Memory & Cognition
2006, 34 (1), 125-137

Effect of delay on recognition decisions: Evidence for a criterion shift

MURRAY SINGER
University of Manitoba, Winnipeg, Manitoba, Canada

and

JOHN T. WIXTED
University of California, San Diego, La Jolla, California

Recent evidence indicates that in intermixed recognition testing of different stimulus classes, people can apply different decision criteria (a *criterion shift*) to stimulus classes distinguished by the study–test delay (Singer, Gagnon, & Richards, 2002), but not by a conspicuous strength manipulation (Stretch & Wixted, 1998b). In an attempt to reconcile these differences, we applied Singer et al.'s text retrieval method to word recognition. People first studied blocked items from each of five categories. After a delay, five new category lists were presented. After each one, the participants recognized intermixed targets and distractors from the current category and one of the earlier ones. At delays of up to 40 min, the answering criteria for immediate and delayed categories were indistinguishable. At delays of 2 days, in contrast, however, both *yes–no* and confidence-rating data indicated that more lenient criteria were applied to delayed than to immediate test items. This suggests that people can use the delay between study and test to flexibly adjust the decision criteria of word recognition.

Signal detection theory is a central formulation in the scientific analysis of recognition decisions. In this framework, recognition test probes, such as studied items (*targets*) and unstudied items (*distractors*), are assessed on a scale of strength (Green & Swets, 1966; Macmillan & Creelman, 1991; McNicol, 1972). A probe is accepted if it exceeds a degree of strength that is treated as a decision *criterion*.

The factors that regulate the positioning of recognition decision criteria have recently come under increased scrutiny. In this regard, item factors, such as word frequency, and procedural factors, such as instructions to participants, result in the application of different criteria to stimulus classes that are otherwise identical. Such variation of the decision criterion is labeled a *criterion shift* (e.g., Brown, Lewis, & Monk, 1977; Hirshman, 1995). The present study evaluated people's ability to shift the criteria applied to item classes within a single list.

More specifically, the study addressed an apparent empirical anomaly in this realm. Stretch and Wixted (1998b, Experiments 3–5) observed similar criteria (i.e., no criterion shift) for words that had appeared either frequently (strong) or only once (weak) during learning. This occurred despite the facts that the strong and weak targets appeared in their own respective colors throughout learning and test and that the strength–color association was sometimes explained to the participants (Experiment 5). The main evidence that the criterion did not shift as a function of strength was that the false alarm rates were the same for distractors that appeared in different colors. By contrast, in a study of sentence recognition, Singer, Gagnon, and Richards (2002) detected a criterion shift. Their participants made recognition decisions about sentences with reference to prior stories. Test sentences that were encountered immediately after their stories and others whose testing was delayed were randomly intermixed. Distinct criteria were measured for the immediate and the delayed items, as evidenced by the fact that the false alarm rate was higher for distractors that were associated with stories that were tested after long delays than for those tested after short delays. Each distractor sentence clearly referred to one of the stories but expressed a false idea.

Thus, test delay, one salient variable, provided the basis for a criterion shift (Singer et al., 2002), whereas item strength complemented by a color cue, another salient variable, did not (Stretch & Wixted, 1998b). This was despite the fact that both manipulations had a powerful effect on the strength measure d'. The main goal of this study was to reconcile those findings. To eliminate stimulus class (sentence vs. word) as the basis of the different results, we applied the method of Singer et al. (2002) to

The data of Experiments 1–3 were presented at the 43rd Annual Meeting of the Psychonomic Society, Kansas City, November 2002. This research was supported by Discovery Grant OGP9800 from the Natural Sciences and Engineering Research Council of Canada and a Leave Research Grant from the University of Manitoba to the first author. We thank Kim Mintenko and Launa Leboe for conducting the experimental sessions and Tamara Ansons for technical assistance in the preparation of the manuscript. Please address correspondence to M. Singer, Department of Psychology, University of Manitoba, Winnipeg, MB, R3T 2N2 Canada (e-mail: m_singer@umanitoba.ca).

word recognition. In the next section, the theoretical details of this research domain will be described. Then an overview of the present approach will be provided.

