DOI 10.3758/s13421-014-0397-y

Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory

Jason L. Hicks · Jeffrey J. Starns

Published online: 13 February 2014

Abstract In seven experiments, we explored the potential for strength-based, within-list criterion shifts in recognition memory. People studied a mix of target words, some presented four times (strong) and others studied once (weak). In Experiments 1, 2, 4A, and 4B, the test was organized into alternating blocks of 10, 20, or 40 trials. Each block contained lures intermixed with strong targets only or weak targets only. In strength-cued conditions, test probes appeared in a unique font color for strong and weak blocks. In the uncued conditions of Experiments 1 and 2, similar strength blocks were tested, but strength was not cued with font color. False alarms to lures were lower in blocks containing strong target words, as compared with lures in blocks containing weak targets, but only when strength was cued with font color. Providing test feedback in Experiment 2 did not alter these results. In Experiments 3A–3C, test items were presented in a random order (i.e., not blocked by strength). Of these three experiments, only one demonstrated a significant shift even though strength cues were provided. Overall, the criterion shift was larger and more reliable as block size increased, and the shift occurred only when strength was cued with font color. These results clarify the factors that affect participants' willingness to change their response criterion within a test list.

Keywords Recognition · Mirror effects · Memory strength · Criterion shifts

People constantly encounter information in their environment that cues memories for personal episodic experiences. However, because memories of prior experience cannot be perfectly veridical, people make assumptions and inferences about what their retrieval experiences represent (Johnson et al. 1993). The decision process leading to an interpretation of cued memories involves many possibilities, including whether memories are accurate or false, whether they can be temporally dated, or whether they originated from one possible encoding context or another (Johnson et al. 1993). The rememberer must set a criterion for how strong the evidence from memory must be for the retrieved information to be accepted as a valid prior experience, and this criterion would ideally be adapted to the specific experience in question. For example, imagine that someone asks you "Have you ever driven through Houston?" versus "Did you ever live in Houston?" You should require much stronger memories of Houston to assent to the second question than to the first. In the present experiments, we investigate criterion adjustments in a standard recognition memory task. In particular, we focus on whether people can use expected stimulus strength as a reliable cue for setting and adjusting a recognition memory criterion within a test. In the remainder of this section, we couch this criterion-setting process in the context of a framework for understanding recognition memory decisions—signal detection theory (Green & Swets, 1966).

Although a signal detection analysis of how criterion shifts might be produced by stimulus strength differences has been presented many times (e.g., Hirshman, 1995; Stretch & Wixted, 1998), we briefly summarize it here. Figure 1 depicts a scenario in which people study some items repeatedly and other items once to create strong and weak distributions of target items, respectively. The distribution of stronger items sits farthest to the right on the strength-of-evidence (i.e., familiarity) scale. These items have, on average, more evidence of prior experience. The weak target distribution sits in the middle, being weaker on average, as compared with the repeated targets. Finally, the lure distribution sits farthest to

J. L. Hicks (✉)
Department of Psychology, Louisiana State University, Baton Rouge, LA 70803-5501, USA
e-mail: jhicks@lsu.edu

J. J. Starns
University of Massachusetts Amherst, Amherst, MA, USA

the left, these items having not been encoded in the experimental context. Lures are included on the test to help provide an index of discriminability between old and new items and to lend credence to the possible "old" and "new" decisions on the test. Variability in this distribution exists primarily because of preexperimental familiarity (i.e., because of prior experience outside the experimental context) and because lures vary in their degree of overlap with the studied items. Evidence of prior experience in the experimental context for all stimuli is neither perfectly complete nor perfectly void; all of the distributions overlap to some extent, creating uncertainty for the decision maker. Therefore, one must establish a criterion on the familiarity scale as a dividing point for "old" and "new" decisions. The vertical lines in Fig. 1 represent different possible criterion settings. To the right of a criterion, "old" responses produce hits for targets and false alarms for lures.

Participants rarely make systematic shifts in their criterion from one item to the next within a recognition test. Consider the criterion labeled C_S in Fig. 1. If this were the only criterion, the nature of the target items as strong or weak would have little influence on the false alarm rate to lures, but of course, hit rates for strong targets would exceed those for weak targets. However, if a criterion shift occurred to the left for any reason—such as a shift from C_S to C_W—the false alarm rate would increase because the lure distribution is fixed.[1] This type of criterion difference is readily found when the strength of target items is pure across lists or across experiments (e.g., Criss, 2006, 2009, 2010; Hirshman, 1995; Ratcliff, Clark, & Shiffrin, 1990; Starns, Ratcliff, & White, 2012; Starns, White, & Ratcliff, 2010, 2013; Verde & Rotello, 2007, Experiments 2 vs. 4). In other words, when all of the targets on a test are strong, false alarm rates are typically lower than when all of the targets are weak. This type of pattern—where hits are highest to strong targets and false alarms are lowest following study of strong targets—is known as a strength-based mirror effect (Glanzer & Adams, 1985).

At issue in the present study is whether people can use cues to item strength as a means to consider lures differently *within* a test. Although the shift described above happens when strength is manipulated across lists, it does not happen readily when targets of differing strength are tested within the same list (e.g., Bruno, Higham, & Perfect, 2009, Experiment 1; Higham, Perfect, & Bruno, 2009, Experiment 2; Morrell, Gaitan, & Wixted, 2002; Singer, 2009, rote-study condition; Singer, Fazaluddin, & Andrew, 2011, 2013; Stretch & Wixted, 1998; Verde & Rotello, 2007, Experiments 1–4). People are typically reluctant to set different criteria among test items that differ in potential strength, although some experiments do show within-test shifts (e.g., Bruno et al., 2009, Experiments 2 and 3; Singer, 2009, semantic orienting task conditions; Singer & Wixted, 2006, Experiments 3 and 4). In the following sections, we review factors that might promote or disrupt within-test criterion shifts. We then detail how our own experiments will explore the relative importance of these factors. Our primary focus is on reports in which strength is manipulated via repetition, rather than some other factor such as delay (Singer & Wixted, 2006) or word frequency (e.g., Stretch & Wixted, 1998). We also contrast the criterion shift account with an account based on the differentiation of strong memory traces, and we discuss how our results will inform this theoretical debate.

Factors influencing within-list, strength-based criterion shifts

Within-list shifts have been investigated in a wide range of test formats, from large blocks of strong versus weak test trials with no explicit strength cue at test (Verde & Rotello, 2007) to randomly intermixed strong versus weak trials with cues to clearly mark strength (e.g., Stretch & Wixted, 1998). When explicit strength cues are not provided, Verde and Rotello demonstrated that participants use the first items that they encounter on a test to set their criterion value, but they are unlikely to adjust their criterion as test conditions change. That is, participants who did not receive feedback were more conservative overall when the test began with a strong block (Experiment 2) than when the test began with a weak block (Experiment 4), and this criterion difference persisted throughout the test. In contrast, Starns et al. (2010, Experiment 3) failed to find this difference; that is, participants applied the same criterion value on tests beginning with strong and weak blocks. A critical difference between the two studies that might explain the discrepant results is the size of the strength blocks. Strength alternated every 80 items on the test in Verde and Rotello versus every 24 items in Starns et al. (2010). In the present experiments, we varied both block size and the strength of the first test block to evaluate potential interactions between these two factors.

