Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 2, 305–320

0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.2.305

On the Dynamic Nature of Response Criterion in Recognition Memory: Effects of Base Rate, Awareness, and Feedback

Matthew G. Rhodes
Colorado State University

Larry L. Jacoby
Washington University in St. Louis

The authors examined whether participants can shift their criterion for recognition decisions in response to the probability that an item was previously studied. Participants in 3 experiments were given recognition tests in which the probability that an item was studied was correlated with its location during the test. Results from all 3 experiments indicated that participants' response criteria were sensitive to the probability that an item was previously studied and that shifts in criterion were robust. In addition, awareness of the bases for criterion shifts and feedback on performance were key factors contributing to the observed shifts in decision criteria. These data suggest that decision processes can operate in a dynamic fashion, shifting from item to item.

Keywords: memory, recognition memory, response criterion, awareness, criterion shifts

Theories of recognition memory distinguish between the ability to discriminate among old and new items and the criterion one sets for deciding whether an item is "old" or "new." The focus of the current article is on changes in criterion that result from varying the base rates of studied items presented for a recognition test. In particular, we ask whether the criterion for recognition memory judgments can operate in a dynamic manner, changing to accommodate different base rates of studied items. If criterion setting can be dynamic, does it depend on feedback regarding the correctness of a judgment? In addition, what role does awareness of base rates play in criterion setting?

Signal Detection Theory and Response Criterion

Analyses of recognition memory that use signal detection theory (e.g., Green & Swets, 1966; Lockhart & Murdock, 1970; Snodgrass & Corwin, 1988) assume that old and new items presented for a recognition decision differ in their familiarity (i.e., the general strength of evidence that they were previously studied). The familiarity of studied items reflects both strength accrued from their presentation (signal) and preexisting strength (noise), whereas the familiarity of new items reflects only preexisting strength (noise). Old and new items constitute overlapping distributions on a continuum of strength of evidence, with the strength of old items (as signal plus noise) generally exceeding the strength of new items (as noise alone). The ideal rememberer is presumed to set a decision criterion at the intersection of the old and new distributions. Items whose strength exceeds the decision criterion are called "old," whereas items whose familiarity does not exceed the decision criterion are called "new." The distance between old and new distributions (or the degree to which one can distinguish between old and new items) refers to discriminability.

The manner in which evidence is accrued for recognition decisions has been a popular topic for theorizing (e.g., Hintzman, 1988; Murdock, 1993). However, the decision stage has received considerably less attention and remains poorly understood (but see, e.g., Benjamin & Bawa, 2004; Dobbins & Kroll, 2005; Estes & Maddox, 1995; Healy & Kubovy, 1978; Hirshman, 1995; Stretch & Wixted, 1998; Verde & Rotello, in press). One persistent issue regards the malleability of the criterion that the rememberer sets for recognition decisions. For example, several reports have suggested that participants are sometimes unwilling to adjust their decision criterion from item to item even when there are clear differences in the memorability of items to be judged. Perhaps the most dramatic example of this reluctance was reported by Stretch and Wixted (1998; see also Hirshman, 1995). Participants studied items presented five times in one color (strong items), studied other items presented once in a different color (weak items), and completed a recognition test. In their Experiment 2, participants first studied the strong or weak items and then completed a recognition test for those items, followed by a separate study and test list for the remaining items. If participants' response criterion was sensitive to study strength, one would expect that endorsements of distractors (i.e., false alarms; FAs) would be lower in the test of strong items than in the test of weak items. That is, if participants demanded more evidence before endorsing an item in the strong list, distractors should be endorsed less frequently in such lists. Results were consistent with this prediction, as participants exhibited fewer FAs to distractors from the strong list than they did to those from the weak list (see Verde & Rotello, in press, for an exception to this pattern). Follow-up experiments (their Experiments 4 and 5) used a similar procedure with one major exception: both the study and test lists were intermixed. Thus, participants were administered a single recognition test consisting of strong and weak targets as well as distractors presented in the color of strong or weak targets. In contrast

Matthew G. Rhodes, Department of Psychology, Colorado State University, Larry L. Jacoby, Department of Psychology, Washington University in St. Louis.

We thank John Dunlosky and Charles Brainerd for their comments on a draft of this article. We also thank Tim Bono, Nancy Byars, Carole Jacoby, Emily Norwood, Allison Silvers, Shari Steinman, and Beth Yanco for their assistance with data collection and Daniel Rhodes for his assistance with the supplemental data analyses reported for Experiment 3.

Correspondence concerning this article should be addressed to Matthew G. Rhodes, Department of Psychology, Fort Collins, CO 80523-1876. E-mail: matthew.rhodes@colostate.edu

to their Experiment 2, Stretch and Wixted did not observe shifts in response criterion, with nominally equivalent levels of FAs for "strong" and "weak" lures.

Subsequent experiments (Morrell, Gaitan, & Wixted, 2002) replicated this pattern, with criterion remaining unaffected by different strength manipulations (e.g., differential strengthening of taxonomic categories) when items were tested within a single list. Thus, in summarizing their experiments, Morrell et al. concluded that, ". . .participants are reluctant to shift their criterion when strength is conspicuously manipulated and that any shift that might occur is surprisingly small even when extraordinary steps are introduced to make it happen" (p. 1103). This echoes a suggestion made 25 years earlier by Brown, Lewis, and Monk (1977) that the degree of evidence required for recognition judgments of differentially memorable items might vary if such items were tested separately. However, "if a mixed-list design is used, in which the subject encounters both types of items in a random order, it is reasonable to assume that [criterion] will remain the same for both types of items" (p. 463).[1]

More recent work has revealed several exceptions to this pattern. First, it appears that participants may adjust their decision criterion within a single recognition test when items are associated with different retention intervals. For example, Singer and Wixted (2006; see also Singer, Gagnon, & Richards, 2002) had participants study items from different taxonomic categories, with categories studied either just prior to or up to 2 days before a recognition test. Although they did not observe differences in criterion at short delays (e.g., 20 or 40 min), they reported criterion shifts at delays of 2 days, as a more stringent criterion was applied to items from categories that were most recently studied.

Second, participants may shift their criterion within a single test list as a function of preexperimental familiarity. For example, Dobbins and Kroll (2005) presented participants with scenes from familiar (e.g., pictures from their own university) and unfamiliar (e.g., pictures from another university) locales, followed by a recognition test containing targets and distractors from each class of stimuli. Results showed that whereas "hits" (i.e., correct endorsements of targets) for familiar scenes exceeded those for unfamiliar scenes, there was no difference in FAs to distractors. This suggests that a more stringent criterion was used for familiar scenes than for unfamiliar scenes. That is, if the same criterion were used, one would expect that FAs to stronger, familiar scenes would exceed those to unfamiliar scenes. It is interesting that FAs to familiar scenes did exceed those to unfamiliar scenes when a deadline was imposed on recognition decisions, suggesting that criterion setting may be an effortful process.

Taken together, these data indicate that decision criterion can be altered or adjusted from item to item, but only in limited circumstances. The paucity of studies examining decision processes in memory suggests some caution in this conclusion, as the potential array of manipulations that may affect decision criterion, and possibly allow it to be altered dynamically, has certainly not been exhausted. In particular, those studies that have reported shifts in criterion from item to item have induced such shifts through some variation in item memorability (see e.g., Dobbins & Kroll, 2005; Rhodes & Kelley, 2003; Singer & Wixted, 2006). Thus, in the absence of obvious manipulations of item memorability, it is not clear whether participants are willing to shift their response criterion within a single test list.[2]

The Current Study

In the following experiments, we manipulated the probability that an item was previously studied. Previous work has shown that participants can adjust their responding in accord with differences in the base rate of studied items (e.g., Estes & Maddox, 1995; Healy & Kubovy, 1978), although these differences have only been observed when lists were tested separately. For example, Estes and Maddox (1995) reported that participants used a more liberal response criterion for recognition tests in which 67% of the items had been studied previously than they used for tests in which 33% of the items had been studied previously. This effect was obtained when stimuli were digits and letters (as opposed to words) and only when participants were given feedback on their performance. When feedback was withheld, participants did not adjust their criterion as a function of the base rate of studied items (see also Verde & Rotello, in press). Thus, feedback may be crucial to criterion shifts (an issue to which we return later).

In the current study, we varied base rates within a single list by correlating the probability of a tested item being old with the location in which it was presented. Specifically, words were presented for a recognition test in one of two locations on a computer screen. Words presented in one location on the screen were typically old, whereas words presented in another location were typically new. The order of presentation was random, as words appeared unpredictably in one location or in the other throughout the test. Thus, shifts in criterion could only occur through adjustments made from item to item. A shift in criterion would be evident if participants adopted a different criterion depending on which side of the screen a word was presented. Presumably, such a shift would lead participants' criterion to be more liberal for words presented on the side of the screen associated with predominantly old items and to be more conservative for words presented on the side of the screen associated with predominantly new items. That is, given that items on each side of the screen should be equally memorable, a shift in criterion would be apparent only if participants were sensitive to the association between prior presentation and item location. If participants were sensitive to this association, one would expect that the most liberal responding would be evident for the side in which the majority of items were old.

Participants in the current study were tested over four study–test blocks, allowing for an examination of changes in criterion across

[1] Brown et al. (1977) required participants to make recognition decisions in the form of confidence judgments corresponding to distinct categories (with extremes of *highly probable* and *highly improbable*). In their discussion of criterion, they suggested that for mixed-lists, ". . .category boundaries remain the same for both types of items" (p. 463). We have imputed the word "criterion" in place of "category boundaries" in the interest of clarity.