Criterion Shifts in Recognition Memory

Signal detection theory. A central principle of signal detection theory is that test items, both targets and distractors, are evaluated on a strength scale, such as familiarity, as is depicted in Figure 1 (Banks, 1970; Green & Swets, 1966; Macmillan & Creelman, 1991; McNicol, 1972; Parks, 1966). The familiarity values of all item classes are assumed to be normally distributed. Without any loss of generality, the distractor distribution is treated as the standard normal (mean = 0, variance = 1). The d' measure is the standardized distance between the target and the distractor distributions.

Classic detection theory posits that the variance of the target and distractor distributions are equal. However, there is considerable evidence that human recognition performance is better characterized by an unequal-variance model. In particular, analyses of the receiver operating characteristic (ROC) typically show that the standard deviation of the target distribution is about 1.25 that of the distractor distribution (Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff, Sheu, & Gronlund, 1992). Despite this, an equal-variance depiction is convenient for presenting the predictions of signal detection analysis (e.g., Hicks & Marsh, 1998, p. 1108; Hirshman, 1995, p. 307; Wixted & Stretch, 2004). Also, for convenience, we use *familiarity* to label the decision axis, although the memory strength variable might be a combination of recollection and familiarity (Wixted & Stretch, 2004).

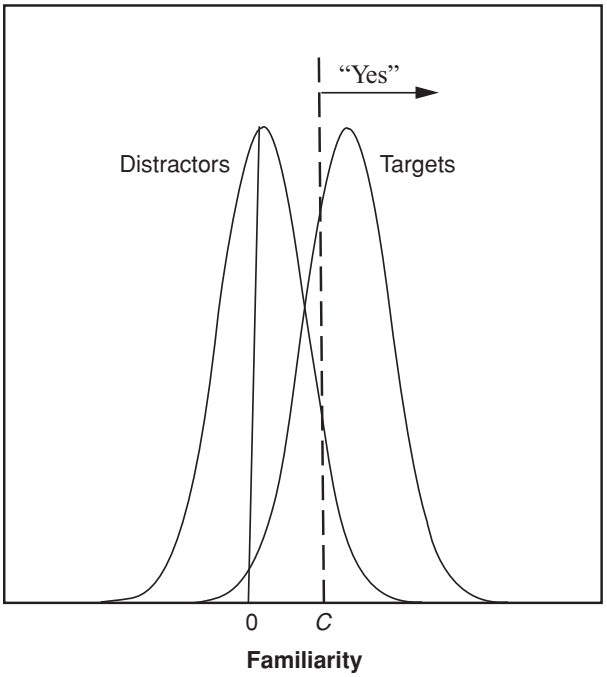

Figure 1. Location of target and distractor distributions and of the decision criterion (*C*) on a signal detection familiarity scale.

The recognition decision about each probe depends on whether its familiarity exceeds a decision *criterion*, shown at value *C* in Figure 1. On the familiarity scale, a numerically *higher* criterion produces *fewer* false alarms ("yes" replies to distractors; see Figure 1), and conversely for a numerically *lower* criterion. Traditionally, criterion placement was considered to be influenced by such factors as task instructions and the costs and benefits of correct and incorrect decisions (Macmillan & Creelman, 1991; McNicol, 1972). As such, criterion placement is considered to be under conscious control (Roediger & McDermott, 1999; cf. Reder & Schunn, 1996).

Criterion shifts: Between-list and within-list effects. The formulation and application of signal detection analysis, originally a perceptual theory, deemphasized the positioning of the decision criterion (Green & Swets, 1966; see Hirshman, 1995). However, recent demonstrations of adjustments and shifts in the decision criterion have begun to clarify the principles governing criterion placement.