Verde and Rotello (2007) and Starns et al. (2010) also reported discrepant results regarding the effect of performance feedback. Verde and Rotello demonstrated across multiple experiments that, with no performance feedback, people did not shift criteria in strong versus weak test blocks. The inclusion of feedback in their fifth experiment apparently made people pay attention to criterion setting, motivating a strength-based shift from strong to weak blocks. However, Starns et al. (2010, Experiment 3) found that participants failed to shift their criterion across strength blocks both with and without feedback, again with no explicit strength cues provided. Again, the size of the strength blocks could be a critical factor

[1] In our case, we assumed a fixed lure distribution, because nothing in our experimental procedures was designed to systematically move the lure distribution around. However, it is reasonable to assume that the lure distribution changes systematically in other circumstances, such as those posited by differentiation models of recognition memory and in cases where the stimuli comprising lure distributions are manipulated (e.g., Criss, 2006).

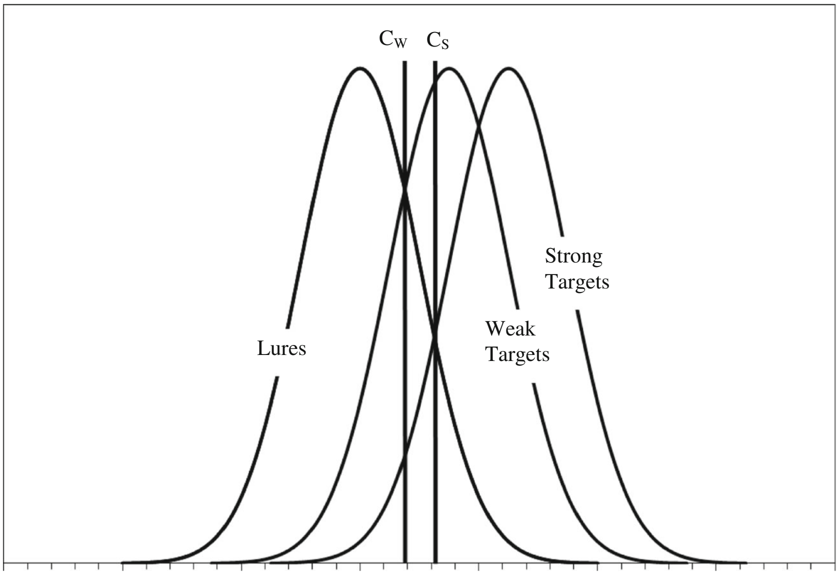

Fig. 1 Theoretical strength distributions for lures, weak targets, and strong targets. Vertical lines represent alternative possible criterion settings for weak test blocks and strong test blocks, respectively. A leftward shift from the strong criterion (C_S) to a weaker criterion (C_W) increases false alarms to lures

in explaining the different pattern of results in these two studies. Without strength marking, it is possible that more than 24 trials of a given strength block are needed for feedback to be effective in promoting shifts.

The Verde and Rotello (2007) and Starns et al. (2010) studies highlight the difficulty in motivating a criterion shift without some obvious cue to signal strength differences across test blocks. For this reason, some feature of the test stimuli is usually correlated with repetition. To list two examples, strength has been correlated with one category versus another (e.g., Bruno et al., 2009; Singer, 2009; Verde & Rotello, 2007, Experiment 3) and with one font color or another (e.g., Stretch & Wixted, 1998). Despite such marking, most of the studies investigating within-list criterion shifts have produced null results. The false alarm rates for lures associated with the strong versus the weak cues are remarkably similar. Notably, nearly all these studies used a random sequence of strong and weak test trials, challenging participants to quickly shift their criterion back and forth from one trial to the next. If criterion shifts require controlled, effortful processing, participants might simply decide that shifting the criterion dozens of times within a test is too taxing. If strength were cued *and* items were blocked as in the Verde and Rotello study, then this would require fewer shifts and, perhaps, produce more reliable differences in false alarms between strong and weak trials.

Study rationale

Our primary goal was to clarify the relative importance of the factors identified above—feedback, block size, and explicit strength cues—by manipulating them within the same set of experiments. We used font color of test items to cue strength at test, and we manipulated the presence/absence of color marking in the first two experiments. This direct comparison of strength cues being available versus unavailable and crossed with different block sizes within an experiment has not been recorded in the literature. We expected the absence of color marking to be a major obstacle to criterion shifting. In Experiments 1 and 2, we presented “strong” and “weak” test blocks that alternated every 10, 20, or 40 trials. As the test blocks vary from long to short, the number of times people would have to shift their criterion increases, and the amount of time they have to set and maintain a given criterion decreases. Thus, the block-size manipulation will allow us to explore how flexible participants can be in making criterion shifts. In Experiments 3 and 4, we present additional attempts at producing criterion shifts, including conditions where strength-cued test items were randomly presented (Experiment 3), as opposed to being blocked (Experiment 4). Random test contexts are the most prevalent in the literature, but they present the most difficult contexts in which to find shifts. To foreshadow, our results reinforce this message.

Finally, Experiments 1 and 2 were identical, except that performance feedback was provided in the second experiment. This allowed us to determine whether feedback facilitates within-test shifts even with no explicit strength cues and whether this effect is dependent on the size of the strength blocks (Starns et al., 2010; Verde & Rotello, 2007). For the participants not given strength cues or feedback, we expected that they would maintain a constant criterion across strength blocks, but they might use a different overall criterion value depending on whether the test begins with a strong block or a weak block. Our design allowed us to determine how large the initial block has to be in order to define participants' expectations across the entire test, thus potentially clarifying the different results of Verde and Rotello and of Starns et al. (2010).

Criterion shifts versus differentiation

Although strength-based changes in the false alarm rate have traditionally been interpreted as a criterion shift, recent recognition models can produce the effect via a process known as differentiation (e.g., Criss, 2006; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Differentiation means that strong memory traces produce a poorer match to lure items than do weak memory traces; thus, the memory strength for lures is lower when they are tested following a study list with well-learned versus poorly-learned words. Better differentiation shifts the lure distribution farther to the left on the strength continuum shown in Fig. 1, producing fewer false alarms. A number of recent experiments purportedly offered evidence against the differentiation mechanism by suggesting that the false alarm rate decline is driven by the expected strength at test, rather than the actual strength of the memory traces (Starns, Ratcliff, & White, 2012; Starns, White, & Ratcliff, 2010, 2012). The critical conditions in these experiments had mixed-strength study lists with half weak words (presented once) and half strong words (presented repeatedly), but the subsequent test contained either only the strong or only the weak words as targets. Participants were told which strength class would be tested before the test began. This manipulation provides participants the chance to change decision standards based on strength while holding the degree of differentiation from the studied traces constant. That is, the study list was identical across all conditions, so lure items were matched against memory traces with the same learning strength regardless of whether the strong or weak targets were tested. All of these experiments showed clear false alarm rate differences in the mixed conditions, which was claimed to be evidence that strength-based changes in the false alarm rate reflect criterion setting, not differentiation.

