Memory strength and the decision process in recognition memory

MICHAEL F. VERDE
University of Plymouth, Plymouth, England

AND

CAREN M. ROTELLO
University of Massachusetts, Amherst, Massachusetts

We investigated the role that memory strength plays in the decision process by examining the extent to which strength is used as a cue to dynamically modify recognition criteria. The study list consisted of strong and weak items, with strength a function of study duration or repetition. The recognition test list was divided into two consecutive blocks; strong items appeared in one block, weak items in the other. If the change in item strength across blocks leads to a shift in criterion, the false alarm rate should change accordingly. In four experiments, the false alarm rates did not change across blocks, even when the difference between the strong and the weak items was magnified and marked with semantic cues. However, the strength of the items in the first test block affected the false alarm rate. Thus, strength cues influence initial criterion placement but fail to induce criterion shifts following permanent and even dramatic changes in item strength. These null findings are contrasted with those in a fifth experiment, in which accuracy feedback produced dynamic criterion shifts.

Sitting down at a café, you notice a woman at the next table who seems very familiar. Although you are not certain, you suspect that she may be your wife's friend, whom you recently met at a party. Should you say hello and risk embarrassment if this turns out, after all, to be a stranger? Or should you ignore her and risk insulting a friend of your wife? As this vignette illustrates, every memory judgment has two components. The *retrieval* process generates information on the basis of a retrieval cue: a feeling of familiarity, remembered details of a past event, and so on. The *decision* process determines how we act on this information. Historically, memory research has focused on the factors that affect retrieval accuracy, often treating other aspects of the judgment as noise or, at best, as things to be accounted for abstractly. However, the role of the decision process has become increasingly prominent in debates surrounding such issues as false and illusory memories (Hekkanen & McEvoy, 2002; Hirshman & Arndt, 1997; Miller & Wolford, 1999; Verde & Rotello, 2003), mirror effects (Glanzer, Kim, & Adams, 1998; Greene, 1996; Hintzman, Caulton, & Curran, 1994; Hirshman, 1995; Stretch & Wixted, 1998), and subjective awareness and phenomenology (Donaldson, 1996; Dunn, 2004; Hirshman & Henzler, 1998; Rotello, Macmillan, Reeder, & Wong, 2005; Verde, 2004). Moreover, growing appreciation of the richly metacognitive nature of memory places pressure on formal memory models to treat decision processes in a substantive way. Metacog-

nitive accounts of recognition suggest that people rely on stimulus or environmental cues that tell them how to decide that something was encountered in the past. The nature of these cues, however, remains poorly understood. In the present study, we considered the role that memory strength plays in setting decision rules over the course of many recognition judgments.

Signal detection theory (SDT) provides a simple way to quantify the distinction between the retrieval and the decision components of recognition memory. We adopt here a *memory strength* framework that is prevalent among SDT models of recognition (for reviews, see Banks, 1970; Clark & Gronlund, 1996) and fits intuitively with many less formal approaches. In this view, the retrieval process compares the test probe with the contents of memory, resulting in some degree of match (similarity, learned association, etc.) that can be represented by a scalar value, memory strength. The greater the memory strength, the more evidence there is that the probe represents something encountered in the past. Each class of probe items (*old* or *new*) is associated with a distribution along the axis of memory strength, as is illustrated in Figure 1. Although old items have greater strength on average, the overlap of the distributions introduces uncertainty about whether a given probe has been studied. Because of this uncertainty, the observer must decide on the minimum amount of evidence required to call a probe "old." In Figure 1, this decision criterion is labeled *C*. The observer responds

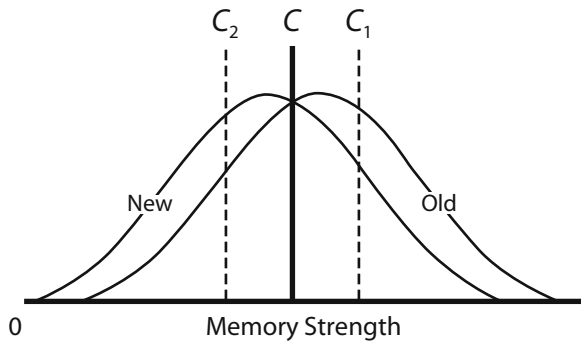M. F. Verde, michael.verde@plymouth.ac.uk

Figure 1. Signal detection theory memory strength model. *C*, decision criterion for new and old item strength distributions; *C_1*, conservative criterion; *C_2*, liberal criterion.

"old" only if probe strength exceeds *C*. Note that *C* may be placed anywhere on the strength axis. A more conservative response bias moves the criterion upward (C_1), decreasing both hits (old items called "old") and false alarms (new items called "old"); a more liberal response bias moves the criterion downward (C_2), increasing hits and false alarms. In this SDT model, the retrieval process generates the evidence distributions, and the decision process governs the placement of the decision criterion.