Two examples stem from investigations of the impact of item strength on recognition memory. In one study (Hirshman, 1995, Experiments 1 and 2), participants studied words for 0.4 sec (weak) or 2.0 sec (strong) each. The study lists included either weak items only (pure) or both weak and strong items (mixed). Hirshman detected a criterion shift: The criteria were lower for the weak items in the pure lists than for those in the mixed lists, even though d' scores were about equal (see also Ratcliff et al., 1992). Likewise, in Experiment 1 of Stretch and Wixted (1998b), the researchers constructed a strong list, which presented some words three times each, and a weak list, in which words appeared only once each. They detected more false alarms (hence, a lower criterion) for the less memorable words (namely, the weak ones).

The latter two item strength criterion shifts received similar explanations (Hirshman, 1995; Stretch & Wixted, 1998b): It was proposed that experimental participants are sensitive to differences in memorability between stimulus categories and that they accordingly set liberal criteria for weak conditions. Hirshman, furthermore, posited that criterion placement is based on the average familiarity of the list items. This average is lower for a pure list of weak items than for a mixed list of weak plus strong items. This accounts for criterion differences between Hirshman's pure-weak lists versus his mixed weak-plus-strong lists.

These and other (e.g., Hicks & Marsh, 1998) criterion shifts have resulted from manipulations *between lists*. For example, what Hirshman (1995) measured was different criteria for weak items in *separate* lists either of (1) weak items only or (2) weak plus strong items. Within the mixed list, however, the criterion was the same for the weak and the strong items.

There is also some evidence of criterion shifts *within a single list*. The involvement of the decision criterion in the recognition of story sentences was inspected by Singer

et al. (2002). Merging the procedures of Reder (1988) and Kintsch, Welsch, Schmalhofer, and Zimny (1990), Singer et al. instructed people to read five stories, in anticipation of a memory test. After a delay, the participants encountered additional stories, each of which was followed by a recognition test that included test items about the current story and one yoked story from prior to the delay. Thus, at test, the participants encountered randomly intermixed immediate and delayed probes (see also Reder, 1988). Each distractor, although clearly false, was related to one of the stories.

Strikingly, the false alarm rate was lower for the immediate condition than for the delayed condition. One explanation for this effect is that the decision criterion was consistently placed at a higher point on the strength scale in the immediate condition than in the delayed condition. This effect was measured at delays of 20 min, 40 min, and 2 days. If this effect does reflect a criterion shift, the participants must have adjusted the criterion on an item-by-item basis, because the short-delay and long-delay items were randomly intermixed on the recognition test. Within-list criterion shifts have been implicated in other recognition phenomena, such as the revelation effect (Hockley & Niewiadomski, 2001).

However, the elusiveness of applying distinct criteria to different stimulus classes within a list was documented by Stretch and Wixted (1998b, Experiments 3–5; see also Morrell, Gaitan, & Wixted, 2002). Within lists, some of the words that were to be learned appeared once, and others appeared five times. To highlight the manipulation, the strong words were colored red (during both study and test), and the weak words were colored blue. Despite the presumed strength difference between the red and the blue words and the salience of the manipulation, similar criteria were in effect for the two sets. Stretch and Wixted proposed that cognitive demands might discourage customizing the criterion to each new probe, since it would require nearly constant shifting of the criterion with each new test item (see also Gillund & Shiffrin, 1984; Hockley & Niewiadomski, 2001).

Overview of the Present Study

We considered the apparent impact of some variables (e.g., delay; Singer et al., 2002), but not of others (e.g., strength; Stretch & Wixted, 1998b), on people's tendency to continually adjust a decision criterion to merit scrutiny. To reconcile these findings and to eliminate stimulus class as the basis of their differences, we applied the method of Singer et al. to word recognition. Taxonomic categories served in place of stories. Category judgments have previously served as a fertile realm for evaluating retrieval strategies in word and sentence recognition (e.g., Lorch, 1981; Singer, 1991), and signal detection analyses have been applied in this domain (Wixted & Stretch, 2000).