Some differentiation models might have a mechanism for accommodating effects in the mixed conditions, however. Although the degree of learning on the study list is equated between groups, learning also occurs during the recognition test (Criss, Malmberg, & Shiffrin, 2011; Malmberg, Criss, Gangwani, & Shiffrin, 2012). In the retrieving effectively from memory (REM) model, a new event is stored in an existing trace if the match between the event and the trace is sufficiently high, producing further differentiation of the trace. If the match is low, a new trace is created to store the new event (Shiffrin & Steyvers, 1997). Because the strong and weak tests have different target items, a test-based differentiation mechanism could potentially explain the false alarm rate difference. That is, each target on the strong test is likely to be stored in the well-learned trace established for this item at encoding, decreasing the match between this trace and any subsequently tested lure. Each target on the weak test is less likely to match its poorly encoded trace than a strong-test target; thus, a new trace will often be created to store the item. This not only precludes further differentiation of the study trace for that item, but also increases the effective list length by establishing extra traces. Both of these factors would contribute to a higher false alarm rate for subsequent lures than would be observed if the target was stored in its existing trace. Our design will permit a test for the criterion shift account independent of differentiation differences created by the test content. The strength-marked and unmarked conditions will have the exact same structure at both study and test, but participants will have more information to support criterion shifts in the former condition than in the latter. Therefore, any differences between the marked and unmarked conditions can be uniquely attributed to a criterion shift and not to differentiation.

Experiments 1 and 2

In these experiments, we explored how within-list criterion shifts are influenced by the availability of strength cues, changes in block size, and the presence of feedback. Participants completed alternating blocks of 10, 20, or 40 test trials with only strong or only weak targets in each block. Some participants experienced strong and weak blocks marked by different colors, whereas others were denied this explicit strength cue. The only difference between experiments was the inclusion of feedback during the recognition test for Experiment 2.

Method

Participants

Two hundred fifty-eight undergraduate students participated in Experiment 1. Between 40 and 46 people were randomly assigned to each of the six conditions in which block size was crossed with the presence or absence of color marking. In Experiment 2, 338 people participated. Between 50 and 60 were randomly assigned to the six conditions. All participants were recruited from undergraduate psychology courses at Louisiana State University (LSU). They participated either to earn credit for a research learning course requirement or to receive extra credit.

Materials

One hundred words were selected from the MRC psycholinguistic database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm). All words were between five and eight letters in length and had familiarity ratings between 500 and 700 and concreteness ratings between 300 and 500 on scales ranging from 100 to 700. A program written in E-Prime 1.2 (Psychology Software Tools, Pittsburgh, PA) controlled stimulus presentation and recorded keyboard responses.

Design and procedure

The design was a 2 (strength) × 3 (block size) × 2 (color marking) mixed factorial, with block size and color marking manipulated between subjects and item strength manipulated within subjects. For each participant, the computer software randomly selected 40 of the stimuli to serve as encoded targets and 40 others as lures. The remaining 20 items were designated as fillers and were shown as the first 10 and last 10 items in the encoding phase but were not tested. Of the 40 targets, 20 were randomly assigned to the strong (presented 4 times) condition and the others were assigned to the weak (presented once) condition. Encoded items were displayed one at a time in black font on a light gray screen background for 700 ms of stimulus presentation and were followed by a blank 100-ms interstimulus interval (ISI). The encoding sequence (120 total presentations) was randomized. Participants were instructed to learn the encoded items for an upcoming memory test and that some of the items would be repeated. Upon the presentation of the final encoded filler item, the test instructions were displayed.

For the marked conditions, strong targets were presented in a red font color at test, and weak targets were presented in a green font color. Of the 40 lures, 20 were randomly assigned to be presented in the "strong" test color (red), and the other 20 were assigned to be presented in the "weak" test color (green). For the unmarked conditions, all test words appeared in black font. The structure of the test list depended on block size condition. Test items were grouped in strength blocks of 10, 20, or 40, with strong and weak blocks alternating. Whether a strong block occurred first or second in the test sequence was counterbalanced across participants, and there was an equal number of strong-first and weak-first participants within each condition. The order of the items within each block was randomized, and each block contained an equal number of targets and lures. Participants used the "z" and "/" keys for "new" and "old" decisions, respectively. Participants in the marked condition were instructed as to the significance of the color cues. Specifically, we instructed participants that when a test cue was shown in red, this color indicated that the item either was studied 4 times or was new. When a test item was shown in green, it indicated that the item either was shown once or was new. We instructed that decisions should be made as quickly and accurately as possible. Participants were not reminded of these instructions once the test phase began. Test items were shown one at a time and remained on the screen until a response was made.

For Experiment 2, feedback was provided during the test phase. Whenever people made an error (i.e., missed a target or committed a false alarm), the software produced a message that their decision was incorrect (***ERROR***).

Results

For ANOVA models, partial-eta squared (η_p^2) is reported as an estimate of effect size. For pairwise comparisons, Cohen's d is offered as the effect size measure. The type I error rate (α) for post hoc comparisons was adjusted with a Bonferroni correction.

Experiment 1 (no feedback)

Table 1 displays hit rates, false alarm rates, and discriminability for the various conditions. We focus primarily on false alarm rates because they reveal criterion shifting most directly, but we also analyze recognition discriminability. For false alarms, the 2 (strength) × 3 (block size) × 2 (color marking) mixed factorial ANOVA produced only two significant effects, and these involved the strength and color-marking factors. False alarms were generally lower for the lures tested in strong test blocks, $F(1, 252) = 6.36$, $p < .05$, $MSE = .014$, $\eta_p^2 = .025$, but this was qualified by a strength × marking interaction, $F(1, 252) = 9.13$, $p < .01$, $MSE = .014$, $\eta_p^2 = .035$. Figure 2 makes clear that the strength difference was apparent only when test cues were marked by color ($M_{\text{strong}} = .25$, $M_{\text{weak}} = .31$), $t(129) = 3.86$, $p < .01$, $d = 0.34$ (Bonferroni-corrected $\alpha = .025$). When strength was not marked, the means were practically identical ($M_{\text{strong}} = .29$, $M_{\text{weak}} = .29$), $t(127) = 0.27$, $p > .50$.

For recognition discriminability in the last two columns of Table 1, only two main effects emerged.[2] Strength was obviously significant, $F(1, 252) = 335.71$, $p < .01$, $MSE = .30$, $\eta_p^2 = .571$. Interestingly, the marking factor was also significant. Color-marked conditions ($M = 1.50$) produced worse discriminability than did unmarked conditions ($M = 1.64$), $F(1, 252) = 6.25$, $p < .05$, $MSE = .46$, $\eta_p^2 = .024$. No other effects for discriminability were significant, $p > .20$.