Theoretical approaches to the recognition decision process that do not explicitly adopt the SDT framework are, nonetheless, often compatible with it. Many take the form of metacognitive or attribution-based explanations for why certain manipulations affect memory performance. A classic example is the fluency attribution account offered by Jacoby and Whitehouse (1989) for their finding that repetition priming can have very different effects on recognition performance, depending on the subjects' response strategies. Repetition priming is thought to increase the fluency or familiarity of the probe, thus increasing the feeling that it has been recently encountered. As long as the prime is quickly presented and masked (making it difficult to identify), subjects become more likely to respond "old." When the prime is easily identified, on the other hand, subjects apparently discount the increase in fluency and become less willing to respond "old." In the SDT framework, priming might be thought to affect memory strength and the manipulation of prime subtlety to affect criterion placement.

Memory Strength As a Decision Cue

How do people choose a decision criterion? The fact that response bias can be manipulated experimentally makes it clear that people do not choose their criteria at random. The rules that guide criterion placement must depend on *decision cues* drawn from the stimulus and the environment. Identifying these cues is a critical step toward understanding the precise nature of the decision rules themselves, and in the present study, we considered the most basic of cues, memory strength. SDT-based models typically assume that response bias is determined by the properties of the strength distributions. Similarly, many metacognitive accounts suggest that recognition

thresholds are based, in part, on a subjective awareness of *familiarity* or *fluency*, terms often used synonymously with *memory strength*. Unfortunately, such assumptions have not been well tested, because strength is usually confounded with other cues present at test. In Jacoby and Whitehouse's (1989) study, for example, one possibility is that the subjects sensed a change in memory strength following the prime and adjusted their decision criterion accordingly. Another possibility, however, is that they simply responded to the presence of the prime itself, perhaps on the basis of some naive belief about the effect of priming. The extent to which people base recognition judgments on direct access to the memory representation (e.g., strength), as opposed to indirect cues or inferences (e.g., the prime or beliefs about priming), is the subject of ongoing debate in the metacognitive literature (Schwartz, Benjamin, & Bjork, 1997; Strack & Forster, 1995).

Hirshman (1995) examined the influence of memory strength on decision criteria in experiments on the list strength effect (Ratcliff, Clark, & Shiffrin, 1990). In a typical list strength experiment, there are three types of study lists, which differ in average memory strength: weak lists containing words studied once or shown for a short duration, strong lists containing words studied several times or shown for a long duration, and mixed lists containing both strong and weak words. A consistent pattern emerged across the studies summarized by Hirshman: Subjects adopted more conservative recognition criteria following higher strength study lists. Hirshman interpreted this correlation as clear evidence that strength controls criterion placement. Moreover, he showed that when the average strength of the *test* list was held constant (only weak items were tested), study list strength still affected response bias. It appeared that the subjects chose their decision criterion solely on the basis of what they had encountered at study.

Other researchers also have failed to find evidence that memory strength affects criterion placement during a recognition test. Stretch and Wixted (1998) replicated the between-list pattern of strength-based criterion shifts but also examined more closely the effect of strength within the test list. Their study items were low-frequency (LF) and high-frequency (HF) words, and the issue was whether subjects use a single decision criterion for both classes of items or separate criteria for LF and HF words. In the weak condition in their Experiment 2, all the words were studied once. In the strong condition, HF words were studied five times, and LF words were studied once. A typical effect of word frequency is that there are fewer false alarms to LF than to HF words. However, if the strength manipulation led the subjects to selectively adopt a more conservative criterion for HF words, the HF false alarm rate would decrease below that of LF words. This did not happen, and Stretch and Wixted concluded that subjects use the same criterion for all classes of items within the same test list. More specifically, when two classes of items differing in average strength are mixed at test, subjects do not use strength as a cue to adjust their criterion from one trial to the next. In other experiments, Stretch and Wixted attempted to make the different strength classes more obvious, first by increasing the strength differential and

then by presenting each class in a different color. Morrell, Gaitan, and Wixted (2002) differentiated strong and weak classes by semantic category and by iconic type (words vs. pictures). None of these experiments succeeded in producing trial-to-trial criterion shifts.

Benjamin (2001) noted a possible case of trial-to-trial shifts in a false memory task. Subjects studied several categories, each composed of words highly associated with a critical (nonstudied) lure. Some categories were repeated once; others were repeated three times. Repetition should increase the familiarity of both studied words and critical lures, leading to an increase in hits and false alarms. The data showed an increase in studied word hit rate in the repeated condition to a similar degree for both younger and older adults. Repetition also increased the rate of critical lure false alarms in older adults but decreased it in younger adults. Benjamin suggested that younger adults adopted a stricter criterion for the stronger categories—in effect, shifting their criterion from trial to trial on the basis of item strength. However, adopting a stricter criterion for a given category should decrease the rate of saying "old" for both studied and lure members of that category. It is thus unclear how a simple criterion shift could produce similar increases in hits for both age groups but very different patterns of false alarms.[1]