Thus, in the present experiments, words derived from 10 categories replaced the story stimuli in Singer et al. (2002). During one phase of the experiments, people studied blocked sets of words from 5 categories (e.g., *occupations*). Later, they studied words from 5 different categories (e.g., *cities*). Each of the latter categories was followed by a recognition test that randomly intermixed words from an earlier category and the current category (e.g., *occupations* and *cities*). Of central concern was whether or not a criterion shift would emerge: That is, would distinct decision criteria be applied to the immediate and the delayed categories? The design of this study is similar to one recently reported by Morrell et al. (2002). Those investigators varied strength between different intermixed semantic categories by means of differential repetition, rather than differential delays. Their results were consistent with the absence of a criterion shift, and the question we ask here is whether or not the same outcome is observed when strength is manipulated by delay.

EXPERIMENT 1

Method

Participants. The participants were 58 female and male students of introductory psychology at the University of Manitoba who were native speakers of English. They took part in partial fulfillment of a course requirement.

Materials. The stimuli were words derived from the category norms of Battig and Montague (1969). We sought 10 categories with over 40 terms that appeared both familiar and representative of the category. The category *part of a building*, to cite one example, was excluded: By its 40th entry, the unrepresentative member *chairs* was listed. The chosen categories were *birds*, *body parts*, *chemical elements*, *cities*, *countries*, *diseases*, *male names*, *mammals/four-footed animals*, *occupations*, and *American states*. For the categories of *cities*, *countries*, and *American states*, the eligible list was based on current population (determined from Internet sources), rather than on the Battig and Montague orderings. We avoided multiword category members (e.g., *blue jay* for *bird*) and items that named their category (e.g., blue*bird*). If a word appeared in two categories, it was retained for one category chosen at random (e.g., *turkey* was a bird and a country). Likewise, one member of synonym pairs was randomly excluded (e.g., *puma–cougar*). All place names and person names appeared capitalized, as usual. The Appendix shows these categories, as well as the eligible stimuli for all of the present experiments.

These stimuli were organized into three lists, each of which had two versions. Each participant encountered only one of these six alternatives. As was mentioned earlier, the first portion of each list version comprised 15 blocked exemplars from each of five categories (disregarding practice). The second portion presented the blocked members of five more categories, but each of the latter categories was followed by an intermixed recognition test about it and one category from the first portion of the list. For all lists, the categories of *colors* and *female names* constituted practice materials.

More specifically, consider the construction of List 1, the first of the three lists. List 1 was derived from the 10 chosen categories as follows. First, the first 30 members of each category that met the criteria described earlier were selected. Then, the two words of each successive pair according to Battig and Montague's (1969) ordering (or else population size ordering) were randomly designated to function as a target and a distractor, respectively. Next, the 10 categories were randomly assigned to five pairs. *Cities*, *countries*, and *states*, however, could not be in the same pair; nor could *mammals* and *birds*.

The five category *pairs* (e.g., *occupations* and *cities*) were then randomly ordered from 1 to 5. One member of each pair (e.g., *occupations*) was randomly assigned to the first portion of the list,

and the other to the second portion. In this way, the second-portion category appeared five beyond its mate among the 10 categories: Category 6 five positions after Category 1, Category 7 five after Category 2, and so forth. If, for example, *occupation* was the first of the 10 categories in a list, the first nonpractice stimuli that the participants encountered were the 15 randomly ordered occupation stimuli.

The *test* items of List 1 were the 30 members of each category, comprising 15 targets and 15 distractors. In a second version of List 1, the roles of target and distractor were reversed for each item pair. Two more lists, with two versions each, were constructed in the same manner as List 1. To repeat, each participant encountered only a single version of one of the lists.

To introduce a delay between study and test, two arithmetic puzzles were presented. The participants were instructed to combine the digits 2, 3, 5, and 7, using ordinary arithmetic operations to generate each of the values from 1 to 25. For example, $(7 + 5 - 3^2)$ yields the answer 3. The second puzzle was identical but used the digits 1, 3, 4, and 9.

Procedure. The sessions were conducted in groups of 1 to 4. Each participant sat in a separate, closed room at a station consisting of a personal computer, monitor, and keyboard.

The session began with 5 min of arithmetic puzzle solving. Thus, the first five experimental categories, like the last five, were preceded by puzzle solving. The participants then studied the first five categories. Each category began with the message "NEXT LIST" for 1 sec. After an additional 1.5 sec, the 15 exemplars were presented in random order for 500 msec each plus a 250-msec interstimulus interval. The participant studied the words in anticipation of a recognition test to follow later. After a 3-sec intercategory interval, the next category was presented.