[2] We also computed discriminability under an assumption that the target distribution is roughly 25% more variable than the lure distribution (i.e., [1.25 * z(HR) − z(FAR)]). The pattern of factors that were significant versus not in Experiments 1 and 2 did not change.

Table 1 Hit and false alarm rates and recognition discriminability (d') in weak and strong test blocks with either color-marked or unmarked strength in Experiment 1

Marking and Block Size	Hit Rate		False Alarm Rate		d'	
	Weak	Strong	Weak	Strong	Weak	Strong
Color-Marked						
10 Items	.65 (.03)	.85 (.02)	.29 (.02)	.25 (.02)	1.02 (.09)	1.93 (.09)
20 Items	.64 (.03)	.85 (.02)	.29 (.02)	.23 (.02)	1.03 (.10)	2.01 (.11)
40 Items	.70 (.02)	.85 (.02)	.33 (.02)	.26 (.03)	1.07 (.09)	1.91 (.10)
Unmarked						
10 Items	.71 (.03)	.93 (.01)	.28 (.03)	.31 (.02)	1.29 (.09)	2.07 (.07)
20 Items	.68 (.03)	.90 (.02)	.31 (.02)	.32 (.03)	1.08 (.08)	1.92 (.11)
40 Items	.70 (.02)	.92 (.01)	.28 (.02)	.25 (.02)	1.26 (.09)	2.24 (.09)

Note. Standard errors of the mean in parentheses, $d' = z(\text{hit rate}) - z(\text{false alarm rate})$.

Experiment 2 (feedback)

These data are reported in Table 2. False alarms were generally lower for the lures in strong test blocks, $F(1, 332) = 23.40$, $p < .01$, $MSE = .013$, $\eta_p^2 = .066$, but this was qualified by a strength × marking interaction, $F(1, 332) = 8.78$, $p < .01$, $MSE = .014$, $\eta_p^2 = .026$. Figure 3 shows that the false alarm difference was apparent only when test cues were marked by color, ($M_{\text{strong}} = .29$, $M_{\text{weak}} = .36$), $t(165) = 5.63$, $p < .01$, $d = 0.44$ (Bonferroni-corrected $\alpha = .025$). When strength was not marked, the means were not different ($M_{\text{strong}} = .29$, $M_{\text{weak}} = .31$), $t(171) = 1.30$, $p = .19$, $d = 0.13$.

A couple of other outcomes emerged that were not present in Experiment 1. False alarms increased overall as block size increased, $F(2, 332) = 3.66$, $p < .05$, $MSE = .037$, $\eta_p^2 = .022$. The 10-item block produced a mean false alarm rate of .28, the 20-item block a rate of .32, and the 40-item block a rate of .33. The 10- and 40-item blocks were significantly different by a Bonferroni-corrected post hoc comparison ($\alpha = .017$), $t(332) = 2.56$, $p = .011$, $d = 0.30$. This false alarm increase across block size was not found in Experiment 1, so whether it represents a reliable effect related to the provision of feedback is unclear. There was also a main effect of marking near conventional significance, $F(1, 332) = 3.75$, $p = .054$, $\eta_p^2 = .011$. Clearly, this is a very small effect size, but it demonstrates a trend toward more false alarms produced overall when color marking was used.

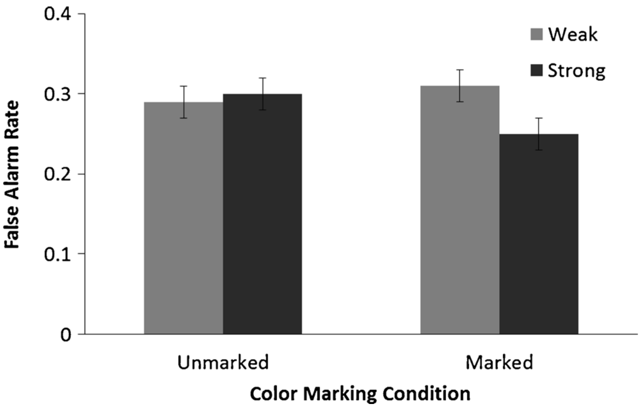

Fig. 2 False alarm rates in Experiment 1 for weak and strong test blocks. Strength was marked by color during the recognition test, or it was unmarked. Weak lures were those tested in blocks with weak targets. Strong lures were those tested in blocks with strong (repeated) targets. Error bars represent 95% within-subjects confidence intervals recommended for mixed factorial designs (Masson & Loftus, 2003)

The outcomes for recognition discriminability were very similar to those in Experiment 1. Stronger items were better recognized, $F(1, 332) = 685.62$, $p < .01$, $MSE = .24$, $\eta_p^2 = .674$. Strength marking was nearly significant: Color-marked conditions ($M = 1.38$) produced worse discriminability than did unmarked conditions ($M = 1.49$), $F(1, 332) = 3.74$, $p = .054$, $MSE = .54$, $\eta_p^2 = .011$. No other potential effects for discriminability were significant, $p > .20$.

Analyses including feedback

The patterns in Experiments 1 and 2 suggest that feedback had little influence on the primary criterion shift outcomes but did have some effect on the false alarm rate in general. To explore this possibility more formally, we performed a 2 (feedback) × 2 (strength) × 3 (block size) × 2 (color marking) ANOVA on the joint data from both experiments. Presence or absence of feedback did not produce significant interactions with strength or block size and, thus, did not influence within-list criterion shifts. Feedback increased false alarms overall ($M = .31$), as compared with no feedback ($M = .28$), $F(1, 584) = 6.13$, $p < .05$, $MSE = .036$, $\eta_p^2 = .01$. More important, this small effect of feedback depended on the color marking factor, $F(1, 584) = 4.02$, $p < .05$, $MSE = .036$, $\eta_p^2 = .007$. Feedback produced higher false alarms ($M = .33$), as opposed to no feedback ($M = .28$), when color marking was present, $t(294) = 3.11$, $p < .01$, $d = 0.37$ (Bonferroni-corrected $\alpha = .025$). This was not true when

Table 2 Hit and false alarm rates and recognition discriminability (d') in weak and strong test blocks with either color-marked or unmarked strength in Experiment 2 (feedback for incorrect test decisions provided)

Marking and	Hit Rate		False Alarm Rate		d'	
Block Size	Weak	Strong	Weak	Strong	Weak	Strong
Color-Marked						
10 Items	.68 (.02)	.88 (.01)	.38 (.02)	.32 (.02)	0.86 (.09)	1.80 (.10)
20 Items	.69 (.01)	.89 (.01)	.38 (.02)	.30 (.02)	0.86 (.07)	1.87 (.10)
40 Items	.65 (.02)	.87 (.01)	.32 (.02)	.25 (.02)	0.93 (.06)	1.99 (.09)
Unmarked						
10 Items	.64 (.02)	.90 (.01)	.33 (.02)	.30 (.02)	0.91 (.08)	1.97 (.09)
20 Items	.67 (.01)	.89 (.01)	.30 (.02)	.30 (.02)	1.07 (.08)	1.92 (.08)
40 Items	.65 (.02)	.89 (.01)	.29 (.02)	.27 (.02)	1.04 (.08)	2.05 (.10)

Note. Standard errors of the mean in parentheses, $d' = z(\text{hit rate}) - z(\text{false alarm rate})$.

marking was absent ($M = .30$ with feedback, $M = .29$ without), $t(298) = 0.43$, $p > .50$.