The picture that emerges from the existing data is that memory strength plays a very limited role in the decision process: People choose an appropriate criterion location on the basis of the average strength of the study list, but they fail to use strength cues at test to adjust this criterion in a dynamic, trial-by-trial fashion. However, there is some new evidence that less frequent criterion changes may occur: Benjamin and Bawa (2004) found that people may shift their criterion when the difficulty of the recognition task changes dramatically and permanently in the middle of the test list. Subjects studied words drawn from several semantic categories. The recognition test consisted of two consecutive blocks. In one block, the test lures were categorically related to the studied words, so that discriminating between old and new words was relatively difficult. In the other block, the test lures were unrelated, and discrimination was easier. The order of these blocks was manipulated between subjects. The subjects were found to have revised their criterion following a switch from unrelated to related lures (easy to hard task) but failed to do so following a switch from related to unrelated lures (hard to easy task). Benjamin and Bawa proposed that people are motivated to change strategies (in this case, their decision criterion) only when a task becomes more difficult and performance drops, as was true for the unrelated-to-related-lure condition.

Benjamin and Bawa's (2004) findings suggest that although people do adjust their decision rules to test conditions, there are limits. Thus, it might be that people find it too difficult or effortful to constantly shift their criterion as item strength varies from one trial to the next but will do so on a one-time basis if the strength conditions change significantly and permanently. Unfortunately, Benjamin and Bawa's study does not answer this question, because their manipulation did not separate strength from the se-

mantic relationship between old and new items. In the present study, we used a variation of their design that allowed us to isolate the contribution of strength. Our study list consisted of both strong and weak items, with memory strength manipulated by study duration or repetition. The test list was divided into two blocks; strong old items were tested in one block, weak old items were tested in the other. Lures were always new words that had never appeared in the study list.

Figure 2 illustrates equal-variance memory strength distributions for the three classes of test items: new, weak old, and strong old. Between-list manipulations of strength have shown that subjects adopt more liberal recognition criteria, meaning that they accept lower standards of evidence, as lists decrease in average strength (and it becomes hard to discriminate between old and new items; Hirshman, 1995). Thus, if C_s represents the location of the criterion in a list composed of only new and strong old items, a criterion lower on the strength axis (C_w) will be expected for a list composed of only new and weak old items. One consequence of lowering the criterion is an increase in false alarms (the area under the *new* distribution above the criterion). In the present experiments, if the change in test item strength from Block 1 to Block 2 led the subjects to modify their decision criterion, the false alarm rate would change as a consequence, because new items were drawn from the same pool in both test blocks.

In Experiments 1–3, we looked for evidence of a strength-based criterion shift under conditions analogous to those in Benjamin and Bawa's (2004) study. All the study lists consisted of both strong and weak items, in equal numbers. However, strong items appeared only in Test Block 1, and weak items only in Test Block 2. If subjects look to test strength as a cue, the increase in recognition difficulty halfway through the test should lead the subjects to adopt a more conservative criterion, just as Benjamin and Bawa observed. We manipulated strength via duration (Experiment 1), repetition (Experiment 2), and both repetition and category membership (Experiment 3), in order to make the change in strength increasingly more obvious to the subjects. In none of the experiments was there evidence of a criterion shift across blocks. Experiment 4 was designed to evaluate whether subjects
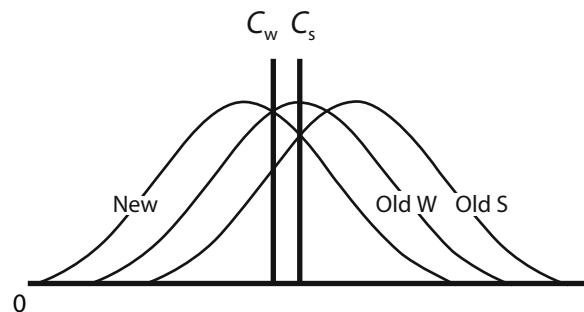

Figure 2. Strength distributions and decision criteria. Assume that C_s is the criterion when the test list contains new and strong old items. It is predicted that the criterion will shift downward to C_w when the test list contains new and weak old items.

choose their criterion on the basis of test conditions but fail to adjust it appropriately when conditions change. Finally, Experiment 5 showed that in contrast to memory strength cues, subjects will use accuracy feedback cues to dynamically adjust their criteria.

EXPERIMENT 1

Benjamin and Bawa (2004) showed that increasing the difficulty of discrimination midway through a memory test can lead subjects to modify their response bias. We would expect similar results if subjects choose a recognition criterion on the basis of the strength of test items. Each of the words in the study list was shown for 500 msec (weak) or 3,000 msec (strong). In the first half of the recognition test, the subjects were tested only on strong old words and new words. In the second half, they saw only weak old words and new words. If the strength of a new item is constant, the false alarm rate serves as a measure of response bias, because any shift in bias will lead to a corresponding change in the false alarm rate. Specifically, if subjects adopt a more liberal bias in Block 2, as predicted by the findings of Benjamin and Bawa (2004) and Hirshman (1995), false alarms should increase in Block 2.

