condition in which participants made strength ratings following lists of all weak items (studied once) or all strong items (studied five times). This condition confounded a change in the differentiation of the studied traces with the opportunity to change decision processes, in that participants knew whether strong or weak items would be tested. We also ran a mixed-study condition in which all of the study lists were a mixture of items studied once and items studied five times, producing the same overall degree of differentiation for all lists. However, the *tests* were pure strength, with only strong or only weak targets appearing on a given test.[1] The participants were informed of the target strength before each test, providing them with the same opportunity to adjust decision processes on the basis of strength that the participants in the pure condition had. In the following section, we present a detailed discussion of the predictions from each account, supported by simulation results from the REM model.

[1] We use the label "mixed-study" to distinguish our procedure from designs that use mixed lists at both study and test, which is by far the most common procedure for conditions that are labeled "mixed" (e.g., Ratcliff et al., 1990). We did not make this distinction in previous studies (Starns et al., in press; Starns et al., 2010), although what we referred to as "mixed" conditions in those studies actually used the same procedure as in the present mixed-study condition (i.e., mixed study lists with pure tests).

Predictions

Criss (2009) simulated predictions from the REM model for the pure-list design with standard parameter values from previous applications of the model (w = 20, g = .35, c = .7, u = .04, t_{WEAK} = 10, t_{STRONG} = 17). The model simulations showed that ratings for both targets and lures diverged from strong to weak lists, just as observed in the data. We simulated REM for our extended design, with the list structure in the simulations matching the lists from our actual experiments. We used all of the parameter values from Criss's (2009) simulation, except that the parameter for the number of learning attempts (t) was reduced to 6 for weak items and 12 for strong items, because our study lists had a faster presentation rate than those in the Criss (2009) experiments. For the simulations, feature vectors for each item in the experiment were randomly sampled, and the items on the study list were probabilistically encoded into memory traces with either the weak or the strong learning rate (depending on condition). To simulate the test phase, the features of each test item were used as a memory probe. This probe was matched to all of the traces encoded on the last list to obtain an odds ratio (for more details on simulating REM, see Shiffrin & Steyvers, 1997). Nineteen criteria were used to divide the odds ratios into the 20 rating responses. The criteria were placed at equal intervals between –1.6 and 4.5 on a log scale. Memory probes with a log-odds ratio below the lowest criterion were assigned a rating of 1; items above the first criterion and below the second were assigned a rating of 2; and so forth.

The simulation results are reported in Fig. 1. The plots show predicted cumulative distributions for the strength ratings. For example, the point farthest to the left shows the proportion of items given a rating of 1, the next point shows the proportion given a rating of either 1 or 2, and so forth. These points can also be interpreted as the proportions of the strength distributions produced by REM that fall *below* a given criterion, with the leftmost point representing the most liberal criterion and the rightmost point representing the most conservative. In the pure condition, the cumulative distributions diverged on the basis of strength for both targets and lures, just as in Criss's (2009) original simulations. That is, additional study shifted the target distribution to higher odds values;

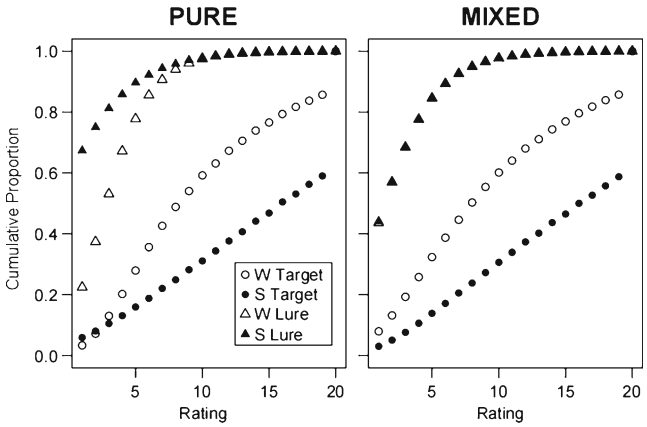

Fig. 1 Cumulative strength-rating distributions predicted by the REM model for pure and mixed study lists in our experiments. The points at each rating value are the proportions of the evidence distributions that fall to the left of a criterion, so conditions with higher average strength will have lower proportions. W = weak tests (target items studied once); S = strong tests (target items studied five times)

therefore, a smaller proportion of the strong than of the weak targets fell below a given criterion (especially for the higher criteria). In addition, traces were more differentiated following strong than following weak lists, which shifted the lure distribution to lower odds values. As a result, more of the strong distribution fell below a given criterion (especially for the lower criteria). The pattern was different for mixed study lists. Although the target distributions still diverged on the basis of strength, the lure distributions did not (i.e., the symbols overlap completely). In other words, the degree of differentiation across the studied traces was the same for the strong and weak tests. Again, in generating these predictions, we applied the same model used by Criss (2009), and only the structure of the study lists changed to match the lists in our experiments (including the added mixed-study condition). In the General Discussion, we discuss an extension of REM in which the test items are also stored in memory, but we will first consider the model that has been previously applied to strength-rating data.