Regarding recognition discriminability, the analysis pooled over Experiments 1 and 2 demonstrated lower discriminability when color marking was present ($M = 1.44$) versus absent ($M = 1.57$), $F(1, 584) = 9.67$, $MSE = .504$, $p < .01$, $\eta_p^2 = .016$. This confirms the results presented earlier within each experiment. In addition, discriminability suffered when feedback was present ($M = 1.44$) versus absent ($M = 1.57$), $F(1, 584) = 10.05$, $MSE = .504$, $p < .01$, $\eta_p^2 = .017$. Thus, strength marking and feedback independently lowered recognition discriminability.

Influence of the initial block in unmarked tests

Verde and Rotello (2007) found that participants set their overall criterion on the basis of the strength of the first test block even when they failed to adjust their criterion across the strong and weak blocks within the same test. Participants in our analogous unmarked conditions might have displayed a similar pattern. To investigate this issue, we compared the overall false alarm rate (across the entire test) for participants who began the test with a strong versus a weak block. We again analyzed unmarked conditions from Experiments 1 and 2 jointly, resulting in a 2 (first-block strength) × 2 (feedback) × 3 (block size) ANOVA. The only significant factors were the main effect of first-block strength, $F(1, 288) = 3.99$, $p < .05$, $MSE = .018$, $\eta_p^2 = .014$, and the interaction of first-block strength with block size, $F(2, 288) = 6.80$, $p = .001$, $\eta_p^2 = .045$. In the 40-trial block conditions, the overall false alarm rate was lower when the first block was strong ($M = .22$) versus weak ($M = .33$), $t(100) = 4.99$, $p < .01$, $d = 1.10$ (Bonferroni-corrected $\alpha = .017$). In contrast, first-block strength had no discernible effect on either 20-trial blocks (.30 vs. .31), $t(96) = 0.55$, $p > .20$, or 10-trial blocks (.32 vs. .29), $t(98) = 0.89$, $p > .20$.

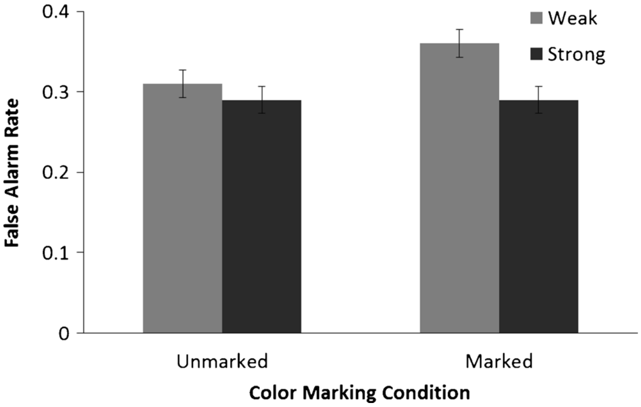

Fig. 3 False alarm rates in Experiment 2 for weak and strong test blocks. Strength was marked by color during the recognition test, or it was unmarked. Feedback was provided during the recognition test for incorrect decisions. Weak lures were those tested in blocks with weak targets. Strong lures were those tested in blocks with strong (repeated) targets. Error bars represent 95% within-subject confidence intervals recommended for mixed factorial designs (Masson & Loftus, 2003)

Discussion

In two experiments, people produced fewer false alarms to lures tested in blocks of strong targets, as opposed to lures tested in blocks of weak targets. Critically, color marking of the strength blocks was required to produce this criterion shift, since no shift occurred when the marking was unavailable. Although participants in the unmarked condition did not change their criterion systematically within a test, they did use a more conservative criterion in the 40-item block condition when the first block was strong versus weak. The 10- and 20-item blocks showed no evidence of this difference. This interaction helps to explain why Verde and Rotello (2007) found an effect of first-block strength with 80-item blocks, whereas Starns et al. (2010) found no effect with 24-item blocks. Feedback in Experiment 2 did not influence the criterion shift results. Thus, unlike Verde and Rotello's (2007, Experiment 5) feedback-induced criterion shift with 80-item blocks of strong versus weak items, feedback did not produce a shift in our unmarked blocks of 10, 20, or 40 items. These results replicate the lack of a within-test shift for 24-item strength blocks reported by Starns et al. (2010) and show that this null result holds in blocks as large as 40 items. Feedback did slightly increase the overall false alarm rate when color marking was present. Also, recognition discriminability was slightly hampered by the presence of color marking in both experiments.

Experiments 3 and 4

One implication of our results is that marking the test items with a strength cue is critical to finding a within-subjects criterion shift, although not always sufficient (e.g., Stretch & Wixted, 1998). Arranging the test items into color-cued blocks was likely important, but it is difficult to say whether such blocking is critical. Therefore, we present five additional experiments that replicate various blocked conditions from Experiments 1 and 2 but also examine whether randomly presented test items can support a criterion shift. Strength cues were provided in all conditions. Experiments 3A–3C represent three independent attempts at using random test presentation. These random conditions essentially replicate Experiments 4 and 5 of Stretch and Wixted's (1998) report. In Experiment 4A, we replicated the 10-, 20-, and 40-item block procedure of Experiment 1. In Experiment 4B, we replicated the 40-item block procedure once again. Blocking test items should facilitate criterion shifting, whereas randomly switching between items should be more taxing. For ease of presentation, Experiments 3 and 4 are presented together.

Method

Participants

Forty-one participants were tested in the random test condition of Experiment 3A, 45 in 3B, and 29 in 3C. Thirty-eight were tested in the 40-item condition of Experiment 4A, 38 in the 20-item condition of 4A, and 40 in the 10-item condition of 4A. Finally, 50 were tested in the 40-item condition of Experiment 4B. Participants in Experiment 3C were recruited from the University of Massachusetts Amherst; all others were recruited from LSU as in Experiments 1 and 2.

Materials and procedure

With only one exception, the procedures were exactly as described earlier. The lone difference in the software was that test items were color-marked but presented in a random order in Experiments 3A–3C. Experiment 4A was an exact replication of the color-marked, blocked test conditions of Experiment 1 (10-, 20-, and 40-item blocks). Experiment 4B was an exact replication of the 40-item blocked test condition of Experiment 1. No feedback was used in any of these experiments.