Method

Subjects. Twenty-five undergraduates from the University of Massachusetts at Amherst participated for extra credit in their psychology courses.

Materials and Design. The stimuli were drawn from a pool of 300 low-frequency (<100 per million; Kučera & Francis, 1967), 5- to 8-letter nouns. The study list consisted of 88 words: 40 short-duration words (weak condition), 40 long-duration words (strong condition), and 8 filler words placed at the beginning and end of the list. The test list consisted of 162 recognition probes: 80 old (studied) words, 80 new words, and 2 new filler words placed at the beginning of the list. Nonfiller words were randomly positioned within each list, with the constraint that an equal number of strong and weak condition words should appear in each half of the study list. The first half of the test list was designated *Block 1*, and the second half *Block 2* (although from the perspective of the subject, there was only a single, continuous test list). There were 40 old strong and 40 new words in Block 1 and 40 old weak and 40 new words in Block 2. The assignment of words to condition and list position was uniquely randomized for each subject. List creation, stimulus presentation, and response collection were computer controlled, and the subjects were assigned to individual computers and testing rooms.

Procedure. The 30-min session consisted of a study phase followed by a test phase. At the beginning of the study phase, the subjects were instructed to learn the words for an upcoming memory test. The words were then presented individually on the computer screen: weak condition words for 500 msec and strong condition words for 3,000 msec.

During the test phase, each trial began with a fixation line displayed in the center of the screen for 500 msec. This was replaced with the test probe, which remained until both responses were made. The subjects first indicated whether the probe was an old or a new word (using the "/" and "z" keys, respectively). They then made a confidence rating on a 6-point scale, ranging from 1 = *very sure new* to 6 = *very sure old* (using the 1–6 keys at the top of the keyboard). A 1,500-msec blank interval concluded the trial. On-screen prompts showing the mapping of keys to response categories appeared below the stimulus display area.

Results

We analyzed recognition performance in two steps. First, we examined accuracy, for which hits and false alarms are considered together. Receiver operating characteristics (ROCs) were constructed from confidence ratings for each subject and each condition. These gave a good index of accuracy, A_z, which estimates the area under the ROC (Macmillan & Creelman, 2005; Verde, Macmillan, & Rotello, 2006). Second, we derived overall hit and false alarm rates by aggregating confidence ratings 1–3 into the *new* response category and confidence ratings 4–6 into the *old* response category (see Table 1). We examined false alarm rates alone to determine the presence of a criterion shift.

As would be expected, the subjects were better at recognizing the strong items in Block 1 than the weak items in Block 2. Accuracy in Block 1 was significantly greater than that in Block 2 [A_z = .80 vs. .71; $t(24)$ = 5.72, $p <$ .001]. However, false alarm rates in Blocks 1 and 2 did not reliably differ [.28 vs. .28; $t(24)$ = 0.10, n.s.]. Thus, the change in memory strength across blocks had no observable effect on response bias.

EXPERIMENT 2

Although the difference in accuracy between strong and weak items indicates that the strength manipulation was successful, it is possible that the strength difference was not readily perceived by the subjects. In Experiment 2, we magnified the difference between weak and strong items by manipulating study repetition, rather than study duration. It is well established that spaced practice (repetition) is more effective than massed practice (duration) at improving memory (for reviews, see Crowder, 1976; Melton, 1970). In this experiment, weak items appeared once in the study list, whereas strong items appeared four

Table 1
Hit and False Alarm (FA) Rates and Accuracy (A_z) by Test Block
in Experiments 1–5 (With Standard Errors)

	Block 1						Block 2					
	Hits		FAs		A_z		Hits		FAs		A_z	
Experiment	M	SE	M	SE	M	SE	M	SE	M	SE	M	SE
1	.73	.02	.28	.02	.80	.02	.60	.04	.28	.02	.71	.02
2	.87	.02	.22	.02	.90	.01	.55	.03	.24	.02	.71	.02
3	.86	.02	.33	.02	.85	.01	.65	.02	.34	.02	.71	.02
4	.68	.02	.33	.02	.73	.02	.85	.01	.32	.03	.86	.01
5	.81	.02	.21	.02	.86	.02	.60	.02	.32	.03	.69	.02

times. The procedure was otherwise identical to that in Experiment 1.

Method

Subjects. Twenty-seven undergraduates from the University of Massachusetts at Amherst participated for extra credit in their psychology course.

Materials and Design. Stimulus words were drawn from the pool used in Experiment 1. However, the study list consisted of 208 items: 40 words presented once (weak condition), 40 words presented four times (strong condition), and 8 filler words placed at the beginning and end of the list. The test list consisted of 162 recognition probes, as in Experiment 1. There were 40 strong old and 40 new words in Block 1 and 40 weak old and 40 new words in Block 2.

Procedure. The procedure was identical to that in Experiment 1, save that during the study phase, each word was shown for 750 msec.