The differentiation account predicts that the effect of strength on ratings for lure items will be different for pure versus mixed study lists. In contrast, if the effect on lure ratings is produced by strength-based changes in decision processes, one would expect to find the same effect with both mixed and pure study lists. That is, both of these conditions provide an opportunity to adjust decision processes on the basis of expected strength, so both should show lower lure ratings for strong than for weak tests. A strength effect only in the pure-study condition, with no effect (or a much smaller effect) in the mixed-study condition, would provide strong evidence against the decision-process account.

Experiments 1 and 2

Method

Design and participants The type of study list (pure vs. mixed) was manipulated between participants. Experiment 1 included 15 participants in the pure-study condition and 18 in the mixed-study condition, whereas Experiment 2 had 21 for pure study and 20 for mixed study. Each participant completed 12 study–test cycles. Within each condition, the target strength at test was manipulated within participants across study–test cycles, with half of the tests including weak targets and half strong targets. For the pure-study condition, the cycles with strong tests also had strong study lists, and cycles with weak tests also had weak study lists. For the mixed-study condition, all cycles had mixed-strength study lists, and only the strength of the targets included on the test changed.

Materials The stimuli were randomly selected from a pool of 814 high-frequency words to create 12 study lists of 24 words each. In the pure condition, each word was presented once on the weak study–test cycles versus five times on the strong study–test cycles. Each test included 12 targets randomly selected from the studied items and 12 lures not seen before in the experiment. In the mixed-study condition, the study lists were always composed of 12 words studied once and 12 words studied five times. On weak study–test cycles, the words presented once were re-presented at test with 12 lure words. On strong study–test cycles, the words presented five times served as the test targets.

Procedure The initial instructions informed participants that they would study a list of words and then be shown a test list with some old (studied) words and some new (unstudied) words. They were told to rate the strength of their memory for each test word on a 1–20 scale, with 1 indicating *very weak memory* and 20 indicating *very strong memory.* They were also informed that there would be a number task between the study and test lists and informed how to complete it (see below). The participants in the first experiment predominantly used the endpoints of the strength-rating scale, so participants in Experiment 2 were asked to distribute their responses across the rating scale. Specifically, they were encouraged to use the different ratings to better reflect differences in memory strength, with the ratings of 20 and 1 reserved for the highest possible certainty that a word was or was not on the list, respectively. They were also explicitly asked to refrain from relying solely on a few ratings for all of their responses. For each study list, words appeared on the computer screen one at a time for 800 ms, with a 150-ms blank screen between. The order of each study list was random with the constraint that at least one word intervened

before a word was repeated. After each study list, the participants completed 30 trials of an *N*-back task. The digits 1–9 appeared on the screen in a random order for 900 ms each, followed by 100 ms of blank screen. Participants were asked to press a key when the current digit was the same as the digit two back in the sequence. After all of the trials, the participants saw a 4-s feedback screen with the number of targets in the last sequence and the number of times they had hit the key when a target occurred. The participants were then prompted to begin the test with a message that informed them that the targets would be items that were studied once (weak cycles) or items that were studied five times (strong cycles). To advance to the test, participants had to respond by hitting the "1" key for weak cycles or the "5" key for strong cycles. Each test word remained on the screen until the participant had entered a strength rating. A TOO FAST message appeared on the screen for any response time under 250 ms.

Results and discussion

Table 1 shows the average ratings from both experiments, and Fig. 2 shows the strength effects with 95 % confidence intervals.[2] As expected, strength ratings for targets increased for words studied five times versus words studied once. As the target differences were clear and predicted by both accounts under investigation, we did not analyze them further. Lure ratings decreased from weak to strong tests, although the differences were much smaller than the target effect. Criss (2009) also found a smaller effect for lure ratings.