Results and discussion

Table 3 presents the relevant recognition decision data for each experiment. A criterion shift was replicated in the blocked conditions of Experiments 4A and 4B but was inconsistent in Experiments 3A through 3C. A paired samples *t* test on false alarm rates for lures presented with the strong versus weak test cue was not significant for Experiment 3A, $t(40) = 0.91$, $p > .05$, $d = 0.14$, or Experiment 3C, $t(28) = 1.92$, $p > .05$, $d = 0.36$, but was significant for 3B, $t(44) = 4.08$, $p < .05$, $d = 0.60$. For Experiment 4A, a 2 (strength) × 3 (block size) mixed factorial ANOVA produced only a significant main effect of strength, since false alarms were generally lower for the lures tested in strong test blocks, $F(1, 113) = 11.95$, $p < .05$, $MSE = .014$, $\eta_p^2 = .096$. The strength × block interaction was not significant, $F(2, 113) = 0.91$, $p > .10$, $MSE = .014$, $\eta_p^2 = .003$. For Experiment 4B, the paired samples *t* test for the false alarm difference was significant, $t(49) = 3.82$, $p < .05$, $d = 0.54$.

Table 3 Hit and false alarm rates and recognition discriminability (d') in weak and strong test blocks with color-marked strength in Experiments 3 and 4

Block Size and Experiment	Hit Rate		False Alarm Rate		d'	
	Weak	Strong	Weak	Strong	Weak	Strong
Random (3A)	.65 (.02)	.88 (.02)	.31 (.03)	.28 (.03)	1.04 (.08)	1.98 (.12)
Random (3B)	.64 (.02)	.84 (.02)	.33 (.02)	.25 (.02)	0.89 (.07)	1.91 (.11)
Random (3C)	.59 (.02)	.84 (.02)	.17 (.02)	.14 (.02)	1.29 (.10)	2.48 (.12)
10 Items (4A)	.67 (.02)	.88 (.02)	.32 (.02)	.28 (.02)	1.01 (.07)	1.93 (.07)
20 Items (4A)	.69 (.02)	.87 (.02)	.29 (.02)	.24 (.03)	1.13 (.08)	2.10 (.12)
40 Items (4A)	.71 (.02)	.86 (.02)	.33 (.02)	.27 (.02)	1.08 (.09)	1.96 (.12)
40 Items (4B)	.64 (.03)	.84 (.02)	.33 (.01)	.19 (.02)	0.93 (.09)	2.09 (.09)

Note. Standard errors of the mean in parentheses, $d' = z(\text{hit rate}) - z(\text{false alarm rate})$.

The results of the blocked conditions of Experiments 4A and 4B replicated the color-marked conditions of Experiment 1. These blocked conditions showed a reliable within-subjects criterion shift, with a slightly decreasing effect as block size decreased. Of the random test conditions in Experiments 3A–3C, one was significant and two were not. The results of the random conditions are not conclusive in and of themselves, especially because the lone significant result produced an effect size commensurate with those of the blocked conditions. However, when these random condition results are compared with similar reports in the literature, it becomes clear that random presentation of strength-marked items is not a reliable method of producing within-subjects criterion shifts. The most similar procedure to ours is that of Stretch and Wixted (1998), and their "strong" and "weak" false alarm rates differed by no more than 2% in their Experiments 4 and 5. Moreover, Olchowski and Starns (2012) presented two random test conditions very similar to ours ("E1B/2-key" and "E2/2-key" conditions) where criterion shift effects were no larger than 2%. These results are summarized in Table 4 and will be discussed in more detail in the General Discussion section. Given these other relevant results, the random procedure appears to produce very small shift effects that are usually not statistically detectable. In contrast, all of our blocked experiments (with strength marking) showed significant criterion shift effects.

Table 4 Summary of within-subjects, strength-based criterion shifts from the present and related studies where strength was defined by repetition and cued with color at retrieval

Block Condition	Source	Shift	95% CI	*N*
40 Items	H&S (Exp. 1)	.075	[.018, .132]	44
	H&S (Exp. 2)	.073	[.030, .115]	60
	H&S (Exp. 4A)	.065	[.003, .126]	38
	H&S (Exp. 4B)	.130	[.062, .199]	50
	Average	**.086**		
20 Items	H&S (Exp. 1)	.058	[.013, .103]	44
	H&S (Exp. 2)	.076	[.032, .119]	56
	H&S (Exp. 4A)	.055	[.010, .100]	38
	Average	**.063**		
10 Items	H&S (Exp. 1)	.043	[−.015, .100]	42
	H&S (Exp. 2)	.055	[.015, .095]	50
	H&S (Exp. 4A)	.041	[−.015, .097]	40
	Average	**.046**		
Random	H&S (Exp. 3A)	.028	[−.035, .091]	41
	H&S (Exp. 3B)	.085	[.043, .128]	45
	H&S (Exp. 3C)	.030	[−.002, .061]	29
	S&W (1998, Exp. 4)	−.015	[−.051, .021]	36
	S&W (1998, Exp. 5)	.019	[−.022, .060]	31
	O&S (2012, Exp. 1B)	−.016	[−.061, .028]	28
	O&S (2012, Exp. 2)	.016	[−.033, .065]	23
	Average	**.021**		

Note H&S refers to the present study (Hicks & Starns); S&W (1998) refers to the Stretch and Wixted (1998) "uncorrelated mixed-strength" conditions collapsed across word frequency; O&S (2012) refers to the Olchowski and Starns (2012) "2-key" conditions. "Shift" is the strong false alarm rate subtracted from the weak false alarm rate. We report only the past studies with designs most similar to our own—that is, with strength marked only by color at test. Experiment 2 of the present study (H&S, Exp. 2) included feedback for recognition decision errors.

General discussion

The following summarizes the key points regarding our investigation. First, providing explicit strength cues facilitates within-subjects strength-based criterion shifts. Experiments 1 and 2 showed that strength marking was crucial by directly comparing marked versus unmarked conditions within an experiment. Second, presenting strength-cued test items in larger blocks produces a more reliable method of obtaining criterion shifts. This conclusion is drawn from the entire pattern of results across Experiments 1 through 4, in tandem with a consideration of the most similar procedures represented in the literature (i.e., Olchowski & Starns, 2012; Stretch & Wixted, 1998). Random presentation of strength-cued test items often fails to produce a significant change in false alarm rate, whereas all of our experiments with blocked lists showed a significant effect. This suggests that forcing participants to continually adjust their criterion throughout a test disrupts their ability to apply separate criteria for strong and weak cues. Third, when items were presented in blocks and strength was unmarked, the strength of the initial test block influenced the overall false alarm rate only when the block was sufficiently large. In the 40-item blocks of Experiments 1 and 2, but not in the 20- or 10-item blocks, beginning with a strong block established a conservative criterion that was maintained throughout the test. Fourth, feedback alone does not consistently produce strength-based criterion shifts in the absence of strength cues at test (cf. Verde & Rotello, 2007). This was the case when the results of Experiment 1 without feedback were compared with those of Experiment 2 with feedback. Feedback also did not alter the size of criterion shifts in the strength-cued test conditions.