Results

Summary statistics are shown in Table 1. We expected that there would be better recognition for the strong items in Block 1 than for the weak items in Block 2 and that the difference would be greater than that observed in Experiment 1. Accuracy in Block 1 was significantly greater than that in Block 2 [A_z = .90 vs. .71; $t(26)$ = 10.76, $p <$ .001]. The difference in A_z was more than double that observed in Experiment 1, and in absolute terms, such discriminability is quite good by most standards. The difference in hit rates (.32) was also dramatic. It is hard to imagine that such large differences in objective performance could fail to produce a subjective awareness of the strength differential. Even so, the false alarm rates in Blocks 1 and 2 did not reliably differ [.22 vs. .24; $t(27)$ = 1.29, n.s.]. Thus, the change in memory strength across blocks had no observable effect on response bias.

EXPERIMENT 3

In Experiment 3, we attempted to make the change in strength from Block 1 to Block 2 more obvious by associating each strength class with a different category. Rather than random nouns, the study materials were male and female names common in the U.S. (e.g., David, Susan). Weak items were drawn from one gender, and strong items from the other gender. Thus, the change in memory strength from Block 1 to 2 was also signaled by a change in the gender category of both old and new test items. Because the association between gender category and memory strength was arbitrary and novel to the experimental context, it can be assumed that a response to the change of gender category would be based on the change in memory strength.

Method

Subjects. Twenty-eight undergraduates from the University of Massachusetts at Amherst participated for extra credit in their psychology course.

Materials and Design. The stimuli were 40 male and 40 female names. For half of the subjects, male names were assigned to the strong condition, and female names to the weak condition; for the other half, the reverse was true. Aside from the nature of the stimuli, the design was identical to that in Experiment 2; the names in the strong condition were shown four times each, and the names in the weak condition were shown once each. There were 40 strong old and 40 new names of one gender in Block 1 and 40 weak old and 40 new names of the other gender in Block 2.

Procedure. The procedure was identical to that in Experiment 2.

Results

Summary statistics are shown in Table 1. As in Experiment 2, manipulating study repetition had a large effect on memory for weak and strong old items. Accuracy in Block 1 was significantly greater than that in Block 2 [A_z = .85 vs. .71; $t(27)$ = 9.63, $p <$ .001]. However, false alarms in Blocks 1 and 2 did not reliably differ [.33 vs. .34; $t(27)$ = 0.12, n.s.]. Once again, the change in memory strength from Block 1 to Block 2 (easy to hard) had no observable effect on response bias even when the strength categories were clearly marked by associating them with different gender categories.

DISCUSSION: EXPERIMENTS 1–3

In three experiments, permanent, dramatic, and clearly marked changes in the average memory strength of test items failed to induce subjects to modify their recognition criteria. In Benjamin and Bawa's (2004) study, subjects who encountered more difficult discriminations midway through the memory test responded by adopting a more liberal response bias. It was unclear, however, whether their subjects were responding to the change in test item strength or to the change in the semantic relationship between old and new items. Our results suggest that it was the latter. Consistent with the experiments that varied strength on a trial-by-trial basis (Morrell et al., 2002; Stretch & Wixted, 1998), our data indicate that people do not rely on strength cues to modify their criterion in a dynamic way. However, there are two important issues that these experiments leave unresolved.

Our results do not rule out the possibility that strength cues at test influence initial criterion placement. Our design makes it possible to change the initial test conditions while holding the study conditions constant, simply by reversing the order in which subjects encounter strong and weak items at test. If subjects encounter only weak old items in Block 1, will they adopt the same decision rule as subjects exposed only to strong items in Block 1? Experiment 4 was identical to Experiment 2, but the order in which strong and weak items were tested was reversed. Finding that the subjects adopted the same criterion in both experiments would be consistent with evidence from Hirshman (1995) that the strength of items during the study phase alone determines the decision rule.

Although one should always be cautious when drawing conclusions from null results, the failure of strength cues to produce dynamic criterion shifts has been consistently observed across several studies. If strength differences can lead to such extreme differences in objective accuracy as those observed in Experiment 2, one is led to doubt that people use strength cues under any normal circumstances. Nevertheless, putting these null results in the context of a set of positive results would make them more compelling. We attempted to do this while addressing two specific

concerns. First, in the context of a lexical decision task, Brown and Steyvers (2005) raised the important point that it takes some time to become aware of changes in test conditions; changes in response bias can lag significantly behind changes in test conditions. In our experiments, 80 trials followed the shift in old item strength. In the studies of Brown and Steyvers (2005) and Benjamin and Bawa (2004), far fewer trials were needed to induce criterion shifts. Still, Experiment 4 can shed light on this question, because any effect of strength on initial criterion placement should be observed within the first 80 test trials. A second concern is that people might simply be unwilling to modify their criterion dynamically, regardless of the cues. In Experiment 5, we used the same materials and procedure as those in Experiment 2 but added accuracy feedback, a cue that we suspected would be more successful at inducing criterion shifts.