We analyzed the average lure ratings from each participant with a 2 × 2 ANOVA with Study List Type (pure vs. mixed) and Strength of the Test Targets (strong vs. weak) as factors. Both experiments showed a significant main effect of strength [$F(1, 31) = 13.88$, $MSE = 1.150$, $p < .001$, for Exp. 1; $F(1, 39) = 19.26$, $MSE = 1.009$, $p < .001$, for Exp. 2], but no main effect of list type [$F(1, 31) = 0.89$, $MSE = 12.753$, n.s., for Exp. 1; $F(1, 39) = 0.02$, $MSE = 6.276$, n.s., for Exp. 2]. The strength effects were quite similar in size for the pure-study and mixed-study conditions, and the interaction did not approach significance for either experiment [$F(1, 31) = 0.10$, n.s., for Exp. 1; $F(1, 39) = 0.12$, n.s., for Exp. 2]. Critically, the mixed-study conditions from both experiments showed a significant strength effect on lure ratings (see the confidence intervals in Fig. 2).

For comparison with previous studies, Table 2 shows the results in terms of traditional HR and FAR measures, with ratings 1–10 defined as "new" responses and ratings 11–20 defined as "old" responses. All comparisons show a mirror pattern, with the HR increasing and the FAR decreasing on the basis of strength. We submitted the FAR data to the same 2 × 2 ANOVA model used for the lure ratings. Both experiments showed a significant strength effect on FAR [$F(1, 31) = 18.21$, $MSE = 0.003$, $p < .001$, for Exp. 1; $F(1, 39) = 19.94$, $MSE = 0.005$, $p < .001$, for Exp. 2]. Neither experiment showed a difference between pure and mixed study lists [$F(1, 31) = 0.62$, $MSE = 0.037$, n.s., for Exp. 1; $F(1, 39) = 0.02$, $MSE = 0.023$, n.s., for Exp. 2]. The strength effect on FARs was as large in the mixed-study condition as in the pure-study condition, and neither experiment showed an interaction of list type with strength [$F(1, 31) = 0.08$, n.s., for Exp. 1; $F(1, 39) = 1.67$, n.s., for Exp. 2].

Figure 3 shows the cumulative rating distributions. The distributions diverged on the basis of strength for both targets and lures, and this pattern emerged with both pure and mixed study lists. The Experiment 2 distributions

Table 1 Average subjective strength ratings

Experiment and Study List Condition	Item Type and Strength			
	Target		Lure	
	Weak	Strong	Weak	Strong
Ex. 1				
Pure	11.71 (0.74)	15.31 (0.51)	6.07 (0.78)	5.00 (0.76)
Mixed	10.44 (0.48)	14.58 (0.38)	5.15 (0.58)	4.25 (0.52)
Ex. 2				
Pure	10.89 (0.37)	14.46 (0.38)	6.58 (0.42)	5.69 (0.42)
Mixed	10.71 (0.63)	14.50 (0.49)	6.74 (0.46)	5.69 (0.38)

Standard errors are in parentheses

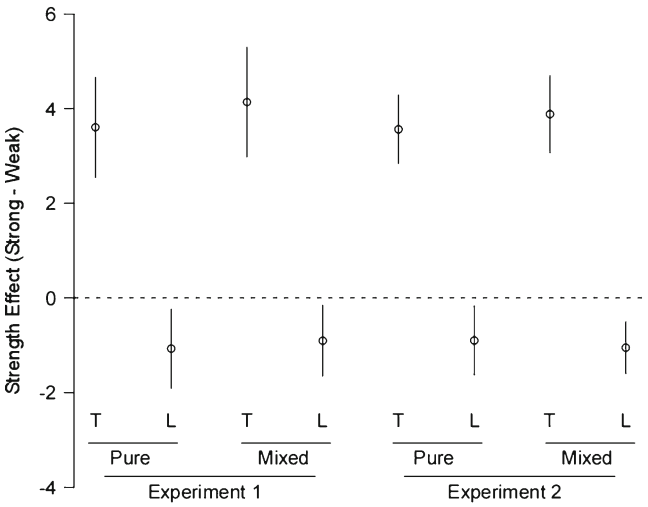

Fig. 2 Strength effects on the rating-scale measure with 95 % confidence intervals on the difference between strong and weak. T = target; L = lure

[2] Of course, any confidence interval that does not include zero indicates that the strength difference is significant by a paired-samples *t* test with $\alpha = .05$.

Table 2 Proportions of "old" responses (ratings 11–20)

Experiment and Study List Condition	Item Type and Strength			
	Target		Lure	
	Weak	Strong	Weak	Strong
Ex. 1				
Pure	.54 (.04)	.75 (.03)	.21 (.05)	.15 (.04)
Mixed	.46 (.03)	.69 (.03)	.18 (.03)	.11 (.02)
Ex. 2				
Pure	.52 (.03)	.74 (.02)	.21 (.03)	.16 (.03)
Mixed	.48 (.04)	.74 (.03)	.23 (.03)	.14 (.02)

Standard errors are in parentheses

showed a more gradual rise, demonstrating that participants responded to the instruction to spread out their ratings. The results were otherwise very similar across experiments.