[2] There are manipulations that can induce criterion shifts without varying item memorability, though these generally have not occurred within a single list. For example, several researchers have shown that decision criterion is sensitive to instructions suggesting that there is either a high or low proportion of old items in a test list (e.g., Estes & Maddox, 1995; Healy & Kubovy, 1978; Hirshman & Henzler, 1998; Strack & Förster, 1995; Verfaellie, Giovanello, & Keane, 2001). In addition, criterion may also be influenced by test instructions to endorse only presented items, distractors related to a study list, or both presented items and related distractors (e.g., Brainerd, Wright, Reyna, & Mojardin, 2001; see also Benjamin & Bawa, 2004, for manipulations of the nature of distractors).

blocks (cf. Estes & Maddox, 1995; Healy & Kubovy, 1978). A change in criterion might occur if participants became aware of the structure of the test and altered their responding accordingly. It is not apparent whether awareness is necessary for criterion shifts. For example, Higham and Brooks (1997) had participants study lists that were created on the basis of an underlying structure of word frequency and part of speech. Results showed that participants became sensitive to the structure of the study lists and, consequently, enhanced their discriminability for items consistent with this structure without explicit awareness of the underlying structure. In contrast, Schunn, Lovett, and Reder (2001) reported that participants who were aware of changes in the base rate of correct solutions to problems were more sensitive to such changes. Prior work in the memory literature has either explicitly informed participants about the nature of the manipulation (e.g., Healy & Kubovy, 1978) or used clear differences in stimuli (e.g., Stretch & Wixted, 1998) and thus has little to say on the issue of awareness.

Given that participants were not explicitly informed of base rates, it is unclear whether they might become sensitive to this information and use it to inform recognition memory judgments. For example, the decision-making literature contains numerous demonstrations in which participants ignore base-rate information (e.g., Kahneman & Tversky, 1973; but see also Koehler, 1996). However, neglect of base-rate information contrasts with other reports of apparent sensitivity to frequency of occurrence (e.g., Hasher & Zacks, 1984). Holyoak and Spellman (1993) have accounted for this discrepancy by suggesting that participants are more likely to use base-rate information when such information is acquired through learning rather than presented explicitly. They further suggested that the use of base rates requires not only acquisition but access to base-rate information. From this perspective, participants in the current study may be able to acquire information about base rates but must have some access to this information to influence recognition decisions. If this was the case, one would expect that only participants who were explicitly aware of the correlation between base rate and location would exhibit differences in response criterion between predominantly old and predominantly new items. In contrast, participants who were unaware of the correlation between base rate and location would be expected to exhibit largely neutral responding.

To summarize, the current study examined response criterion in three experiments in which the probability that an item was studied was correlated with its location at test. In Experiment 1, we attempted to determine whether such a manipulation could induce shifts in criterion. In Experiment 2, we examined participants' awareness of the base-rate manipulation and also varied the method of inputting responses as a method of manipulating awareness. Experiments 1 and 2 replicated and extended results from an unpublished Ph.D. thesis by Dolan (1999). We discuss results reported by Dolan after reporting results from our experiments. In Experiment 3, we manipulated the nature of the feedback provided to participants after each test trial to examine its role in criterion shifts in recognition memory. As will be discussed later, one possibility is that feedback on recognition decisions allows participants to monitor their performance and to use information at test (such as the location of test items) to inform recognition decisions and thus respond optimally.

Experiment 1

In Experiment 1, we examined whether participants would adjust their decision criterion in response to the probability that items presented in different test locations were previously studied. Specifically, participants studied 72 items and were then given a recognition test for the 72 studied items and 72 unstudied distractors, with the procedure repeated over four unique study–test blocks. Each test item was randomly presented on either the left or right side of the screen. In one location, 67% of the items had been previously studied, whereas, in the other location, 33% of the items had been previously studied. If participants can shift their response criterion in accord with the probability that an item is old, one would expect that criterion should be more liberal for the location in which the majority (67%) of test items are old than for the location in which the minority (33%) of test items are old. Such data would suggest that criterion can operate dynamically, changing as a function of shifting text contexts.

Method

Participants. Twelve Washington University psychology students (6 women and 6 men) participated for course credit or pay ($10). Participants were tested individually.

Materials and design. Materials consisted of 612 nouns (mean frequency = 34.88, *SD* = 20.43; mean number of letters = 5.8, *SD* = 1.14; mean number of syllables = 1.77, *SD* = 0.69) taken from the Kucera and Francis (1967) norms. These were randomly divided into four sets of 144 items (further subdivided into eight sets of 72 items; 576 items in total) to serve as study and test items for each of four blocks, with the remaining items serving either as primacy or recency buffers (24 items) or as items for the practice phase of the experiment (12 items). For each set of 144 items, half of the items were studied and presented as old items at test, whereas the remaining half of the items served as distractors on the recognition test. Thus, the study list for each block consisted of 72 items, and the test list was made up of 144 items.

Test items were presented such that 67% (48 of 72) of the items on one side of the screen were old and 33% (24 of 72) of the items on the other side of the screen were old. To ensure balance, we further subdivided each set of 144 items for each block into six sets of 24 items, equated for frequency, number of letters, and number of syllables. Half of these sets (72 items) were presented as studied items. Of these three studied sets, two sets (48 items) were presented on the side with a studied base rate of 67%, and the remaining set (24 items) was presented on the side with a studied base rate of 33%. The other half of the 144-item set (72 items) made up the distractor list. Two of the distractor sets (48 items) were presented on the side with a studied base rate of 33%, and the remaining distractor set was presented on the side with a studied base rate of 67%. Items were counterbalanced across base rates and old–new status. The computer randomly selected which particular set would be presented within each block.

Test items were presented on either the far right or far left side of the screen, centered vertically. The side in which the majority of studied items was presented was counterbalanced such that items on the right side of the screen were predominantly old for half of the participants and items on the left side of the screen were predominantly old for the other half of the participants. For sim-

plicity, items presented at test on the side in which the majority of items were old will be termed *mostly old* items, whereas items presented on the side of the screen for which the majority were new will be termed *mostly new* items.

Procedure. All study and test stimuli were presented in white, lowercase letters in 30-point Arial font in the center of a black background on an IBM-compatible computer. After providing informed consent, participants began the practice phase of the experiment. They were first shown a list of six words presented at a 1-s rate with a 250-ms interstimulus interval, with instructions to remember these words for a forthcoming memory test. All study items were presented in a freshly randomized order for each participant. Immediately following the practice study list, participants were given instructions for the 12-item practice recognition test. They were informed that their memory for the preceding list would be tested and that each test item would appear on either the left or right side of the screen. Participants were instructed that for each item they were to press the key designated "old" (the *B* key) if the word had been studied or the key designated "new" (the *N* key) if the word had not been studied. Further, participants were informed that a running score, centered in the top portion of the screen, would be maintained to track their performance. Each correct answer was denoted with the feedback "+1," and 1 point was added to the overall score. Each incorrect answer was denoted with the feedback "−1," and 1 point was deducted from the overall score. Following this feedback, the next test trial appeared. As the practice test was only intended to familiarize participants with the general procedure, the base rate of studied items was equated (i.e., 50%) for each side of the screen. Participants were not informed of the probability that an item was old nor were they informed in subsequent test lists about the distribution of old and new items. All test items were presented in a freshly randomized order for each participant.

Following the practice phase of the experiment, participants began the first of four study–test blocks. Each block was identical to the practice phase, with the exception that participants were presented with longer study (72 item) and test (144 item) lists. In addition, each study list was preceded and followed by a buffer of three items intended to control for primacy and recency effects. Items from these buffers were not tested. For each block of the test phase, participants began with a running score of zero. Following the fourth and final block of test items, participants were debriefed and thanked for their participation. The experiment took approximately 45 to 60 min to complete.

Results

Recognition data are summarized in Table 1. Hit and FA rates were calculated for each block and were used to calculate measures of discriminability and criterion for each participant. In the interest of brevity, only analyses of signal detection estimates are reported.[3] Following Snodgrass and Corwin (1988), all hit and FA rates were first adjusted by adding 0.5 to each frequency and dividing by $N + 1$, where N is the number of trials for a particular type of item. All signal detection analyses are reported using d' and C as measures of discriminability and response criterion, respectively. The measure d' reflects the standardized difference between old and new distributions. The measure C calculates criterion on the basis of its distance from the intersection of the old and new distributions (i.e., $d' \div 2$) and is measured in standardized units ($C = z_{FA} - d' \div 2$). Neutral responding is indicated by a value of 0, with values above 0 indicative of conservative responding and values below 0 indicative of liberal responding. Aside from its direct relation to hits and FAs, the measure C has the added feature of requiring fewer assumptions about participants' knowledge of the familiarity of old and new distributions (Snodgrass & Corwin, 1988). However, the use of other signal detection measures (e.g., *Pr*, A', *Br*, B''_D) did not change the pattern of results reported in this or subsequent experiments. The alpha level was set to .05 for all statistical analyses.

Table 1
Means (and Standard Deviations) of Recognition Performance in Experiments 1–3

Condition	Hits	FA	d'	C
Experiment 1				
Mostly old	.75 (.08)	.31 (.15)	1.21 (.41)	−.07 (.27)
Mostly new	.69 (.08)	.28 (.13)	1.12 (.37)	.07 (.25)
Experiment 2				
Same keys				
Mostly old	.69 (.08)	.30 (.13)	1.08 (.48)	.03 (.24)
Mostly new	.62 (.13)	.25 (.08)	1.00 (.41)	.19 (.26)
Different keys				
Mostly old	.77 (.10)	.45 (.23)	0.93 (.52)	−.33 (.44)
Mostly new	.57 (.19)	.22 (.11)	1.02 (.54)	.32 (.36)
Experiment 3				
Block 1 & 2 feedback				
Mostly old	.71 (.09)	.33 (.16)	1.03 (.59)	−.04 (.25)
Mostly old	.62 (.14)	.29 (.12)	0.90 (.55)	.13 (.24)
Block 3 & 4 feedback				
Mostly old	.70 (.10)	.37 (.17)	0.90 (.52)	−.08 (.32)
Mostly old	.60 (.13)	.28 (.12)	0.87 (.51)	.18 (.25)

Note. FA = false alarm; d' = discriminability; C = response criterion.