EXPERIMENT 4

The previous experiments suggested that strength cues do not influence subjects to change their response bias. The purpose of Experiment 4 was to determine whether strength at test affects the initial choice of a criterion. In one experiment, Hirshman (1995, Experiment 4) had subjects study lists that varied in average strength (they contained different proportions of strong and weak items). However, only weak old items were presented at test, so that test strength was held constant across study conditions. Differences in response bias were still observed between conditions. Although this result shows that study conditions may influence criterion placement, it does not logically rule out the possibility that test conditions may also have an influence. In Experiment 4, the subjects were shown a study list similar to that in Experiment 2. However, the order of weak and strong old items was reversed; weak items were tested in Block 1, and strong items were tested in Block 2. If the subjects attended to the strength conditions at test when choosing their initial criterion, the Block 1 false alarm rates of the two experiments should differ. In line with what has been observed when between-list strength at study has been manipulated, a more liberal criterion should result in a higher false alarm rate when weak items were initially encountered at test.

Method

Subjects. Twenty-five undergraduates from the University of Massachusetts at Amherst participated for extra credit in their psychology course.

Materials and Design. The study and test lists were identical to those in Experiment 2, except that weak (one-repetition) words appeared only in the first half (Block 1) of the test list, whereas strong (four-repetition) words appeared only in the second half (Block 2) of the test.

Procedure. The procedure was identical to that in Experiment 2.

Results

Summary statistics are shown in Table 1. Recognition for strong items was better than that for weak items. Accuracy in Block 1 was significantly lower than that in Block 2 [$A_z = .73$ vs. .86; $t(24) = 10.50, p < .001$]. How-

ever, false alarms in Blocks 1 and 2 did not reliably differ [.33 vs. .32; $t(24) = 0.66$, n.s.]. As in the previous experiments, the change in memory strength from Block 1 to Block 2 had no observable effect on response bias.

To examine the influence of test strength on initial criterion placement, we compared the Block 1 false alarm rates in Experiments 2 and 4, which used identical study lists and procedures but differed in whether weak or strong items appeared in the first half of the test. If the decision criterion is based solely on what was encountered at study, response bias should be identical for these experiments. In fact, the significant difference in false alarm rates [.22 vs. .33; $t(50) = 3.77, p < .001$] indicated that the subjects adopted a more liberal bias in Experiment 4. This could not be tied to differences in overall accuracy between the experiments: Collapsing over blocks, accuracy was nearly identical in Experiments 2 and 4 [$A_z = .80$ vs. .79; $t(50) = 0.40$, n.s.]. The difference in false alarm rates is consistent with that observed when strength is manipulated between study lists and shows that strength conditions at test do influence response bias, contrary to Hirshman's (1995) findings.

These results also show that the length of the test list was not a critical factor in the failure of the subjects to modify their criterion across blocks. The difference in bias between Experiments 2 and 4 indicates that the subjects were able to determine the average strength of test items within the 80 trials in Block 1 and, thus, should have been able to do so within the 80 trials in Block 2 following the strength change. In sum, strength cues did influence the subjects to choose a decision rule at the outset of the test, but not to change this rule during the course of the test.

EXPERIMENT 5

Accuracy feedback has been shown to influence the calibration of confidence and decision criteria in a number of domains. In recognition memory, Estes and Maddox (1995) found that feedback influenced base-rate-related bias shifts with some materials. Their data suggest that feedback has the potential to be an effective signal of changes in test conditions. Experiment 5 was identical to Experiment 2, save for the addition of accuracy feedback after each trial of the recognition test.

Method

Subjects. Twenty-six undergraduates from the University of Massachusetts at Amherst participated for extra credit in their psychology course.

Materials and Design. The study and test lists were identical to those in Experiment 2. Four-repetition (strong) studied words appeared only in the first half (Block 1) of the test list, whereas one-repetition (weak) studied words appeared only in the second half (Block 2) of the list.

Procedure. The procedure was identical to that in Experiment 2 in most respects. However, prior to the recognition test, the subjects were told that they would be given feedback as a way to improve their accuracy. During the test, following each confidence rating response, the subjects were informed whether their recognition response was correct or incorrect with a message that replaced the test probe. In addition, after each quarter of the test list (40 trials), the

subjects were shown a summary of the total correct and incorrect responses for that quarter of the test.

Results

Summary statistics are shown in Table 1. In line with previous results, accuracy in the easier Block 1 was significantly higher than that in Block 2 [A_z = .86 vs. .69; $t(25)$ = 10.49, $p <$ .001]. However, accuracy feedback produced a large and reliable increase in the false alarm rate from Block 1 to Block 2, consistent with a liberal shift in response bias [.21 vs. .32; $t(25)$ = 4.44, $p <$ .001]. Thus, the null results from the previous experiments were apparently tied to the type of cue provided, rather than to the subjects' inability to shift their decision criterion.