We used two different approaches with complementary strengths and weaknesses to analyze the full distributions of strength ratings. The primary concern for both analyses was whether ratings for lures changed between strong and weak tests within the pure-study and mixed-study conditions. Our first strategy was to analyze the rating results by applying a signal detection model. This model assumes that memory strength follows Gaussian distributions. All of the parameters were scaled relative to the lure distribution, which was assumed to have a mean of zero and a standard deviation of 1. Both the mean (μ) and standard deviation (σ) of the target distribution were estimated as free parameters that could vary on the basis of target strength. The model used 19 criteria (λ) to map strength values onto the 20-point rating scale. For each list type, we compared a *no-shift model*, in which the same 19 λ parameters were used on both the strong and weak tests, to a *shift model*, in which the λ parameters could differ between strong and weak. To eliminate the need to dramatically increase model complexity by estimating 19 free λ parameters for each strength condition, the shift model had a single parameter for the difference between strong and weak (λ_{SHIFT}) with all of the λ parameters moving in lockstep. We compared the shift and no-shift models with both the Akaike and Bayesian information criteria (AIC and BIC, respectively), with lower values indicating the preferred model (Akaike, 1979; Schwarz, 1978). If lure ratings followed the same distribution for strong and weak tests, the no-shift model should provide the best account of the data. If the distributions of lure ratings were affected by strength, the shift model should then be preferred.

To be clear, using the λ shift model does not necessarily imply that the change in lure ratings was produced by a decision process. The λ parameters represent the distance of each criterion *from the mean of the lure distribution*; thus, they would be expected to increase if differentiation shifted

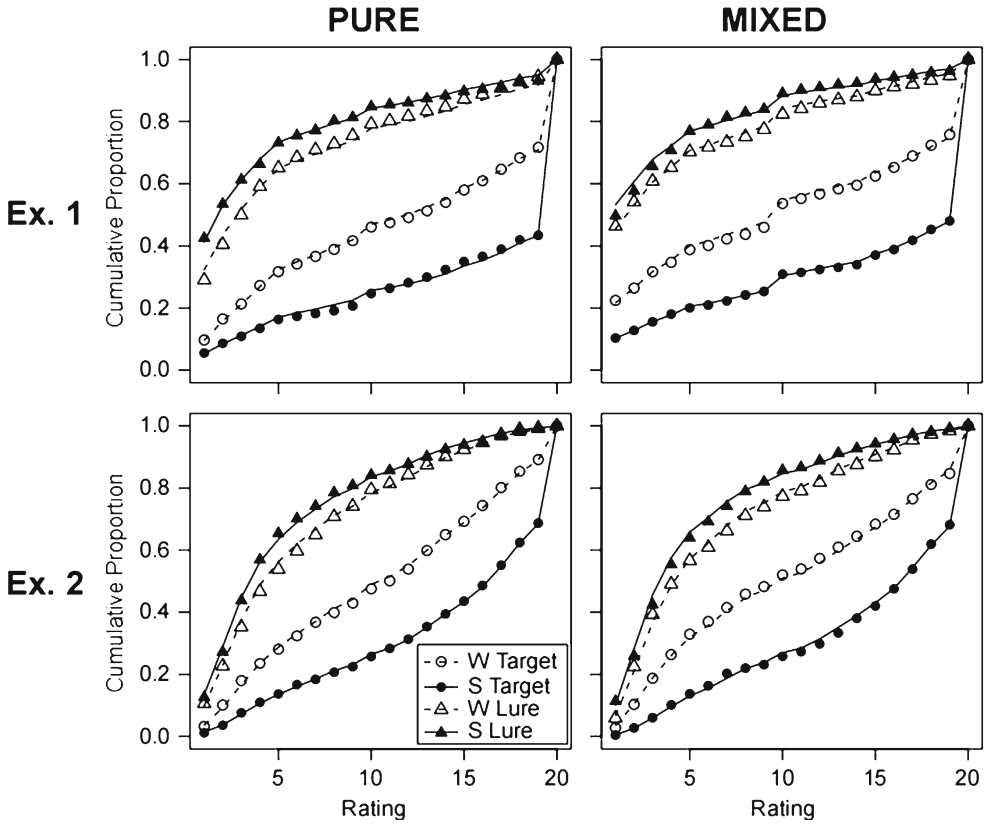

Fig. 3 Empirical cumulative distributions of strength ratings from Experiments 1 and 2. The lines are the fits of the signal detection model with λ parameters allowed to shift between strong and weak tests. W = weak tests (target items studied once); S = strong tests (target items studied five times)

the mean of the lure distribution to a lower value without affecting the absolute positions of the criteria. Of course, the λ values would also increase if the response criteria shifted to higher values without a change in the position of the lure distribution. The critical result for distinguishing these two accounts was whether the shift in the λ parameters was tied to differentiation from the study list. If so, the shift should be observed only in the pure conditions. If not, the shift should be observed even with mixed study lists.