Discriminability. Inspection of Table 1 indicates that discriminability (d') did not differ for items that were predominantly old (mostly old items) compared with those that were predominantly new (mostly new items). This was confirmed by a 2 (item type: mostly old, mostly new) × 4 (block: 1, 2, 3, 4) repeated measures analysis of variance (ANOVA) on mean discriminability estimates. Specifically, there was no main effect of item type, $F(1, 33) = 1.16$, $p = .30$, $\eta^2_p = .10$. Discriminability did not vary between blocks ($F < 1$), nor did block interact with item type ($F < 1$).

Response criterion. Analyses of mean response criterion estimates (using the same factors as the analysis of discriminability) showed that criterion (C) was sensitive to the probability that an item was old in a given location. In particular, participants' estimated response criterion was significantly more liberal for mostly old items than it was for mostly new items, $F(1, 33) = 9.44$, $\eta^2_p = .46$. Response criterion estimates also varied to some degree across blocks, $F(3, 33) = 3.25$, $\eta^2_p = .23$, but block did not interact with item type ($F < 1$).

[3] Analyses of hits and FAs are available on request from Matthew G. Rhodes.

Discussion

Results from Experiment 1 showed that presenting test items in contexts in which an item was typically old or typically new had a singular effect on recognition memory. Specifically, participants' criterion for responding was markedly more liberal when test items were presented in a context for which items were predominantly old. The change in response criterion was evident even though test items were presented randomly in one context or the other. These results replicate and extend those reported by Dolan (1999). Dolan correlated the base rate of studied items with the color in which test items were presented (instead of screen location) and found that participants' responding was sensitive to base rate.

The changes in response criterion evident in Experiment 1 might reflect a form of implicit learning (cf. Nissen & Bullemer, 1987; Reber, 1967; see Seger, 1994, for a review) or a more explicit basis for altering response criterion. Experiment 2 explores these possibilities.

Experiment 2

Experiment 1 demonstrated that participants can dynamically alter their response criterion from item to item in a recognition memory task as a function of different probabilities that a test item presented in a particular context was studied. However, it is unclear whether the change in response criterion is implicit and occurs without the participant's awareness or whether it reflects explicit awareness that the probability that an item was studied is correlated with test context. Presumably, awareness of this correlation would foster greater sensitivity to variations in base rate (cf. Holyoak & Spellman, 1993). In much of the prior work examining the influence of base rates on response criterion (e.g., Healy & Kubovy, 1978), participants were explicitly informed of base rates. Thus, there is little evidence regarding whether changes in criterion for recognition decisions can occur without awareness and whether awareness influences responding.

In Experiment 2, we examined this issue by administering a posttest questionnaire, taken from Dolan (1999), to assess awareness (see Appendix) immediately following completion of the final block of test items. In addition, Dolan reported that participants who were instructed to input their answers using different keys (depending on the color of a test item) were more likely to be aware of a manipulation of the base rate of studied items than were participants who used the same input keys regardless of the color of a test item. Thus, this manipulation of using the same or different keys to input answers at test was included in Experiment 2 as a method of manipulating awareness. The distribution of old and new items was also altered somewhat in Experiment 2 in the interest of maximizing possible effects. Specifically, 60 of 72 items in the mostly old condition were old, whereas the reverse relationship (i.e., 12 of 72 items were old) held for the mostly new condition. Therefore, 83% of the items presented on one side at test were old, whereas 17% of the items presented on the opposite side at test were old.

Method

Participants. Forty-eight Washington University psychology students (34 women, 14 men) participated for course credit or pay ($10). Participants were tested individually.

Materials and design. The materials used in Experiment 2 were identical to those used in Experiment 1. However, the set of 144 items used for each study–test block was subdivided into 12 sets of 12 items (instead of 6 sets of 24 items as in Experiment 1), which were equated for frequency, number of letters, and number of syllables. This was done to accommodate the distribution of mostly old (60 out of 72 items were old) and mostly new (12 out of 72 items were old) items used in Experiment 2. Thus, the study list consisted of six sets of 12 items (72 items). During the test, five of these sets (60 items) were presented on the mostly old side, whereas the remaining set (12 items) was presented on the mostly new side. The remaining six sets of 12 items were used as distractors on the recognition test. Five of these sets (60 items) were presented as distractors on the mostly new side with the remaining set (12 items) presented as distractors on the mostly old side. Each item was presented equally often as a mostly old or mostly new item and was also presented equally often on the left or right side.

Procedure. The procedure used in Experiment 2 was identical to that used for Experiment 1, with two exceptions. First, half of the participants were given instructions to use different response keys at test depending on which side of the screen a test item was presented. Specifically, participants were instructed that when test items were presented on the left side of the screen they should use keys on the left side of the keyboard to make their recognition decision, denoting an item as old (*A* key) or new (*S* key). For items presented on the right side of the screen, participants were likewise instructed to make their recognition decision of old (*K* key) or new (*L* key) using keys on the right side of the keyboard. The other half of the participants used the same input keys regardless of which side a test item was presented on and were given the instructions described in Experiment 1. All other aspects of the presentation, including the method of feedback, were identical to Experiment 1. The only other difference from Experiment 1 is that immediately following the fourth and final block of test items, participants were administered a questionnaire examining their knowledge of the experiment (see Appendix). Participants gave their answers orally as a research assistant wrote down their responses.

Results

Discriminability. Mean discriminability data (Table 1) were submitted to a 2 (item type: mostly old, mostly new) × 2 (input keys: same, different) × 4 (block: 1, 2, 3, 4) mixed-factor ANOVA. As in Experiment 1, discriminability did not differ between mostly old and mostly new items ($F < 1$). Discriminability also did not differ between participants who used the same or different input keys ($F < 1$), but a marginal main effect of block was present, $F(3, 138) = 2.28$, $p = .08$, $\eta^2_p = .05$. Block did not interact with input keys, $F(3, 138) = 1.01$, $p = .39$, $\eta^2_p = .02$, or with item type ($F < 1$). In addition, the triple interaction of block, item type, and input keys was not reliable, $F(3, 138) = 1.41$, $p = .24$, $\eta^2_p = .03$. However, a marginal Item Type × Input Keys interaction was evident, $F(1, 46) = 3.20$, $p = .08$, $\eta^2_p = .07$. This reflects the fact that discriminability was numerically, but not reliably, better for mostly old items compared with mostly new items in the same keys condition, $F(1, 23) = 2.00$, $p = .17$, $\eta^2_p = .08$. For participants using different keys, the opposite pattern was evident, as discriminability was somewhat, but not reliably, better for mostly new items compared with mostly old items, $F(1, 23) = 1.35$, $p = .26$, $\eta^2_p = .06$.

Response criterion. Mean response criterion estimates (see Table 1) were analyzed using the same factors as the analysis of discriminability and were plotted across blocks in Figure 1. These data indicated that responding was considerably more liberal for mostly old items than it was for mostly new items, $F(1, 46) = 25.02$, $\eta^2_p = .35$. The main effect of block was not reliable ($F < 1$), but a reliable main effect of input keys, $F(1, 46) = 3.93$, $p = .05$, $\eta^2_p = .08$, was present. More important, several significant interactions were evident. In particular, item type interacted with block, $F(3, 138) = 9.36$, $\eta^2_p = .17$, and with input keys, $F(3, 138) = 9.44$, $\eta^2_p = .17$. These interactions are subsumed by a significant three-way interaction of item type, block, and input keys, $F(3, 138) = 8.37$, $\eta^2_p = .15$. This reflects the fact that, in the same keys condition, the difference in response bias for mostly old items compared with mostly new items was largest in Blocks 1 and 4, ts(23) $\geq$ 2.39, Cohen's $d > .62$, and was smaller in the second and third blocks, $t(23) = 1.91$, $p = .07$, Cohen's $d = .43$, and, $t(23) = 1.36$, $p = .19$, Cohen's $d = .38$, respectively. A different pattern was evident for the different keys condition. In that condition, the difference in estimated response criterion between mostly old and mostly new items became progressively larger across blocks. For example, the effect size (i.e., Cohen's d) for the difference in estimated criterion for mostly old items versus mostly new items was 0.76 in Block 1, 0.90 in Block 2, 1.47 in Block 3, and 1.91 in Block 4 (ts $\geq$ 2.60, ps $\leq$ .02).

As these data also indicate, the difference in response criterion between mostly old and mostly new items, although robust across all conditions, was larger in the different keys condition compared with the same keys condition. For example, in the same keys condition, the effect size for the difference in estimates response criterion in the fourth test block was .62, whereas the comparable value for the different keys condition was 1.91. Such differences may be related to awareness, which we examine next.