After the fifth category, a further 11.5-min period of puzzle solution ensued, which, coupled with task instructions, generated the desired 20-min delay between study and test. Then screen instructions reminded the participants that they would encounter additional categories and that each one would be followed by test items about it and one of the earlier categories. Each new category was presented in the same manner as the original five. After each one, there was a 3-sec interval. Then a message such as "Recognition test items about occupations and cities" signaled, for 5 sec, that a test would occur. Next, the 15 target and 15 distractor items for each of the two categories were presented in a randomly intermixed order. Each test item was preceded by a fixation "x" for 0.5 sec and was followed by a 0.5-sec intertrial interval. The participants registered their responses with their index fingers, using the "." and "x" keys for *yes* and *no*, respectively. There was no answer time limit, and no feedback was provided. After the last test item of a pair of categories, there was a 3-sec interval, and then the next block was presented. A message signaled the end of the experiment.

After the initial puzzle solving but before any category study, the participants viewed two practice categories (*colors*, *female names*). The method for these categories was identical to the others, except that each was followed by test items from that category alone.

Results

Table 1 presents the d' scores, hits, and false alarms in Experiment 1. It also shows the signal detection criterion, in the form of a z score (z_c) relative to the position of the mean of the distractor distribution (viz., 0). The z_c is the position on the strength continuum that cuts off the right-hand portion of the distractor distribution in a proportion equal to the false alarm rate. According to the standard-normal table, for example, a false alarm rate of 3% corresponds to a z score of 1.88. This manner of expressing the criterion is central to our evaluation of the criterion shift proposal (e.g., Morrell et al., 2002; Treisman & Williams, 1984). Lower values of z_c on the strength axis tend to diagnose more lenient criteria. Alternative signal detection criteria, in contrast, were not suitable for our purposes. Statistic *C*, for example (Macmillan & Creelman, 1991), evaluates the respondent's bias to favor the *yes* or no response. Suppose that, as we hypothesize, the criterion shifted *locations* from the immediate to the delay condition. Both of these two criteria might, however, be positioned at the intersection of their respective target and distractor distributions. By its mathematical definition, *C* would then have the identical value of 0 (diagnosing the absence of bias) in both instances, masking the fact that the criterion had shifted locations.

An ANOVA was applied to the data: For each measure, delay was a within-participants variable, and list and list version were between-participants variables. The list and version variables were included in the design for the purpose of counterbalancing and held no theoretical interest. Furthermore, they entered into few significant effects. They will be reported but are not shown in Table 1. Finally, the ANOVA results for raw scores (hits and false alarms) will be reported only if they are inconsistent with those for the signal detection measures. An alpha level of .05 was used throughout.

Table 1 shows that the criterion z_c had similar values of 0.47 and 0.52 in the immediate and the delay conditions, respectively. The z_c ANOVA revealed no significant effects at all. The strength measure d' was greater in the immediate than in the delay condition [$F(1,52) = 84.28$,

Table 1
Experimental Measures as a Function of Delay (With Standard Errors)

		Measure							
		Hits		False Alarms		d'		z_c	
Experiment	Delay	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
1	Immediate	.763	.011	.336	.012	1.23	.076	0.47	.06
	Delay	.558	.012	.318	.011	0.67	.055	0.52	.06
2	Immediate	.855	.016	.208	.301	2.04	.129	0.89	.105
	Delay	.759	.031	.228	.029	1.60	.124	0.83	.107
3	Immediate	.831	.021	.159	.017	2.20	.124	1.11	.068
	Delay	.623	.020	.313	.017	0.84	.056	0.53	.058
4	Immediate	.901	1.124	.163	1.902	2.58	.116	1.11	.076
	Delay	.680	2.090	.379	1.881	0.82	.054	0.32	.055

$MS_e = 0.10$]. The d' ANOVA also revealed a significant delay × version interaction [$F(1,52) = 5.71$, $MS_e = 0.10$]. The only deviation from these outcomes in the raw score ANOVAs was that the delay × version interaction was not significant in the hit analysis.