Table 3 reports the best-fitting parameter values and the fit statistics. Figure 3 shows the cumulative rating distributions predicted by the signal detection model as the lines running through the data points. The fits were to group data, but the model results converged on the same conclusions as the analyses on average ratings for individual participants reported above. As expected, strong targets had a higher average strength than did weak targets. Replicating previous work (e.g., Wixted, 2007), the σ values consistently indicated that target evidence was more variable than lure evidence (the only exception was the weak targets in the pure condition of Exp. 1). The λ-shift parameters indicated that the criteria were closer to the mean of the lure distribution on weak than on strong tests. Crucially, the size of this shift was very similar in the pure-study and mixed-study conditions (.24 vs. .19 in Exp. 1, .19 vs. .23 in Exp. 2). AIC statistics preferred the shift model over the no-shift model across all comparisons. BIC preferred the shift model for all conditions except the mixed-study condition in Experiment 1, in which the BIC values for the two models were nearly identical. Therefore, the modeling results are consistent with the analyses on average strength ratings, and both show a strength effect in lure ratings with both the pure and mixed study lists.

Table 3 Parameters and fit statistics from signal detection model fits

Data Set and Submodel	Parameter and Fit Statistic							
	μ_S	μ_W	σ_S	σ_W	λ_{SHIFT}	G^2	AIC	BIC
Ex. 1 Pure Study								
Shift model	1.85	0.84	1.29	1.00	0.24	85.2	505.0	653.3
No-shift model	1.78	0.98	1.34	1.03	0*	106.2	524.0	666.1
Ex. 1 Mixed Study								
Shift model	1.96	0.86	1.49	1.24	0.19	73.6	496.0	648.5
No-shift model	1.88	0.96	1.50	1.25	0*	81.8	502.2	648.3
Ex. 2 Pure Study								
Shift model	1.82	0.82	1.34	1.18	0.19	69.3	536.4	692.7
No-shift model	1.73	0.92	1.35	1.19	0*	87.4	552.5	702.3
Ex. 2 Mixed Study								
Shift model	1.77	0.77	1.21	1.25	0.23	78.6	537.6	692.2
No-shift model	1.65	0.88	1.21	1.24	0*	101.9	559.0	707.1

Parameters marked with an asterisk were fixed in the fits

The signal detection approach is attractive, in that it captures the effect of strength on the lure distributions in terms of a single parameter (λ_{SHIFT}). However, the model also makes simplifying assumptions that might not be appropriate for our rating distributions (although note the close correspondence between the model and the data shown in Fig. 3). Most critically, REM does not produce Gaussian distributions of memory strength, as assumed in the signal detection model. To ensure that these simplifying assumptions did not distort our conclusions, our second strategy for analyzing the rating distributions made no distributional assumptions. We simply performed a χ^2 test on the lure rating data to compare two models: one that estimated different distributions of rating probabilities on strong and weak tests, and one that used the same distribution for strong and weak. In detail, the same-distribution model had 19 free parameters; that is, free parameters for the probability of using each of the 20 levels of the rating scale (one parameter was fixed to ensure that the probabilities would sum to 1). The different-distribution model had 38 free parameters: two sets of 19 free parameters for lures on the strong and weak tests (note that this model is saturated; i.e., it has as many free parameters as there are degrees of freedom in the data). The test between the models has 19 degrees of freedom.

The χ^2 tests showed that allowing different lure distributions on strong and weak tests significantly improved the fit, and this result was obtained for both pure and mixed study lists in each experiment [Exp. 1 pure, $\chi^2(19) = 68.60$, $p < .001$; Exp. 1 mixed, $\chi^2(19) = 32.82$, $p < .05$; Exp. 2 pure, $\chi^2(19) = 46.16$, $p < .001$; Exp. 2 mixed, $\chi^2(19) = 53.04$, $p < .001$]. In Experiment 1, the χ^2 statistic was larger for the pure-study condition (68.60) than for the mixed-study condition (32.82), perhaps indicating a greater deviation in the distributions with pure study lists. However, the opposite pattern was observed in Experiment 2 (46.16 and 53.04 for pure and mixed study lists, respectively). Most importantly, the χ^2 tests confirmed our key conclusion from the average rating and signal detection analyses: Lure ratings differed between strong and weak tests even in the mixed-study condition.