Awareness data. Participants were classified as aware or unaware using a questionnaire administered immediately after the final test block, with classifications made on the basis of responses to the first eight questions. We classified participants as aware if they explicitly and accurately described the distribution of test items or their basis for responding in a manner that corresponded to the distribution of test items. For example, in response to the question, "What thoughts did you have while performing the test as to the purpose of having the test items on the left or right side of the screen?", participants classified as aware gave answers such as "Got used to having old on left and new on right" or "The left side was generally old and the right side was generally new." In response to the question, "Did the side of the test word have any influence on your responding?", aware participants gave answers such as "If I was unsure and the word was on the left I was more likely to respond old" or (when mostly old items were on the right) "If it was on the right and I didn't know it, I would push old and if it was on the left I would push new." Classifications of aware or unaware were made independently by one of the authors (Matthew G. Rhodes) and a research assistant, with agreement on 45 of 48 (94%) cases. Cases in which there was a disagreement were resolved through discussion. Results showed that 6 of 24 participants (25%) in the same keys condition were classified as aware, whereas 17 of 24 participants (71%) in the different keys condition were classified as aware. A chi-square test confirmed that participants in the different keys condition were significantly more likely to be aware of the correlation between old–new status and test context than were participants in the same keys condition, $\chi^2(1, N = 48) = 10.10$, $p < .01$.

Recognition data broken down by awareness are presented in Table 2, with estimated response criterion depicted across blocks in Figure 2 for the same (top panel) and different (bottom panel) keys conditions. Inspection of Figure 2 suggests that participants classified as aware of the manipulation of base rate exhibited more liberal responding than those participants classified as unaware. In addition, differences between aware and unaware participants were greatest when participants used different keys to input their responses. This was

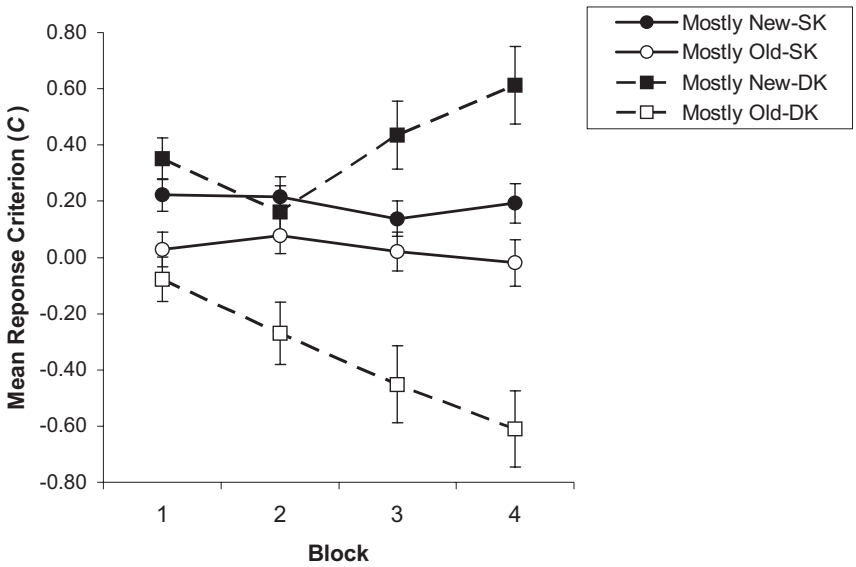

Figure 1. Mean estimated response criterion estimates by block and item type in Experiment 2. Error bars represent standard error. SK = same input keys; DK = different input keys

Table 2
Means (and Standard Deviations) of Recognition Performance by Awareness in Experiment 2

Condition	Hits	FA	d'	C
Same keys-aware (n = 6)				
Mostly old	.72 (.06)	.31 (.13)	1.11 (.41)	−.03 (.19)
Mostly new	.59 (.09)	.25 (.07)	0.93 (.32)	.23 (.16)
Same keys-unaware (n = 18)				
Mostly old	.68 (.09)	.30 (.13)	1.07 (.51)	.05 (.26)
Mostly new	.63 (.15)	.25 (.09)	1.03 (.44)	.18 (.28)
Different keys-aware (n = 17)				
Mostly old	.79 (.11)	.49 (.26)	0.93 (.60)	−.42 (.50)
Mostly new	.52 (.20)	.18 (.09)	1.02 (.61)	.45 (.32)
Different keys-unaware (n = 7)				
Mostly old	.72 (.05)	.37 (.09)	0.93 (.32)	−.13 (.12)
Mostly new	.69 (.09)	.31 (.10)	1.01 (.33)	.00 (.22)

Note. FA = false alarm; d' = discriminability; C = response criterion.

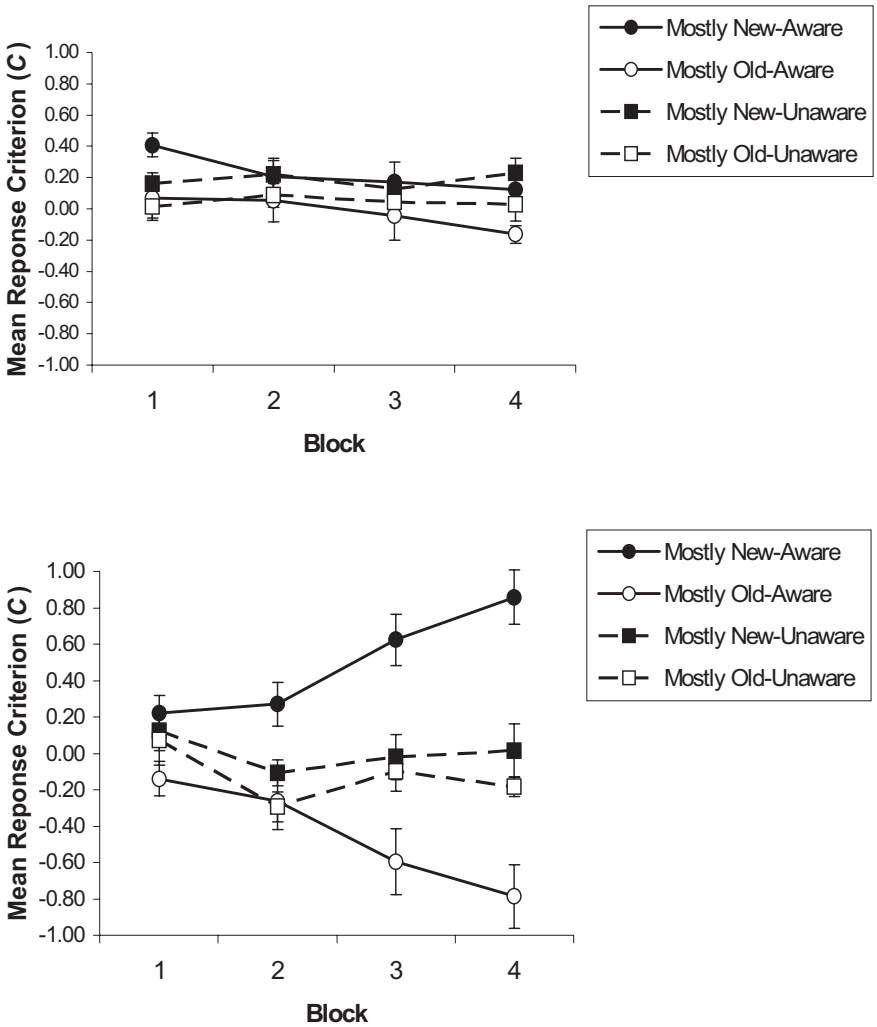

Figure 2. Mean estimated response criterion estimates by block and item type in Experiment 2 for participants classified as aware and unaware. Top panel: Data for participants who used the same input keys during the recognition test. Bottom panel: Data for participants who used different input keys at test. Error bars represent standard error.

confirmed in a 2 (item type: mostly old, mostly new) × 4 (block: 1, 2, 3, 4) × 2 (input keys: same, different) × 2 (awareness: aware, unaware) mixed-factor ANOVA.[4] In the interest of brevity, we do not report results for all 15 main effects and interactions and instead focused on only those higher order interactions that were reliable, particularly as they pertain to the impact of awareness on response criterion. Results showed that awareness interacted with item type, $F(1, 46) = 6.45$, $\eta^2_p = .13$. Specifically, unaware participants exhibited a significantly more liberal response criterion for mostly old items ($M = -.04$, $SE = .08$) than they did for mostly new items ($M = .09$, $SE = .07$), $F(1, 23) = 5.54$, $\eta^2_p = .19$. Participants classified as aware likewise exhibited significantly more liberal responding for mostly old items ($M = -.23$, $SE = .09$) than they did for mostly new items ($M = .36$, $SE = .07$), $F(1, 21) = 11.00$, $\eta^2_p = .34$, but with differences of a greater magnitude than those evident for unaware participants. A marginal triple interaction of item type, input keys, and awareness was also evident, $F(1, 44) = 3.83$, $p = .06$, $\eta^2_p = .08$. This stems from the fact that, in the same keys condition, estimated response criterion did not differ between aware and unaware participants for either mostly old items or mostly new items ($ts < 1$). In contrast, aware participants ($M = -.45$, $SE = .09$) in the different keys condition exhibited more liberal responding to mostly old items than did unaware participants ($M = -.13$, $SE = .14$), a difference that was marginally reliable, $t(22) = 1.89$, $p = .07$, Cohen's $d = .89$. Likewise, aware participants' ($M = .49$, $SE = .07$) estimated criterion for mostly new items in the different keys condition was significantly more conservative than that exhibited by unaware participants ($M = .003$, $SE = .07$), $t(22) = 3.62$, Cohen's $d = 1.70$.