Discussion

The main result of Experiment 1 was that the decision criterion did not vary with delay: The criterion location z_c was slightly higher in the delay condition than in the immediate condition, a difference opposite in direction to the criterion shift hypothesis that the participants would assign a more liberal criterion to the delayed items. Thus, the results correspond to those recently reported by Morrell et al. (2002). Those authors suggested that participants did not use different criteria for the different strength conditions. The outcome of Experiment 1 could mean that (1) the criterion shift hypothesis accurately characterizes recognition decisions about text (Singer et al., 2002), but not about words, or (2) the particular parameters in Experiment 1 in some manner prohibited the application of distinct criteria to immediate and delayed recognition probes. Experiment 2 particularly addressed the second alternative: Several procedural changes were implemented to facilitate item-by-item criterion changes on the participant's part. First, at their 20-min delay, Singer et al. detected only a small, although significant, difference between the immediate and the delayed criteria. Therefore, the delay between the early and the late categories was increased to 40 min. Second, the readers in Singer et al.'s study can be assumed to have derived the meaning of the texts they examined. To likewise encourage the semantic processing of the present stimuli, the participants in Experiment 2 were instructed to perform a pleasantness rating of each studied stimulus (e.g., Shiffrin, Huber, & Marinelli, 1995). This entailed an increase in stimulus presentation time. Third, each category was signaled with its name during study. In addition, the number of items per category was increased to 20 targets and 20 distractors, from the 15 of each used in Experiment 1. This change was not considered to promote a criterion shift but, rather, was intended to prevent ceiling effects that might have resulted from the semantic orienting task.

EXPERIMENT 2

Method

Participants. There were 33 naive participants selected from the same participant pool as that used in Experiment 1.

Materials. The stimuli comprised all 40 words from each of 10 experimental categories and 30 words from each of the two practice categories (see the Appendix) that served as the pool of stimuli for Experiment 1.

The list and list version variables entered into only one significant effect in Experiment 1, and we attached no theoretical importance to that outcome. Therefore, only one list was constructed for Experiment 2, following the same principles as those for List 1 in Experiment 1. Two versions of the list were used. In each version, there were 20 targets and 20 distractors in each category, and the target–distractor roles were reversed between the versions. *Colors* and *female names* again functioned as the practice categories.

Procedure. In Experiment 2, the study phase for each category list was signaled by a message such as "OCCUPATION WORDS," rather than by the "NEXT LIST" warning in Experiment 1. During study, the participants used the top-left keyboard buttons of 1 to 4 to rate the pleasantness of each word (4 = *highly pleasant*). Word presentation time was 3 sec, and if no response was registered during this time, a tone sounded for 250 msec. An interstimulus interval of 500 msec preceded the next word.

The interval between study and test was 40 min, split between the solving of the arithmetic puzzles in Experiment 1 and acrostic word puzzles. The word puzzles were introduced because the Experiment 1 participants found even 20 min of arithmetic puzzle solving to be tedious.

When the participants encountered the final five categories plus tests, they had to move their fingers from the rating keys 1 to 4 (study) to the *yes* and *no* keys (test). Therefore, messages about impending study or test were accompanied by instructions to relocate one's fingers appropriately. The message remained on the screen until the participant pressed the space bar to signal readiness. For the sake of consistency, the two practice categories and the first five categories were also preceded by messages to the participants to place their fingers on the correct keys. In all other respects, the procedure was identical to that in Experiment 1.

Results

The mean pleasantness rating time during study was 1,093 msec ($SD = 455$ msec). The recognition results of Experiment 2 appear in Table 1. In all ANOVAs, delay was a within-participants variable, and list version was a between-participants variable. The immediate and delay z_c criterion scores did not differ significantly, and an ANOVA of these scores yielded no significant effects. However, the ANOVA revealed that d' was significantly higher in the immediate than in the delay condition [$F(1,31) = 32.87$, $MS_e = 0.10$]. The d' ANOVA also revealed a delay × version interaction [$F(1,31) = 6.22$, $MS_e = 0.10$]. Tests of simple main effects to pursue the latter interaction revealed that there was a significant effect of delay for both versions of the list (Fs ≥ 5.41, $MS_e = 0.10$). Finally, comparable to Experiment 1, the raw score ANOVAs were completely consistent, except that the delay × version interaction was not significant in the hits ANOVA.