General discussion

These results demonstrate that subjective strength ratings are sensitive to changes in decision processes. Even when the content of the study list was held constant (i.e., the mixed-study condition), participants gave lower strength ratings for lures when they were informed that only strong

versus only weak targets would appear at test. This effect was confirmed by three different strategies for analyzing the rating data. The results in the mixed condition were very similar to those in the pure condition, even though differentiation from the study list was equated in the former but not the latter. Thus, effects in strength ratings cannot be directly interpreted as changes in the evidence distributions produced by a differentiation mechanism.

Comparison with Criss (2009)

In contrast to Criss (2009), participants in our experiments were informed of the target strength just before the test. Of course, this design was critical in the mixed-study condition, because the study lists were always the same and could not guide participants' expectations for the test. However, this aspect of our design might have encouraged changes in decision processes that do not characterize experiments without a strength message. For example, it might be that Criss's (2009) participants did not change their decision processes on the basis of strength, so the observed changes in lure ratings could be attributed to differentiation.

Although we acknowledge this possibility, the participants in a pure-list condition have all the information they need to adjust their decision processes, even without a strength-specific test message. We gave pure-study participants the same test message as the mixed-study participants to keep the procedures as similar as possible between conditions, but the message reported no new information in the pure condition. Of course participants would expect to have stronger memories following an entire list of items studied five times versus an entire list of items studied once, and it is safe to assume that they were capable of noticing the difference between strong and weak lists. For example, if participants were asked just before the test whether they had studied the items on the last pure-strength list once or multiple times, we would expect them to show perfect accuracy. Moreover, even if participants only change decision processes when they are given a strength-specific test message, our results still show that decision processes can influence subjective strength ratings. So, although one can say that the Criss (2009) results *might* reflect differentiation, there is no point in saying that they *must* reflect differentiation because they were obtained using a dependent variable that was not sensitive to decision processes. Using strength ratings does not help to discriminate the differentiation and decision accounts.

Expanding on the points above, we think it is appropriate to regard the decision-process account as a default explanation. The type of adjustment involved is conceptually simple, and changes of this sort are evidenced across a wide variety of decision tasks. The differentiation account is specific to a class of memory models, and it is central to how the models accommodate recognition data. Thus, we think it is fair to say that the onus is on supporters of differentiation to provide forms of evidence that are not easily explained in terms of decision processes.

Our results do not help to explain why Criss (2009) found that strength ratings were not affected by the proportions of targets on the test. Previous results have shown an effect of this variable—as well as of other types of decision biases—on rating scales that have fewer than 20 levels (Mueller & Weidemann, 2008; Ratcliff & Starns, 2009; Van Zandt, 2000). Future research will be needed to better define the effects of proportions on rating scales and to explore the potential role of the number of levels. However, we think that this issue is best considered orthogonal to the cause(s) of the SBME. Target proportion and expected strength might affect the decision process in different ways, and only the effect of expected strength is directly relevant for the SBME.

Differentiation at test

We derived our predictions from the basic REM model that Criss (2009) used to accommodate strength-rating data. Some applications of REM have included the assumption that test items are added to memory as the test proceeds (Criss, Malmberg, & Shiffrin, 2011), and it is possible that only some repetitions are stored in the same memory trace (Shiffrin & Steyvers, 1997). In this sort of application, the model matches each test item to memory in order to determine the odds that the item was previously presented. If the odds ratio indicates that the item does not match any previously presented item, the model returns a "new" response *and* creates a new memory trace that stores the item's features. If the odds ratio indicates that the item was previously presented, the model returns an "old" response and stores the features of the test probe in the existing trace with the highest match (thus further differentiating this trace). Together, these assumptions introduce the possibility that differentiation could produce effects based on the test content, even with identical study lists. For example, target items on a strong test will be more likely to be stored in a trace established at encoding than will targets on a weak test. Thus, the degree of differentiation will be higher by the end of the strong test than by the end of the weak test, perhaps accounting for the difference in lure responding that we observed in the mixed-study condition. We acknowledge that our design cannot definitively rule out the differentiation-at-test explanation, and it is an interesting idea that should be more carefully tested. However, several results from our own data and from outside studies are problematic for this account, and we list these in detail below.