Several other interactions were also present. For example, a reliable Block × Input Keys × Awareness interaction was evident, $F(3, 132) = 3.66$, $\eta^2_p = .08$. This occurred because, for aware participants, overall response criterion estimates in each block did not vary between the same and different keys conditions ($ts \leq 1.68$, $ps \geq .11$, Cohen's $d < .87$). However, for unaware participants, overall responding was reliably more liberal when different rather than the same response keys were used in both the second and fourth blocks, $t(23) = 2.76$, Cohen's $d = 1.33$, and, $t(23) = 2.76$, $p = .05$, Cohen's $d = .99$, respectively. Of greater importance, a significant triple interaction of block, item type, and awareness, $F(3, 132) = 3.26$, $\eta^2_p = .07$, was present. In particular, estimated criterion did not reliably differ between aware and unaware participants for mostly old ($ts < 1$) and mostly new ($ts \leq 1.66$, $ps \geq .10$, Cohen's $d < .73$) items over the first two blocks of testing. However, aware participants were significantly more liberal in their responding to mostly old items than were unaware participants in the third and fourth test blocks, $t(46) = 1.76$, $p = .08$, Cohen's $d = .52$, and, $t(46) = 2.38$, Cohen's $d = .70$, respectively. In addition, aware participants were more conservative in their responding to mostly new items than were unaware participants over the third and fourth test blocks, $ts(46) \geq 2.48$, Cohen's $d > .73$. A significant (but largely uninterpretable) four-way interaction of item type, input keys, block, and awareness, $F(3, 132) = 3.43$, $\eta^2_p = .07$, was also present.

Discussion

Consistent with Experiment 1, results from Experiment 2 demonstrated that participants' response criteria were sensitive to the correlation between old–new status and test context (i.e., the location of the test item), whereas discriminability was largely unaffected. Further, results showed that the difference in estimated response criterion for items that were predominantly old compared with those that were predominantly new was largest when different input keys were used (cf. Dolan, 1999). This may be related to awareness. That is, participants were more likely to exhibit awareness of the correlation between old–new status and test context when different response keys were used. In turn, participants who were aware of the manipulation exhibited a larger difference in estimated response criterion for mostly old versus mostly new items than did participants who appeared to be unaware. However, because awareness was not explicitly manipulated, this conclusion should be treated with some caution.

Given that a greater number of participants in the different keys condition were explicitly aware of the correlation between base rate and location, it appears that responding with different input keys aided the learning of base-rate information. This may reflect the degree of compatibility between items presented for recognition and the method of inputting answers (cf. Fitts & Seeger, 1953; Kornblum, Hasbroucq, & Osman, 1990). For example, responses were made on either the left or right side of the keyboard and items were presented on the left or right side of the screen. Such compatibility may have facilitated the acquisition of base-rate information.

In addition, several participants in Experiment 2 reported that because the *old* key was on the left side of the keyboard (with respect to the *new* key), the majority of items presented on the left side of the screen must have been old. As mostly old items were presented on the left side of the screen for only half of the participants tested, this conclusion would have been valid for only those participants. However, to ensure that the patterns of data reported are robust with respect to which side of the screen mostly old items are presented in, we presented test items in Experiment 3 (see below) in either the top or bottom portion of the screen instead of on the left or right side.

In all, results from Experiments 1 and 2 demonstrated that participants can adjust their criterion for recognition decisions within a single test list. Such data must be reconciled with evidence from other reports that participants are often unwilling to make such adjustments (Morrell et al., 2002; Stretch & Wixted, 1998). One possibility is that feedback on performance plays an important role in shifts in response criterion. We examined this issue in Experiment 3 by manipulating feedback at test.

Experiment 3

Previous work has suggested that feedback may be important for criterion shifts. For example, Estes and Maddox (1995) reported that shifts in response criterion were evident only when feedback was provided to participants in a recognition experiment with different base rates of old items. Verde and Rotello (in press) have likewise reported that feedback may be crucial to criterion shifts. In particular, they observed criterion shifts as a function of differ-

[4] Analyses of discriminability using the same factors indicated that discriminability did not differ as a function of awareness ($F < 1$). In addition, awareness did not reliably interact with any other variables ($ps > .29$).

entially strengthened test items only when participants were provided with feedback on the accuracy of their responses. However, Dobbins and Kroll (2005) observed criterion shifts that were based on the familiarity of different classes of items and that occurred without feedback.

Is feedback necessary for participants to exhibit shifts in response criterion in the current study? The experiments reported thus far do not allow for any conclusions. We examined the role of feedback in Experiment 3 by manipulating the presence of feedback. Specifically, half of the participants in Experiment 3 received feedback for the first two test blocks but did not receive feedback for the final two test blocks. In contrast, the other half of the participants in Experiment 3 did not receive feedback for the first two blocks of test items but did receive feedback for the final two blocks of test items. If feedback is irrelevant to shifts in criterion, one would expect to find almost identical patterns of data across the two conditions. However, feedback may be important for criterion shifts in the current study, possibly because it facilitates learning about the distribution of old and new items and thus allows participants to respond optimally (e.g., with a liberal criterion for predominantly old items). If this is the case, one would predict that participants given feedback in the first two blocks should continue to adjust their criterion depending on the context of the test item (e.g., more liberal responding to mostly old items) even when feedback is removed for the final two blocks of testing. Likewise, if feedback is necessary to learn about the distribution of old items, participants given feedback in the final two blocks should only exhibit a shift in criterion in those final test blocks.

It must also be noted that, as described previously, test items in Experiment 3 were presented in the top or bottom portion of the screen rather than on the left or right side. This was done as several participants in Experiment 2 used the apparent symmetry between the position of the response keys (with the *old* key on the left) and that of the test items to infer the distribution of studied items. Finally, given that the strongest effects on response criterion in Experiment 2 were evident when participants used different input keys, only this condition was tested.

Method

Participants. Forty-eight Washington University psychology students (27 women, 21 men) participated for course credit or pay ($10). Participants were tested individually.

Materials and design. The materials used in Experiment 3 were identical to those used in Experiment 2.

Procedure. The procedure used in Experiment 3 was identical to that used for the different input keys condition of Experiment 2 with two exceptions. First, test items were presented in a different manner. Specifically, items were centered horizontally in either the top or bottom portion of the screen instead of on the left or right side, as in Experiments 1 and 2. Because items were presented in the top or bottom portion of the screen, the running score used when feedback was implemented was placed in the center of the screen rather than at the top as in previous experiments. Second, the presence of feedback was manipulated. In particular, half of the participants were given feedback for the first two test blocks in the manner used in previous experiments. However, for the third and fourth test blocks, feedback was removed. When this occurred, participants were informed that the next test item would appear immediately following their recognition decision. For the other half of the participants, feedback was withheld for the first two test blocks, with the next test item appearing immediately after a participant's response. Feedback was then provided for the final two test blocks in the manner described for previous experiments. Participants in both conditions were given four practice trials prior to beginning the third block of testing for familiarization with the new procedure. The items for this practice trial consisted of two distractors, one item studied in the recency buffer, and one item studied in the primacy buffer.

Results

Discriminability. Mean estimates of discriminability (see Table 1) were submitted to a 2 (item type: mostly old, mostly new) × 4 (block: 1, 2, 3, 4) × 2 (feedback: Blocks 1 & 2, Blocks 3 & 4) mixed-factor ANOVA. Results showed that there was a main effect of block, $F(3, 138) = 2.98$, $\eta^2_p = .06$, in addition to a Feedback × Block interaction, $F(3, 138) = 3.17$, $\eta^2_p = .06$. This occurred because, for participants given feedback in the first two blocks, discriminability was better in the first block ($M = 1.20$, $SE = .11$) than in the proceeding blocks. In contrast, for participants given feedback in the third and fourth blocks, discriminability was relatively stable across the first two blocks ($M = .97$, $SE = .11$) and declined somewhat for the last two blocks ($M = .82$, $SE = .13$). Results also revealed a main of effect of item type, $F(1, 46) = 4.15$, $\eta^2_p = .08$, as discriminability was somewhat better for mostly old items ($M = .97$, $SE = .08$) than for mostly new items ($M = .90$, $SE = .08$). Item type did not interact with feedback, $F(1, 46) = 1.07$, $p = .31$, $\eta^2_p = .02$, or with block ($F < 1$). In addition, the triple interaction of item type, block, and feedback was not reliable, $F(3, 138) = 1.83$, $p = .15$, $\eta^2_p = .04$.

Response criterion. Mean estimates of response criterion (see Table 1) are plotted across blocks in Figure 3 and were submitted to an analysis using the same factors as the analysis of discriminability. These data showed that estimated response criterion was significantly more liberal for mostly old items ($M = -.07$, $SE = .04$) than it was for mostly new items ($M = .16$, $SE = .04$), $F(1, 46) = 29.66$, $\eta^2_p = .39$. Mean estimated criterion did not vary across blocks, $F(3, 138) = 1.49$, $p = .22$, $\eta^2_p = .03$, but a Block × Item Type interaction was present, $F(3, 138) = 5.01$, $\eta^2_p = .10$.

Of greater importance, inspection of Figure 3 shows that feedback was an important determinant of response criterion, as reflected by a significant triple interaction of item type, block, and feedback, $F(3, 138) = 4.94$, $\eta^2_p = .10$. In particular, participants given feedback in the first two blocks exhibited a significantly more liberal criterion for responding to mostly old items than they did for mostly new items by the second block, $t(23) = 4.12$, Cohen's $d = 1.21$. Once feedback was removed for the final two blocks, participants in that condition initially exhibited a small difference in response criterion that was not reliable in Block 3, $t(23) = 1.38$, $p = .18$, Cohen's $d = .32$, but was reliable by Block 4, $t(23) = 2.27$, Cohen's $d = .44$. Participants given feedback in Blocks 3 and 4 exhibited a different pattern, as differences in criterion became progressively larger across blocks. Specifically, the difference in estimated response criterion for mostly old items compared with mostly new items was not reliable in the first block of testing, $t(23) = 1.55$, $p = .14$, Cohen's $d = .30$, but was reliable by the second block of testing, $t(23) = 3.22$, Cohen's $d = .46$.