Should accuracy feedback be qualitatively distinguished from memory strength, or does it merely enhance the cuing properties of strength? There are two points to consider. First, if the efforts in Experiments 1–4 to make strength information available to the subjects still failed to do so, one has to doubt the ability of memory strength by itself to cue bias shifts under reasonable circumstances. Second, if feedback only increases the effectiveness of strength cues, one would expect the increasing efforts over the previous experiments to result in a graded effect on bias, rather than in the abrupt discontinuity observed when Experiment 5 is compared with the previous experiments. Although the effects of feedback depend on the changes in discriminability caused by the strength shift, it is not clear that subjects consciously consider strength when they respond to feedback.

GENERAL DISCUSSION

In recent years, a great deal of interest has focused on the decision processes that determine how we use information retrieved from memory. Metacognitive theorists, for example, argue that attribution and subjective interpretation are as critical as the objective properties of the retrieval cue in creating feelings of familiarity and experience (Jacoby & Whitehouse, 1989; Whittlesea & Williams, 1998). The role of response bias in producing false and illusory memories (Hekkanen & McEvoy, 2002; Hirshman & Arndt, 1997; Miller & Wolford, 1999; Verde & Rotello, 2003), the word frequency effect, and other puzzling mirror effects (Glanzer et al., 1998; Greene, 1996; Hintzman et al., 1994; Hirshman, 1995; Stretch & Wixted, 1998), as well as other phenomena critical to building theories of recognition, continues to be the focus of vigorous debate. Response bias is of special concern in applied areas such as eyewitness memory (for a review, see Wells & Olson, 2003), where bias has direct, real-world consequences. Finally, formal models of memory have long incorporated decision parameters alongside mechanisms of encoding and retrieval, and understanding the conditions of criterion-setting is necessary to constrain these models.

The notion that memory strength (sometimes referred to as *familiarity*, *fluency*, etc.) directly influences how people choose their recognition criterion is common among SDT-based models and metacognitive accounts of mem-

ory. Unfortunately, the few studies in which strength has been examined in isolation from other cues have offered little evidence that memory strength influences criterion placement in a dynamic way. The question of interest to us was whether people can make use of strength cues at all during the course of a memory test. Previous studies have shown that people are unwilling or incapable of changing response bias from trial to trial when items from different strength classes are intermixed within a test list (Morrell et al., 2002; Stretch & Wixted, 1998). Although one might intuitively predict that people lack the motivation or the information necessary to constantly modify their criterion, trial-to-trial criterion shifts have been observed in other circumstances, such as when the nature of the task changes dramatically from one trial to the next (Heit, Brockdorff, & Lamberts, 2003; Hicks & Marsh, 1998; Hockley & Niewiadomski, 2001; Verde & Rotello, 2003) or when typical words are intermixed with unusual classes of items (Whittlesea & Williams, 1998; Windmann & Kutas, 2001; Wixted, 1992). In all of these examples, strength was not separated from other aspects of the task or materials.

In the present study, we investigated the more intuitively plausible possibility that people will use strength as a cue to modify their criterion on a one-time basis in response to a significant and permanent change in the strength characteristics of the test probes. Benjamin and Bawa (2004) observed this to happen when a recognition test was made more difficult partway through. In Experiments 1–3, we attempted to replicate this finding, using a design that allowed us to look at strength independently of other stimulus characteristics. The study list consisted of strong and weak items, with strength manipulated by study duration or repetition. During the test, only strong old items were shown in the first half (Block 1), and only weak old items were shown in the second half (Block 2). Previous studies in which the average strength of study lists was manipulated showed that subjects responded more liberally after studying weak lists (Hirshman, 1995). If people also pay attention to the strength conditions at test, they should likewise adopt a more liberal criterion when the test shifts to weaker items in Block 2. In all three experiments, we failed to find evidence of a criterion shift, despite attempts to make the change in strength very obvious. Experiment 4 also failed to produce a criterion shift when test item strength changed from weak to strong across blocks. However, the subjects responded more liberally in this experiment than in Experiment 2, which was identical in all respects, save for the direction of the strength change. In other words, the subjects were influenced, at least in part, by the strength of the items first encountered during the test phase. This result is contrary to Hirshman's (1995) results, which suggest that the strength conditions at study alone determine the decision criterion at test.

The stable response bias across blocks in Experiments 1–4 contrasts with the findings of Experiment 5, in which the subjects readily changed bias across blocks when provided with accuracy feedback. Apparently, people are motivated to adapt their decision rule to changing conditions at test but do not normally use strength cues to do so. This insight has broader implications for the development of

theory. First, recent debates in the false memory literature illustrate the difficulty of distinguishing the effects of retrieval and the effects of decision on recognition performance (Miller & Wolford, 1999; Roediger & McDermott, 1999; Verde & Rotello, 2003; Wixted & Stretch, 2000). Accounts that emphasize the latter are often tentative and mainly theoretical, simply because of the paucity of empirical studies aimed at defining the normal parameters of the decision process. An explanation that invokes shifts in criterion can be judged plausible by the presence of known decisions cues. Given the evidence thus far, any explanation that requires subjects to shift their criteria in a dynamic fashion on the basis of strength cues alone is subject to question on grounds of plausibility. These findings are also relevant to one of the central issues in the study of metacognition: Are judgments based primarily on direct access to memory representations or on indirect cues and inferences? The bias differences between Experiments 2 and 4 suggest that strength cues influence initial criterion placement, and this is evidence of direct access. However, the limited role of strength cues over the course of the test, as compared with feedback cues (Experiment 5), suggests that indirect or inferential cues may normally exert greater control over the decision process.