The first evidence against the differentiation-at-test explanation comes from a reanalysis of our own data. To assess the influence of differentiation at retrieval, we separately analyzed items from the first half and the second half of each test in the mixed-study condition. If test differentiation created our FAR differences, then these differences should be larger for later test items. That is, with mixed study lists, there is no differentiation difference at the *beginning* of strong and weak tests, but a differentiation difference can build up as more targets are tested. For the first experiment, the strength effect in FAR was .08 for the first half of the test and .06 for the second half. For Experiment 2, the effect was .06 on the first half and .10 on the second. Thus, there was no consistent evidence for the test differentiation account: A strength difference was observed even early in the tests, and the effect sizes were quite similar in the early and later portions of the test (bigger effect for first half in one experiment, smaller in the other).

Second, the differentiation-at-test account cannot explain why the strength effects on lure ratings were the same size with pure and mixed study lists. These conditions had identical tests for strong and weak, so differentiation from the test items was equated. However, the pure condition had an additional differentiation difference created by studying the list items once versus five times. If differentiation were truly the basis for the difference in lure ratings, one would expect a larger strength effect with pure than with mixed study lists. This prediction is contradicted by the present data, and Starns et al. (in press) also showed very similar FAR effects with pure and mixed study lists. Of course, our conclusion that strength effects are no larger with pure lists is based on a null effect, and thus should be viewed with caution. Nevertheless, both of the studies in question showed significant strength effects. If these strength effects are to be interpreted in terms of differentiation at test, some explanation is needed as to why the effect of differentiation from the study list was so difficult to detect, even though the same process produces clear effects when it operates at retrieval.

Finally, some experiments have reported changes in FAR with the degrees of differentiation controlled at study *and* test. A few studies have shown mirror effects when strength is signaled on an item-by-item basis during the recognition test (Singer, 2009; Singer & Wixted, 2006), although this effect is generally difficult to obtain (e.g., Stretch & Wixted, 1998). The tests in these experiments contained an equal mix of strong and weak targets in a random order, so there was no basis for changes in differentiation at test. In a recent study, participants learned mixed lists of strong and weak items, followed by a test in which alternating blocks of trials contained only strong or only weak targets (Starns & Hicks, 2011). For some participants, the strong and weak test blocks were in two different colors to make them easy to distinguish, and other participants saw all of the test words in black. With larger block sizes (e.g., 20 items per block), the participants with color cues had a lower FAR on strong blocks than on weak blocks, but participants without color cues showed no hint of this effect, even though they encountered the exact same study and test structures. These results cannot be attributed to differentiation in traces established at study or test, but they are consistent with the notion that participants use more conservative decision standards when they expect stronger items at test.

Qualitative versus quantitative decision standards

Scimeca, McDonough, and Gallo (2011) found no difference in FARs when participants were tested on words studied three times versus words studied once, but they did find a lower FAR with tests on pictures versus words. The authors suggested that qualitative changes in decision processes are more critical than quantitative changes. We acknowledge that qualitative changes are important to consider, but it is unclear why study repetition failed to produce a FAR decrease in Scimeca et al., in contrast to the present results and those of previous studies (Marsh et al., 2009; Starns et al., in press; Starns et al., 2010; Stretch & Wixted, 1998; Verde & Rotello, 2007). Several procedural differences may explain the difference in results, the most obvious being that Scimeca et al. required participants to discriminate between different classes of studied items (i.e., to make source decisions) as opposed to simply recognizing the studied items. Specifically, Scimeca et al. had participants study lists of black words that were seen once, red words that were seen once, and green words that were seen three times. At test, the participants were asked to identify either words studied in red (weak) or words studied in green (strong), and in both cases the lures on the test included both nonstudied words and words studied in black. The need to exclude a class of studied items might have prompted participants to disregard strength in setting their retrieval standards, creating a result different from the one found in standard recognition studies.

Broader implications for REM

Our results suggest that differentiation is not required in order to produce the SBME, but they are not necessarily problematic for the REM model in general. REM includes a mechanism that can produce the SBME in terms of decision processes; that is, the criterion for the odds ratio needed to support an "old" response could be moved to a higher value for strong tests than for weak tests. The criterion is often fixed at a value of 1 in applications of the model (such that anything more likely to be a target than a lure is called "old"). However, Shiffrin and Steyvers (1997) noted that

the criterion might be affected by some experimental variables.

Other results have also put in question the importance of differentiation in memory. Without differentiation, REM incorrectly predicts a positive list-strength effect (Shiffrin et al., 1990; Shiffrin & Steyvers, 1997), which was the primary motivation for developing the differentiation concept in the first place. More recent work has suggested that differentiation might not be necessary to explain null list-strength effects; instead, item interference in recognition memory may simply not be strong enough to produce either a list-strength or list-length effect (Dennis & Chapman, 2010; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011).