Considering the results for unmarked strength blocks, our results strengthen some conclusions from the previous literature and help to address some (but not all) of the discrepancies. Verde and Rotello (2007) found that participants applied a more conservative criterion throughout the test when the first block contained strong, as opposed to weak, targets, but Starns et al. (2010) found no such difference in their Experiment 4. Our results support the contention that the critical difference between the two studies is the size of the strength blocks (80 vs. 24 items per block). Our 40-item block participants maintained a more conservative criterion overall when they began the test with a strong block than with a weak block, replicating Verde and Rotello. However, the 20- and 10-item blocks showed no hint of this difference, replicating Starns et al. (2010). This suggests that the initial criterion-setting process takes some time. That is, participants do not stabilize their criterion in the first 10 or 20 trials, but getting a consistent and high expectation of strength for 40 trials produces a criterion difference that persists throughout the test.

Both Verde and Rotello (2007) and Starns et al. (2010) found that participants did not change their criterion between test blocks with weak versus strong targets when no feedback was provided. Our results replicate this finding. However, the two past experiments differed in that introducing feedback produced a difference between blocks in Verde and Rotello, but not in Starns et al. (2010). Our results do not completely resolve this discrepancy, but they do replicate the findings of Starns et al. (2010) and show that the null interaction effect between feedback and block strength holds with blocks of up to 40 items each. Thus, although feedback might sometimes facilitate strength-based criterion shifts without explicit strength cues, participants often fail to make shifts even when feedback is available.

As has been noted by others (e.g., Singer, 2009), the difficulty in producing within-subjects, within-list, strength-based criterion shifts stands in stark contrast to investigations showing easily induced shifts either between lists or between subjects/experiments (e.g., Hirshman, 1995; Starns, Hicks, & Marsh, 2006; Stretch & Wixted, 1998; Verde & Rotello, 2007). When people establish a criterion based on encoding and test characteristics, it becomes difficult to manipulate within the test, especially when strength is marked with arbitrary cues and not with intrinsic characteristics of the stimuli, such as category or stimulus type (e.g., Stretch & Wixted, 1998). Our results demonstrate that organizing the test into blocks facilitates within-test criterion shifts in response to arbitrary strength cues. We found significant false alarm rate differences in all of our color-marked conditions with 20- and 40-item blocks, which contrasts with reports showing null effects when strength cues switch randomly from one item to the next. Organizing the test into blocks reduces the number of criterion shifts required and allows participants to maintain a given criterion value for an extended period of time. Brown and Steyvers (2005) reported evidence that participants need around 14 trials to adjust to a new decision environment in perceptual and lexical decision tasks. Careful readers will note that the 95% confidence intervals for our 10-item block conditions contain a value of zero in two out of three cases (see Table 4). Thus, our block size results are quite consistent with the Brown and Steyvers work, albeit with strength-marked cues to facilitate shifting.

Comparison with previous studies

In our study, repetition of information was the only procedure used to signify strength differences at encoding. At test, the only strength cue, when made available, was color to mark the potential strength differences. A review of the published attempts to produce such criterion shifts shows that encoding strength is typically correlated with some other type of stimulus attribute—most often, with category membership (e.g., Bruno et al., 2009, Experiments 1 and 3; Morrell et al., 2002; Singer, 2009; Singer et al., 2013). We regard this as a potentially important difference, so we exclude these studies from the following comparison of our results with the existing literature. Stretch and Wixted (1998) and Olchowski and Starns (2012) used random test procedures very similar to our own. These studies incorporated color cues at retrieval to cue strength differences in target items. Stretch and Wixted also presented items at encoding in different colors to signify strength, but the procedures in their fourth and fifth experiments are otherwise close enough to permit a comparison with our studies. In Table 4, we present a summary of the criterion shift produced in these various experiments and conditions, including the relevant conditions of the present study. The mean shift was calculated by subtracting the "strong" false alarm rate from the "weak" false alarm rate in each condition, with positive values representing a strength-based shift. Ninety-five percent confidence intervals around the mean shifts are also presented. The pattern, as one scans the table from top to bottom, is average decreases in criterion shifting when test items are randomly presented, as opposed to blocked. Of the seven random conditions presented, only one was significant by traditional standards ($\alpha = .05$). All four 40-item conditions were significant, all three 20-item conditions were significant, and one of the three 10-item conditions was significant. The pattern across this table suggests two important conclusions. First, the effect of strength-marking is more reliable as block size increases. Second, criterion shifting in the random conditions is rarely successful, indicating that this effect size is very small.

Randomized test conditions and the global subjective memorability hypothesis

Bruno et al. (2009) offered the *global subjective memorability* (GSM) hypothesis to explain when conditions would be most favorable for within-list criterion shifting. They tested a variety of conditions in which strength was manipulated via repetition and test items were randomly intermixed. Across a series of experiments, they demonstrated that when the likely perceived memorability of the entire encoded set was low (i.e., low GSM), a criterion shift occurred. For example, when nonwords were intermixed with words in the encoding list, a criterion shift was present when 80 of the 120 stimuli were nonwords (Experiment 2), but not when 20 of the 92 stimuli were nonwords (Experiment 1). In their third experiment with words only, a criterion shift occurred when the words were encoded for 0.5 s each, but not when they were encoded for 3 s each. There was a 1-s ISI, which meant that the respective effective encoding times were roughly 1.5 and 4 s in each condition. The second experiment from Higham et al. (2009), also with 3-s encoding and otherwise similar to Bruno et al.'s third experiment, showed no criterion shift based on expected strength. Bruno et al. interpreted the pattern across these experiments to indicate that when overall subjective memorability of the stimuli is low, a criterion shift is more likely to appear.

However, the broader criterion shift literature provides little support for the GSM hypothesis. This account predicts that short encoding conditions should promote low GSM in general. We used 0.7-s encoding times with a very short 0.1-s ISI, making the effective encoding time about 0.8 s. This overall encoding time is even less than Bruno et al.'s (2009) short encoding time (0.5 s + 1-s ISI). Stretch and Wixted's (1998) relevant conditions also used a very short encoding time (0.5 s + 0.25-s ISI). Yet in Table 4, the seven random test conditions listed demonstrated only one significant shift. Furthermore, other work using repetition-induced strength apparently runs counter to the GSM prediction. Singer (2009) had people encode members of some categories twice and others once (i.e., strength-correlated category membership). People produced criterion shifts across category lures only when they made pleasantness ratings for words during encoding, but not following rote encoding instructions. Clearly the latter condition should have promoted lower relative GSM. In a follow-up study, Singer et al. (2013) manipulated encoding time very similarly to Bruno et al.'s third experiment, again using a rote encoding instruction. In conditions ranging from 0.75- to 3-s encoding times in two experiments, Singer et al. (2013) reported no strength-based criterion shifts (see also Singer et al., 2011, Experiment 1a). In fact, *opposite* shifts occurred in which the false alarm rate was higher for "strong" category lures.