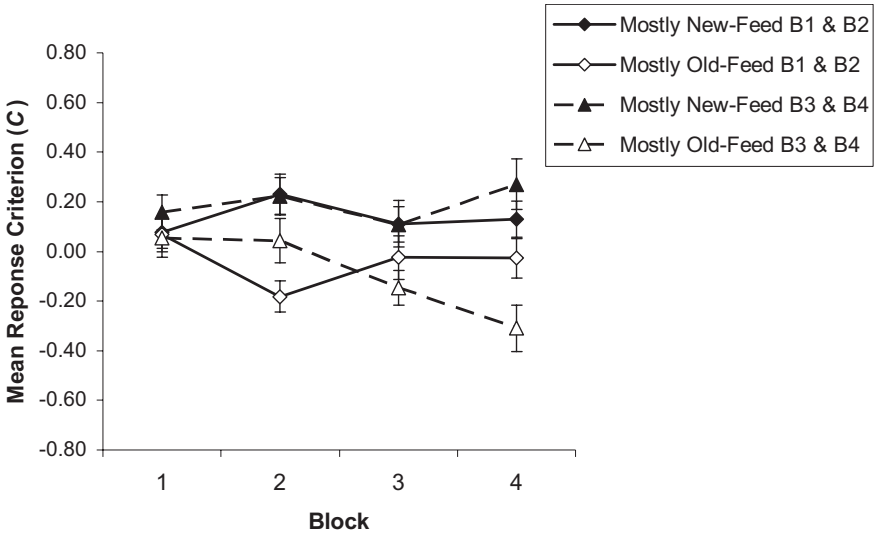

Figure 3. Mean estimated response criterion estimates by block and item type in Experiment 3. Error bars represent standard error. Feed = Feedback; B = Block.

When feedback was introduced, a reliable difference was evident in both the third, $t(23) = 2.45$, Cohen's $d = .77$, and fourth blocks, $t(23) = 3.58$, Cohen's $d = 1.30$. Thus, feedback appears to play a critical role in the robust criterion shifts exhibited by participants (cf. Estes & Maddox, 1995; Verde & Rotello, in press), as the presence of feedback was generally associated with larger effect size differences. However, one might presume that removing feedback would have a minimal impact on the performance of participants who were aware of the manipulation. We consider that issue next.

Awareness data. We classified participants as aware or unaware of the manipulation of base rate by using the same method described in Experiment 2. Results showed that 11 of 24 participants (42%) who received feedback in Blocks 1 and 2 were classified as aware, whereas 15 of 24 participants (63%) who received feedback in Blocks 3 and 4 were classified as aware. The number of aware participants did not differ between the two feedback conditions, $\chi^2(1, N = 48) = 1.34$, $p = .25$. Recognition data for participants (broken down by awareness) are presented in Table 3, with mean estimated response criterion depicted in Figure 4 for participants given feedback in the first two (top panel) and final two (bottom panel) blocks.

Mean response criterion estimates were subjected to a 2 (item type: mostly old, mostly new) × 4 (block: 1, 2, 3, 4) × 2 (feedback: Blocks 1 & 2, Blocks 3 & 4) × 2 (awareness: aware, unaware) mixed-factor ANOVA.[5] As in the previous analysis that examined awareness, we primarily report only those main effects or interactions that pertain to awareness. Inspection of awareness data shows that responding varied to a greater extent on the basis of whether feedback was present rather than whether participants were aware or unaware of the manipulation of base rate. For example, participants who were unaware of the manipulation exhibited a significantly more liberal response criterion for mostly old items ($M = -.06$, $SE = .06$) than they did for mostly new items ($M = .10$, $SE = .10$), $F(1, 20) = 9.49$, $\eta^2_p = .32$. Like the unaware participants, aware participants' criterion for responding to mostly old items ($M = -.06$, $SE = .06$) was significantly more liberal than their criterion for responding to mostly new items ($M = .21$, $SE = .05$), $F(1, 24) = 18.04$, $\eta^2_p = .43$. However, awareness did not interact with feedback ($F < 1$), nor did awareness interact with block, $F(3, 132) = 1.40$, $p = .25$, $\eta^2_p = .03$, or item type, $F(3, 132) = 1.74$, $p = .19$, $\eta^2_p = .04$. This pattern also held for all other higher order interactions (Fs ≤ 2.01, ps ≥ .12). Thus, although aware participants exhibited a somewhat larger effect size difference in response criterion, awareness had little impact on estimated response criterion.

Discussion

Results from Experiment 3 showed that feedback plays an important role in the criterion shifts reported. In particular, participants were more likely to exhibit large differences in estimated response criterion, with criterion being more liberal for mostly old items than for mostly new items when feedback was present. When feedback was absent, a much weaker trend remained for participants to respond more liberally to predominantly old items. It is interesting that this pattern of data was also present for participants classified as being aware of the base-rate manipulation. This may suggest that such participants are able to distinguish between various sources of information (e.g., location of test item, feed-

[5] Analyses of discriminability using the same factors indicated that discriminability did not differ as a function of awareness ($F < 1$). However, a marginally reliable Feedback × Awareness interaction was present, $F(1, 44) = 4.00$, $p = .05$, $\eta^2_p = .08$. This reflects the fact that discriminability was numerically, but not reliably, poorer for unaware ($M = 0.82$) compared with aware ($M = 1.17$) participants when feedback was given in the first two blocks, $t(46) = 1.63$, $p = .11$, Cohen's $d = .49$. In contrast, unaware participants' ($M = 1.05$) discriminability was numerically, but not reliably, better than that of aware participants ($M = 0.80$) when feedback was given in the final two blocks, $t(46) = 1.20$, $p = .24$, Cohen's $d = .36$. Awareness did not interact with any other variables (ps > .15).

Table 3
Means (and Standard Deviations) of Recognition Performance by Awareness in Experiment 3

Condition	Hits	FA	d'	C
Feedback Blocks 1 & 2-aware (n = 11)				
Mostly old	.73 (.08)	.29 (.15)	1.22 (.55)	−.01 (.26)
Mostly new	.63 (.14)	.25 (.12)	1.08 (.54)	.18 (.29)
Feedback Blocks 1 & 2-unaware (n = 13)				
Mostly old	.69 (.09)	.37 (.17)	0.87 (.60)	−.07 (.24)
Mostly new	.61 (.14)	.33 (.11)	0.75 (.54)	.10 (.21)
Feedback Blocks 3 & 4-aware (n = 15)				
Mostly old	.69 (.10)	.40 (.19)	0.80 (.46)	−.10 (.37)
Mostly new	.56 (.13)	.27 (.10)	0.77 (.40)	.23 (.24)
Feedback Blocks 3 & 4-unaware (n = 9)				
Mostly old	.71 (.10)	.32 (.14)	1.07 (.59)	−.05 (.23)
Mostly new	.66 (.12)	.29 (.15)	1.02 (.66)	.10 (.26)

Note. FA = false alarm; d' = discriminability; C = response criterion

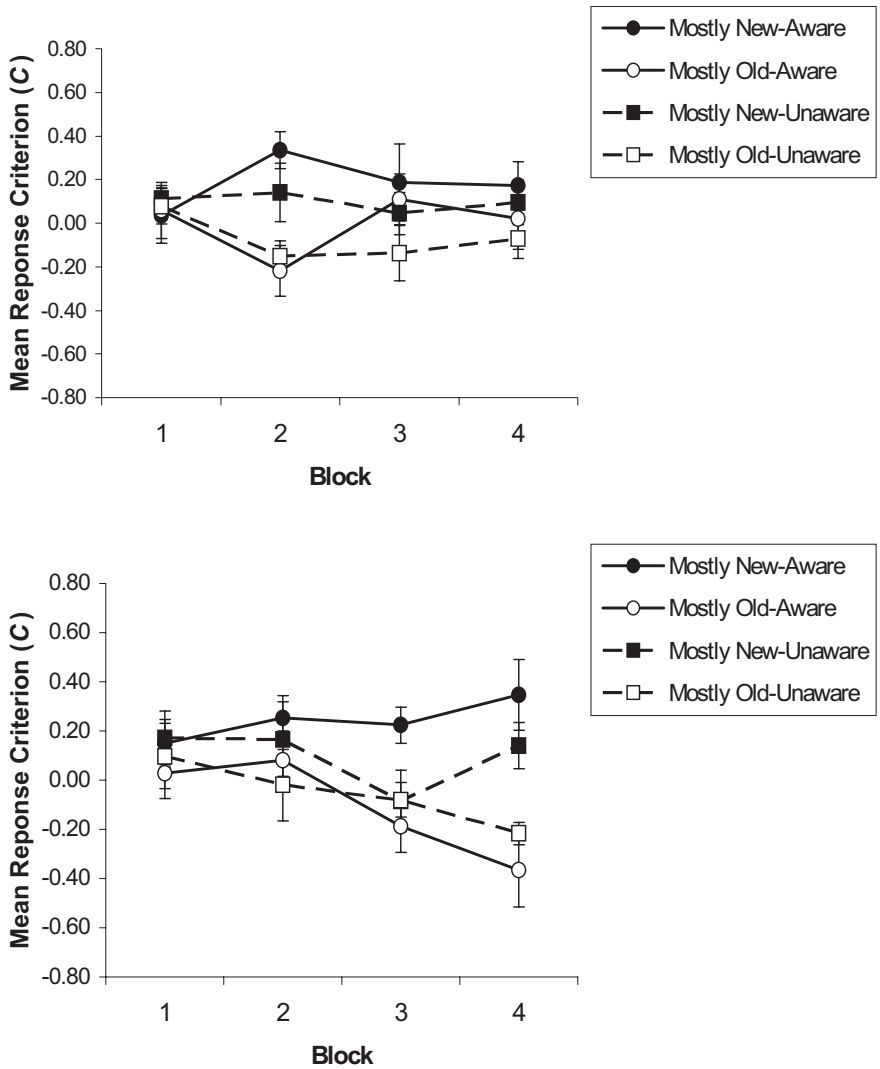

Figure 4. Mean estimated response criterion estimates by block and item type in Experiment 3 for participants classified as aware and unaware. Top panel: Data for participants given feedback in Blocks 1 and 2 during the recognition test. Bottom panel: Data for participants given feedback in Blocks 3 and 4 during the recognition test. Error bars represent standard error.

back) and use that information when appropriate in order to respond optimally. We examine these and other issues in the General Discussion.