Distinguishing the accounts

A general strategy for supporting the differentiation account over the decision-process account has been to find a measure of memory strength that is not affected by decision biases, and this property has been claimed for both subjective strength ratings (Criss, 2009) and diffusion model drift rates (Criss, 2010). By demonstrating significant strength effects following mixed study lists, our findings suggest that both of these measures can be influenced by decision processes (the present experiments; Starns et al., in press). We propose that finding a "pure" measure of strength will not be an effective way to discriminate the accounts; instead, the focus should be on isolating the differentiation and decision mechanisms experimentally.

Currently, the literature provides stronger support for the decision-process account than for the differentiation account. A number of existing studies have demonstrated an SBME with no difference in the degree of differentiation from the study list (Marsh et al., 2009; Starns et al., in press; Starns et al., 2010), and some studies have shown an SBME with no difference in differentiation at study *or* test (Singer, 2009; Singer & Wixted, 2006; Starns & Hicks, 2011). All of these studies have reported positive effects that are best explained by changes in decision processes.

Several studies have now reported a failure to obtain effects predicted by the differentiation account. As mentioned, even if differentiation at test is granted, the differentiation account predicts a bigger strength effect on FARs with pure than with mixed study lists. The present results and those of Starns et al. (in press) both violate this prediction. Starns et al. (2010) also reported a failure to observe differentiation effects in another design. In one condition, participants studied mixed lists of strong and weak items but were informed that only the weak items would appear at test. Separate groups studied the (nontested) strong items either two times or five times, which should have created more differentiated traces in the latter condition. However, the two groups showed no difference in FARs. These are all null results, which are always difficult to interpret. However, all of the studies mentioned did find significant strength effects on FARs when participants had an opportunity to adjust their decision processes, suggesting that experimental power was adequate. If differentiation produces effects so small that they are difficult to observe, then this process probably is not the main source of the robust and easily detected SBME.

Conclusion

The present results add to growing evidence that decision processes play a primary role in the SBME (Marsh et al., 2009; Singer, 2009; Singer & Wixted, 2006; Starns & Hicks, 2011; Starns et al., in press; Starns et al., 2010; Stretch & Wixted, 1998; Verde & Rotello, 2007). None of these studies have decisively ruled out the differentiation process. However, the studies do clearly demonstrate that results that have been used as support for differentiation can be produced by decision processes alone. Thus, different forms of evidence will be required to establish a role for differentiation in the SBME.

Author note This research was funded by AFOSR Grant FA9550-06-1-0055 and NIA Grant R01-AG17083.

References

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika, 66,* 237–242.

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

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. doi:10.1016/j.cogpsych.2009.07.003

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. doi:10.1037/a0018435

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316–326.

Dennis, S., & Chapman, A. (2010). The inverse list length effect: A challenge for pure exemplar models of recognition memory. *Journal of Memory and Language, 63,* 416–424.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. doi:10.1037/0033-295X.108.2.452

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.

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

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.

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4,* 577–581. doi:10.3758/BF03214352

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39,* 348–363.

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. doi:10.1016/j.jml.2009.06.005

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. doi:10.1037/0033-295X.105.4.734-760

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865. doi:10.3758/BF03194112

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494. doi:10.3758/PBR.15.3.465

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92,* 212–225. doi:10.1037/0033-295X.92.2.212

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. doi:10.1037/0278-7393.16.2.163

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59–83. doi:10.1037/a0014086

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300. doi:10.1037/0033-295X.106.2.261

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Scimeca, J. M., McDonough, I. M., & Gallo, D. A. (2011). Quality trumps quantity at reducing memory errors: Implications for retrieval monitoring and mirror effects. *Journal of Memory and Language, 65,* 363–377.

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. doi:10.1037/0278-7393.16.2.179

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

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. doi:10.3758/BF03193392

Starns, J. J., & Hicks, J. L. (2011, November). *Strength-based criterion shifts within a single test: How flexible are they?* Poster presented at the 52nd Annual Meeting of the Psychonomic Society, Seattle, WA.

Starns, J. J., Ratcliff, R., & White, C. N. (in press). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/a0028151

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63,* 18–34. doi:10.1016/j.jml.2010.03.004

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.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. doi:10.1037/0278-7393.26.3.582

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254–262. doi:10.3758/BF03193446

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152–176. doi:10.1037/0033-295X.114.1.152