With cues that do influence dynamic shifts in criterion, people most likely respond not to the identity of the cues per se but, rather, to some underlying stimulus property that the cues reveal. Benjamin and Bawa (2004) found that altering the identity of new lures while holding old items constant induced a criterion shift. In our Experiments 2 and 4, we observed a difference in criterion placement when new items were held constant but old item strength was manipulated. Taken together, these findings suggest that subjects must be attending to the properties of both new and old items. This makes sense, given the standard notion that optimal observers base their criterion on discriminability, which is defined by the relative distance and shapes of new and old distributions. Discriminability may be the underlying property reflected by indirect cues, such as the semantic relationship between old and new items in Benjamin and Bawa's study.

REFERENCES

BANKS, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, **74**, 81-99.

BENJAMIN, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 941-947.

BENJAMIN, A. S., & BAWA, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory & Language*, **51**, 159-172.

BROWN, S., & STEYVERS, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 587-599.

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

CROWDER, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

DONALDSON, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, **24**, 523-533.

DUNN, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review*, **111**, 524-542.

ESTES, W. K., & MADDOX, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1075-1095.

GALLO, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 120-128.

GLANZER, M., KIM, K., & ADAMS, J. K. (1998). Response distribution as an explanation of the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 633-644.

GREENE, R. L. (1996). Mirror effect in order and associative information: Role of response strategies. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 687-695.

HEIT, E., BROCKDORFF, N., & LAMBERTS, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review*, **10**, 718-723.

HEKKANEN, S. T., & MCEVOY, C. (2002). False memories and source-monitoring problems: Criterion differences. *Applied Cognitive Psychology*, **16**, 73-85.

HICKS, J. L., & MARSH, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1105-1120.

HINTZMAN, D. L., CAULTON, D. A., & CURRAN, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 275-289.

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

HIRSHMAN, E., & ARNDT, J. (1997). Discriminating alternative conceptions of false recognition: The cases of word concreteness and word frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 1306-1323.

HIRSHMAN, E., & HENZLER, A. (1998). The role of decision processes in conscious recollection. *Psychological Science*, **9**, 61-65.

HOCKLEY, W. E., & NIEWIADOMSKI, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition*, **29**, 1176-1184.

JACOBY, L. L., & WHITEHOUSE, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, **118**, 126-135.

KENSINGER, E. A., & SCHACTER, D. L. (1999). When true memories suppress false memories: Effects of ageing. *Cognitive Neuropsychology*, **16**, 399-415.

KUČERA, F., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

MACMILLAN, N. A., & CREELMAN, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

MELTON, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning & Verbal Behavior*, **9**, 596-606.

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, & Cognition*, **28**, 1095-1110.

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

ROEDIGER, H. L., III, & MCDERMOTT, K. B. (1999). False alarms and false memories. *Psychological Review*, **106**, 406-410.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The *remember* response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, **12**, 865-873.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, **6**, 132-137.

Strack, F., & Forster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science*, **6**, 352-358.

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, & Cognition*, **24**, 1379-1396.

Verde, M. F. (2004). Associative interference in recognition memory: A dual-process account. *Memory & Cognition*, **32**, 1273-1283.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics*, **68**, 643-654.

Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 739-746.

Watson, J. M., McDermott, K. B., & Balota, D. A. (2004). Attempting to avoid false memories in the Deese/Roediger–McDermott paradigm: Assessing the combined influence of practice and warnings in young and old adults. *Memory & Cognition*, **32**, 135-141.

Wells, G. L., & Olson, E. A. (2003). Eyewitness identification. *Annual Review of Psychology*, **54**, 277-295.

Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica*, **98**, 141-165.

Windmann, S., & Kutas, M. (2001). Electrophysiological correlates of emotion-induced recognition bias. *Journal of Cognitive Neuroscience*, **13**, 577-592.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 681-690.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, **107**, 368-376.

NOTE

1. An alternative account is suggested by findings that recalling studied members of a category can reduce false alarms to related lures (Gallo, 2004). Repetition should increase recall, thus reducing false alarms. In Benjamin's (2001) study, perhaps younger adults were better able to use a recall-to-reject strategy to reduce false alarms, an idea consistent with other findings that older adults are less prone to use control processes to reject false lures (Kensinger & Schacter, 1999; Watson, McDermott, & Balota, 2004).