General Discussion

The current study investigated decision processes in recognition memory, operationalized as the criterion the rememberer sets for endorsing a test item as previously studied (Macmillan & Creelman, 1990). Results from all three experiments demonstrated that participants can shift their criterion dynamically when features of the test context are correlated with the probability that an item was studied. Such a finding is relatively unique (but see Dobbins & Kroll, 2005; Singer & Wixted, 2006), as participants often appear unwilling to continuously shift their response criterion within a single test list (e.g., Morrell et al., 2002; Stretch & Wixted, 1998).

The experiments also highlight two factors that influence the magnitude and likelihood of criterion shifts in recognition memory judgments. First, participants appeared more likely to exhibit shifts (or at least stronger shifts) in their decision processes when they were explicitly aware of the bases for doing so. For example, participants in Experiment 2 who were aware of the manipulation ($\eta^2_p = .34$) demonstrated a larger effect size difference in estimated response criterion between mostly old and mostly new items than did participants who were unaware ($\eta^2_p = .19$). Differences based on awareness were particularly strong when different keys were used to input responses. This may reflect the degree of stimulus–response (S–R) compatibility (cf., Kornblum et al., 1990) fostered by presenting items on the left or right side of the screen and likewise by requiring participants to input their responses on the left or right side of the keyboard. When items were presented in the top and bottom portion of the screen in Experiment 3, the effects on response criterion were not of the magnitude apparent for participants in Experiment 2 who used different input keys. Thus, the degree of S–R compatibility may influence the nature or degree of awareness that participants exhibit. At the extreme, high levels of S–R compatibility may foster sensitivity to base rates even in the absence of feedback. Though not tested in the current study, it remains an important question for future research.

Data from Experiment 3 are also important in that they highlight the role of the second important factor in shifts in criterion: feedback. Specifically, when participants in Experiment 3 were given trial-by-trial feedback on their performance, they used a substantially more liberal response criterion for items from predominantly old contexts than they did for items from predominantly new contexts. However, when feedback was removed, even after some participants had already completed two blocks (288 trials) with feedback, the difference in response criterion for mostly old versus mostly new items was markedly diminished. Thus, feedback appears to be an important factor in the criterion shifts reported.

Bases for Criterion Shifts

Results from the current study contrast with previous reports that participants appear highly unlikely to exhibit shifts in response criterion on an item-by-item basis (e.g., Morrell et al., 2002; Stretch & Wixted, 1998). Data from Experiment 3 are perhaps useful in accounting for this discrepancy as they showed that removing feedback diminished differences in estimated response criterion between mostly old and mostly new items (cf. Estes & Maddox, 1995; Verde & Rotello, in press). This suggests that feedback may serve to highlight those dimensions of the test that are predictive of whether an item is old and may influence decision processes by emphasizing that dimension. For example, optimal responding would require that participants use a more liberal decision criterion (i.e., *C* less than 0) when items are predominantly old and use a more conservative criterion (i.e., *C* greater than 0) when items are predominantly new. Inspection of Figure 3 shows that this pattern was obtained only when feedback was present. In the absence of feedback, participants' responding approached neutrality (i.e., *C* equal to 0), a finding evident even for those participants classified as aware of the base-rate manipulation. Whether this reflects the difficulty of monitoring base rates with continuously changing test contexts or a lack of motivation in the absence of feedback is not clear at present.[6]

However, data reported for Experiment 3 do indicate that, even without feedback, a small difference in estimated response criterion remained between mostly old and mostly new items. What accounts for this difference? One possibility is that, without feedback, participants were only sensitive to base rates associated with the test context when context remained the same for several test trials. Alternatively, criterion change might require an explicit change in context. That is, participants might have only attended to base rates when there was a context change. Given that test items appeared in one context or the other randomly, participants would have experienced many cases in which test items appeared consecutively in one context (i.e., "no switch" in context) or in which test items switched back and forth between contexts (i.e., a "switch" in context). Therefore, analyses of such trials may shed light on other bases for criterion shifts.

To examine the possibility that context change impacted criterion, we reanalyzed data from Experiment 3 and classified trials on the basis of whether they were preceded by at least one trial in the same context (no-switch trials) or involved a switch from one context to another (switch trials).[7] Identical analyses undertaken for Experiments 1 and 2 showed that the difference in estimated response criterion for mostly old versus mostly new items did not vary on the basis of the type of trial (switch or no switch). Likewise, for blocks in which feedback was provided in Experiment 3, no difference in response criterion appeared as a function of the type of trial. However, when analyzed only for blocks in which feedback was absent (collapsed across feedback conditions), a different pattern appeared (see Figure 5). Specifically, when feedback was absent, the difference in estimated response criterion for mostly old versus mostly new items was greater for switch trials in comparison with no-switch trials. For example, the average estimated response criterion for no-switch trials was −.06 for mostly old items and .06 for mostly new items, $F(1, 46) = 5.84$,

[6] We thank two anonymous reviewers for suggesting these possibilities.

[7] For participants given feedback in Blocks 1 and 2, 45.7% of the trials were switch trials, and 53.6% of the trials were no-switch trials. For participants given feedback in Blocks 3 and 4, 44.4% of the trials were switch trials, and 54.9% of the trials were no-switch trials. Trials that started a block were excluded from the analyses (0.7% of trials in both conditions).

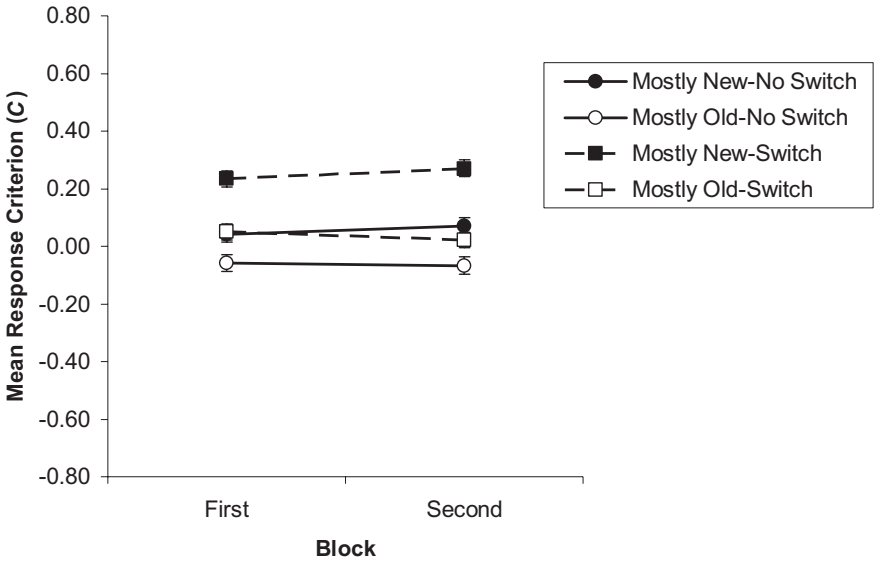

Figure 5. Mean estimated response criterion estimates for blocks in which feedback was withheld (collapsed across feedback conditions) in Experiment 3. Trials were classified on the basis of whether there was a change in context (switch trials) and/or test context stayed the same (no-switch trials). Error bars represent standard error.

η^2_p = .11. On switch trials, the comparable means are .04 for mostly old items and .25 for mostly new items, $F(1, 46) = 21.80$, η^2_p = .32. Thus, evaluated in terms of effect size, participants exhibited almost three times the difference in estimated criterion between mostly old and mostly new items for switch compared with no-switch trials. These data indicate that shifts in criterion were significantly more likely to occur or were simply stronger in response to a specific change in test context. Therefore, when feedback is present (as in Experiments 1 and 2, and blocks of Experiment 3), participants may chronically attend to test context as a dimension that informs recognition decisions and facilitates optimal responding. In the absence of feedback, test context may exert greater influence when there is an explicit change in context.

If participants do seek to maximize performance when given feedback, such feedback may be sufficient for item-to-item shifts in response criterion to occur using the methodology of Stretch and Wixted (1998; see also Morrell et al., 2002). Participants in their experiments studied items that were differentially strengthened. In particular, items in one color were presented five times (strong items), and items in a different color were presented once (weak items). Given that these colors were also used at test, it seems unlikely that participants were not explicitly aware of which colors were indicative of five presentations and which were indicative of only one presentation. However, Stretch and Wixted reported (in their Experiments 4 and 5) that criterion shifts were nonexistent. In contrast, using a similar procedure, Verde and Rotello (in press) observed shifts in criterion based on strength (manipulated by varying the number of times items were studied) when participants were provided with feedback on the accuracy of responses. Providing feedback might have the effect of emphasizing differences in the number of presentations. Such a conclusion suggests that, in the absence of feedback, participants do not explicitly attend to this information. For example, the study phase of Stretch and Wixted's experiments was ostensibly a manipulation of frequency of presentation, but the recognition test called for the same decision regardless of whether an item had been studied once or five times. That is, participants only needed to make a determination of whether an item had been presented and not a decision based on the number of study presentations. Therefore, establishing a single criterion for whether an item had simply been presented would be a reasonable strategy.