Criterion shifts versus differentiation

Our results strengthen support for the criterion-shift explanation of the strength-based mirror effect and demonstrate that criterion shifts can produce the mirror effect without any contribution from differentiation. Previous results (Starns, White, & Ratcliff, 2010, 2012, 2013) demonstrated mirror effects with no differentiation difference in the memory traces established at encoding, but these experiments did not control for potential differentiation differences produced by learning at test. The present results with 20- and 40-item blocks provide evidence for criterion shifts with differentiation controlled at study and test: Participants with strength cues showed a strength-based mirror effect with the exact same study and test lists as the uncued participants who showed no false alarm difference. These results do not rule out the possibility that differentiation plays some role in recognition memory, but they do show that the strength-based mirror effect does not rely on differentiation and cannot be interpreted per se as evidence for this process. Differentiation models also have parameters for the response criterion, so they can accommodate false alarm rate changes produced by a criterion shift (Criss, 2006; McClelland & Chappell, 1998).

Criterion shifts in pure- versus mixed-list designs

The results of our experiments and our analysis of prior work using within-list manipulations of strength apparently suggest that people do not bother shifting criteria in response to expected strength differences unless given strong motivation to do so. Without explicit strength cues and blocked tests, they fail to show strength effects on false alarm rates even though strength has a large effect on hit rates. From this perspective, criterion shifts as explanatory mechanisms for findings such as the mirror effect might seem difficult to defend. However, a different perspective is that people are usually quite willing to use expected strength as a cue, but they are usually unable to notice changes in strength at test if these differences are not explicitly signaled, and they become unwilling to use expected strength when test items appear randomly with regard to the strength variable. Importantly, pure lists (e.g., Hirshman, 1995) have both of the characteristics we suggest are needed to motivate criterion shifts: explicit knowledge of the expected strength at test and stable strength expectations across the entire test. Thus, our results suggest that criterion shifts should play an influential role in producing the strength-based mirror effect in pure-list designs, although isolating the unique influence of shifts, as opposed to processes like differentiation, is very difficult with pure lists.

Qualitative versus quantitative shifts

Our findings also qualify some recent claims that people depend on expected retrieval of more *qualitative* memorial details, rather than *quantitative* information, to reduce false recognition. For example, Scimeca, McDonough, and Gallo (2011) showed that quantitative differences in strength induced by repetition—similar to what we tested—did not encourage reductions in false alarms for separate tests in which people expected either strong or weak targets. In contrast, they showed that people could show such reductions when expecting test blocks that included targets encoded as pictures, as opposed to words (see also Gallo, Weiss, & Schacter, 2004), or with deeper processing, as opposed to shallow processing (Gallo, Meadow, Johnson, & Foster, 2008). Scimeca et al. concluded that reductions in false recognition are more easily induced by expectation of *qualitative* differences that predict more memorable, versus less memorable, targets. Shifts based on qualitative expectations may be easier to produce, but our results show that shifts based on quantitative differences are also possible (also see Starns, White, & Ratcliff, 2010, 2012).

Individual differences

The criterion shifts produced in this report clearly relied on people's metacognitive control and willingness to use diagnostic information in the test cue. One potentially fruitful avenue of investigation will be to study individual differences in cognitive control. For example, within-subjects criterion shifts induced by a target probability manipulation have been reported as highly consistent within individuals across separate tests (Aminoff et al., 2012). Whether or not similar reliability across individuals occurs for strength-based shifting, as in our first two experiments, remains an open question. Working memory capacity may be a predictor of the presence and/or magnitude of within-list criterion shifts, because information must be kept in mind while applying expectations for memory strength in a given block of items. Keeping these conditions in awareness in a rapidly shifting test environment would be theoretically taxing. It is worth noting again that false alarms increased in color-marked test blocks of Experiment 2 and that recognition discriminability was lower in marked test blocks of Experiments 1 and 2, suggesting that using color cues made the test more challenging overall. This unexpected result was one benefit of directly comparing marked and unmarked conditions. Inducing criterion shifts between weak and strong blocks may also increase criterion noise, thereby impairing performance overall (Benjamin, Diaz, & Wee, 2009). The unexpected lower discriminability with feedback, as opposed to no feedback, and higher false alarm rates overall may suggest that feedback also produces some criterion noise. This may be especially true when feedback is considered along with other instructions intended to guide/adjust strength-based criterion setting. However, our results should not be viewed as strong evidence that feedback impairs performance, given that our effect was small and many null effects of feedback have been reported (e.g., Kantner & Lindsay, 2010).

Conclusion

In closing, we emphasize that our study demonstrates a reliable within-list, strength-based criterion shift in conditions where arbitrary test cues signaled differences in strength and were presented in sufficiently large blocks. Memory strength—a quantitative variable—can be used effectively as a cue for changes in retrieval standards. The key appears to be giving people a reliable source of information to guide their expectations for memory strength and allowing them to maintain the same standards over a number of trials, instead of switching randomly. Our results also bolster the claim that people are often unwilling to systematically change their criterion on a trial-by-trial basis, which is problematic for theories that rely on item-based criterion adjustments to explain phenomena such as word frequency effects (e.g., Brown, Lewis, & Monk, 1977) and false memory (e.g., Miller & Wolford, 1999).

Acknowledgments Portions of these experiments were reported at the 52nd Annual Meeting of the Psychonomic Society, Seattle, Washington, November 2011. We wish to thank Alexandra Adams, Datonye Charles, Stephenie Holzhalb, Keyahna Hudson, and Tiffany Scott for their dedicated help in collecting the data. We are grateful for helpful comments from John Wixted and two anonymous reviewers in the publication process.

References

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., … Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition, 40*, 1016–1030.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115.

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.

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition, 37,* 807–818.

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

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

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

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

Gallo, D. A., Meadow, N. G., Johnson, E. L., & Foster, K. T. (2008). Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language, 58,* 1095–1111.

Gallo, D. A., Weiss, J. A., & Schacter, D. L. (2004). Reducing false recognition with criterial recollection tests: Distinctiveness heuristic versus criterion shifts. *Journal of Memory and Language, 51,* 473–493.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using Type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 57–80.

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

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389–406.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23,* 115–119.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

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.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106,* 398–405.

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.

Olchowski, J. E, & Starns, J. J. (2012, November). *Item-by-item criterion shifts in recognition memory: All you need is an extra button*. Poster session presented at the 53rd Annual Meeting of the Psychonomic Society, Minneapolis, MN.

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.

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

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

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

Singer, M., Fazaluddin, A., & Andrew, K. N. (2011). Distinctiveness and repetition in item recognition. *Canadian Journal of Experimental Psychology, 65,* 200–207.

Singer, M., Fazaluddin, A., & Andrew, K. N. (2013). Recognition of categorised words: Repetition effects in rote study. *Memory, 21,* 467–481.

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.

Starns, J. J., Ratcliff, R., & White, C. N. (2012a). 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, 38,* 1137–1151.

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.

Starns, J. J., White, C. N., & Ratcliff, R. (2012b). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition, 40,* 1189–1199.

Starns, J. J., White, C. N., & Ratcliff, R. (2013). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory and Cognition, 40,* 1189–1199.

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.

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