A similar explanation can be extended to data recently reported by Singer and Wixted (2006; see also Singer et al., 2002). They observed criterion changes in a single test when items from different taxonomic categories were studied either just prior to or 2 days before a recognition test but did not observe such effects when the delay occurred within a single session. As Singer and Wixted note, at shorter delays there may generally be a greater degree of contextual overlap between the study and test phases, or participants may presume that all items will be equally memorable. Consequently, participants may not attend to the retention interval at shorter delays. With a substantial delay, category becomes an important dimension to consider for recognition judgments, denoting items that were recently studied (and should be quite familiar) or studied sometime previously. Given evidence that predictions of future memory performance can be sensitive to different retention intervals (Koriat, Bjork, Sheffer, & Bar, 2004), it is not surprising that, with a longer delay, participants would attend to such information. Other reports that criterion is influenced by the inherent memorability of items (e.g., Brown et al., 1977; Dobbins & Kroll, 2005) is amenable to a similar explanation. For example, Dobbins and Kroll (2005) demonstrated shifts in response criterion as a function of whether pictures were high or low in preexperimental familiarity. Such differences in familiarity likely led participants to attend to that dimension when making recognition decisions and to use a more stringent criterion when judging familiar versus unfa-

miliar pictures. Thus, in the absence of feedback, shifts in response criterion may occur only when a dimension or feature of the target set is necessary or viewed as necessary for a recognition decision.

Using the Test Context

The fact that participants were less likely to exhibit item-by-item shifts in response criterion when feedback was removed in Experiment 3 suggests that test context may be a separable bit of information that informs recognition. That is, test context may be one source of information that contributes to recognition judgments, in addition to the strength of evidence accumulated from prior presentation. Models of memory such as the search of associative memory model (e.g., Gillund & Shiffrin, 1984) do allow for context to contribute to memory judgments, but context is typically construed as those elements of encoding that are present or recapitulated during retrieval (see also Murnane & Phelps, 1993). The current study in fact suggests that test context is sometimes a nonobligatory context that is or is not integrated with retrieved information to make a memory decision. In particular, participants who were aware of the manipulation appeared to make the most use of context information when feedback was present. When feedback was absent, the influence of test context was less pronounced. Awareness may reflect a conscious use of that information (i.e., information about base rates as a function of test context) to inform memory judgments and thus integrate that information (cf. Waltz et al., 1999) during recognition. Lacking awareness, tacit knowledge of base rates may seep into recognition decisions but exert less influence. This account is, of course, somewhat speculative given that awareness has been estimated in a less-than-ideal fashion in the current study. However, it does provide a possible intersection of implicit learning (e.g., Nissen & Bullemer, 1987; Reber, 1967) and explicit memory that warrants further attention.

In addition, the notion that test context may be a source of information that informs recognition decisions raises the possibility that test contexts shifted target distributions (cf. Wixted & Stretch, 2000). In particular, one could argue that test context simply added to the strength of target and lure distributions, with items from mostly old test contexts (as the condition with the higher base rate of studied items) being subjected to greater accrual of strength than were items from mostly new contexts. If this were the case, the results reported could be accommodated by an account which assumes a fixed criterion. That is, the higher level of hits and FAs evident for mostly old items could occur if targets and lures in that condition possessed greater strength than did mostly new items and if the same criterion for endorsement was used for both test contexts. The experiments reported were explicitly designed to minimize this possibility. For example, all items were studied in an equivalent manner, and targets and distractors on the recognition test were equated on several dimensions known to affect memorability (e.g., frequency). Thus, the items used do not lend themselves to a fixed-criterion account.

Instead, the choice of a fixed- or dual-criterion account may depend on how one interprets the use of test context in recognition memory judgments. Old–new recognition decisions are essentially categorization judgments, though the link between categorization and recognition memory has remained largely unexplored (but see, e.g., Estes & Maddox, 1995; Nosofsky, 1988). Test context may either function as the basis for a categorization rule (i.e., as a basis for response criterion) or as a source of category information for the item to be classified as old or new (i.e., as a source of familiarity in the underlying distributions). One intriguing possibility is that context may serve both functions depending on whether one has explicitly learned the association between context and base rate. For example, participants who were aware of the manipulation often suggested that, when in doubt, they answered in a manner consistent with the test context (e.g., "If I was unsure and the word was on the left, I was more likely to respond old."). For these participants, the test context may serve as the basis for a decision rule that produces two different criteria depending on the particular context of a test item. Unaware participants, in contrast, may experience the test context as an extra, though unidentified, source of familiarity that leads to differences in item strength. We cannot currently resolve these two competing ideas but point to them as critical for understanding recognition memory criterion.

Summary and Conclusions

The current study investigated decision processes in recognition memory vis-à-vis changes in response criterion as a function of the probability that items appearing in a particular test context were old. Results from all three experiments indicate that participants can shift their criterion in response to the probability that an item is old and can do so by shifting their criterion on an item-by-item basis. The most dramatic shifts in estimated criterion were apparent for participants who were explicitly aware of the manipulation used, often those who used different keys to input responses. In addition, feedback appears to play a crucial role in criterion shifts, as the absence of feedback was associated with a significant decline in the magnitude of observed criterion differences.

Although decision processes have been relatively understudied by memory researchers, we suggest that the issues raised have import in a number of domains. For example, memory for out-group versus in-group individuals may reflect differences in criterion, with memory for in-group individuals potentially being characterized by a more conservative criterion (e.g., Anastasi & Rhodes, 2005; Meissner & Brigham, 2001). Memory performance in individuals with dementia may also be characterized by a liberalization of responding (Snodgrass & Corwin, 1988). In addition, decision criteria may influence control over memory accuracy (Koriat & Goldsmith, 1996) or in the type, rather than the strength, of information that is sought by the rememberer (Jacoby, Kelley, & McElree, 1999; Jacoby, Shimizu, Daniels, & Rhodes, 2005). Such data suggest that continued examination of decision processes is well worth the focus of memory researchers.

References

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review, 12,* 1043–1047.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51,* 159–172.

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 307–327.

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

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1186–1198.

Dolan, P. O. (1999). Response biases, implicit learning, and the effects of age. (Doctoral dissertation, New York University, 1999). *Dissertation Abstracts International, 60,* 381.

Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1075–1095.

Fitts, P. M., & Seeger, C. M. (1953). S–R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46,* 199–210.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

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

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372–1388.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff locations in recognition memory. *Memory & Cognition, 6,* 544–553.

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *Quarterly Journal of Experimental Psychology, 50A,* 199–215.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace model. *Psychological Review, 93,* 411–428.

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

Hirshman, E., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science, 9,* 61–65.

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. *Annual Review of Psychology, 44,* 265–315.

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 383–400). New York: Guilford Press.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12,* 852–857.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643–656.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review, 97,* 253–270.

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

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

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and nonparametric indexes. *Psychological Bulletin, 107,* 401–413.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7,* 3–35.

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

Murdock, B. B. (1993). TODAM2: A model for storage and retrieval of item, associative, and serial order information. *Psychological Review, 100,* 183–203.

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

Nissen, M. J., & Bullemer, P. (1987). Attentional requirement of leaning: Evidence from performance measures. *Cognitive Psychology, 19,* 1–32.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 5,* 855–863.

Rhodes, M. G., & Kelley, C. M. (2003). The ring of familiarity: False familiarity due to rhyming primes in item and associative recognition. *Journal of Memory and Language, 48,* 581–595.

Schunn, C. D., Lovett, M. C., & Reder, L. M. (2001). Awareness and working memory in strategy adaptivity. *Memory & Cognition, 29,* 254–266.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin, 115,* 163–196.

Singer, M., Gagnon, N., & Richards, E. (2002). Strategies of text retrieval: A criterion shift account. *Canadian Journal of Experimental Psychology, 56,* 41–57.

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

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

Strack, F., & Förster, J. (1995). Reporting recollective experiences: Direct access to the 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, and Cognition, 24,* 1379–1396.

Verde, M. F., & Rotello, C. M. (in press). Memory strength and the decision process in recognition memory. *Memory & Cognition.*

Verfaellie, M., Giovanello, K. S., & Keane, M. M. (2001). Recognition memory in amnesia: Effects of relaxing response criteria. *Cognitive, Affective, & Behavioral Neuroscience, 1,* 3–9.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science, 10,* 119–125.

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

(*Appendix follows*)

Appendix

Posttest Questionnaire Administered to Participants

1. Do you have any thoughts on the experiment?
2. What thoughts did you have while performing the test as to the purpose of having the test items on the left or right side of the screen?
3. Did you do better on one side or the other or were there more words on one side than the other?
4. Was that distracting or helpful in any way?
5. Did the side the test word was on have any influence on your responding?
6. Did you notice any relationship between the side the word was on and its correct answer?
7. Did you consider that while you were performing the task?
8. If a test word came up on the right/left side and you had no idea what the answer was, were you more likely to respond one way or the other, simply because of what side it was on?
9. There was a relationship such that most of the words you studied were presented on the right/left side - Did you notice that?
10. Does that seem right, thinking back?
11. Obviously one strategy you could adopt is to respond 'Old' when it was on the right/left side and 'New' when it was on the left/right side if you were unsure. Did you do that at all?
12. So the side of the screen had no influence on how you responded?

Note. For Experiment 3, questions pertained to the top–bottom portion of the screen rather than the right–left side of the screen.

Received April 21, 2006
Revision received September 21, 2006
Accepted September 27, 2006 ■