Discussion

Experiment 2 was designed to overcome features of Experiment 1 that may have obstructed trial-by-trial criterion adjustment by the participants. In contrast with Experiment 1, each study category was explicitly labeled; the participants performed semantic judgments about the stimuli, rather than examining them passively (and concomitantly, maximum study time per word, including the interstimulus interval, was 3.50 sec, rather than 0.75 sec), and study and test were separated by an interval of 40 min, rather than 20 min. The Experiment 2 z_c difference was in the direction predicted by the criterion shift hypothesis, but it was not significant. This outcome invited the conclusion that people do not continually adjust decision criteria among word stimulus categories within a single list (e.g., Stretch & Wixted, 1998b; cf. Hockley & Niewiadomski, 2001). However, in previous studies, it has been at a study–test interval of 2 days that dramatic

differences have been detected between the false alarm rates associated with the immediate and the delayed conditions in an intermixed list (Kintsch et al., 1990; Reder, 1988; Shiffrin et al., 1995; Singer et al., 2002). Experiment 3 constituted a third attempt to provide evidence for within-list adjustment of decision criteria. It combined most features of the method of Experiment 2 with a 2-day study–test interval.

EXPERIMENT 3

Method

Fifty-six naive participants were selected from the same participant pool as that used in the previous experiments. The materials were identical to those in Experiment 2. On Day 0, two practice categories and five experimental categories were presented in a manner identical to that in the corresponding phases in Experiment 2. Two days later (Day 2), the participants returned to the laboratory. They first solved arithmetic puzzles for 5 min, in order that the Day 2 categories, like the Day 0 categories, be preceded by puzzle solving. They then encountered the final five categories, each followed by an intermixed recognition test of words from the current category and its yoked counterpart from Day 0.

Results

The mean pleasantness rating time was 1,125 msec (SD = 443 msec). The recognition results appear in Table 1. Z_c was appreciably lower, or more lenient, in the delay condition than in the immediate condition, a pattern dramatically different from those in Experiments 1 and 2. The delay main effect was significant [$F(1,54) = 78.96$, $MS_e = 0.114$]. The z_c ANOVA also revealed a delay × version interaction [$F(1,54) = 7.69$, $MS_e = 0.007$]. Follow-up tests of simple main effects indicated that the effects of delay were significant for both versions (Fs > 7.89, $MS_e = 0.11$). In the immediate condition, d' was greater than in the delay conditions [$F(1,54) = 223.25$, $MS_e = 0.23$].

Discussion

The z_c criterion was much lower in the delay condition than in the immediate condition. This suggests that the participants adjusted their response criterion on an item-by-item basis. This outcome reproduces the pattern that is detected when immediate and delayed test items are intermixed in *text* recognition (Reder, 1988; Singer et al., 2002).

It is noteworthy that decision criteria are likewise considerably more lenient (and false alarm rates higher) for delayed than for immediate probes when recognition testing is entirely separate for probes representing the different delays (*uniform-delay* testing, unlike in Experiments 1–3). This pattern has been measured in both text recognition (Kintsch et al., 1990) and category word recognition (Shiffrin et al., 1995, Appendix D, Experiment 3). Singer et al. (2002) posited that at any given delay, there ought to be some resemblance between recognition performance for uniform-delay and mixed-delay testing. They further proposed that in mixed-delay lists, the probes of the two delays mutually influence the decision criteria of one another. In accord with these proposals, Table 2 shows that they measured (1) more false alarms for delayed than for immediate stories in both uniform and mixed testing and (2) less extreme criteria under mixed testing than under uniform testing (Singer et al., 2002, Experiment 2). The latter outcome was corroborated by a significant test procedure × delay interaction. Singer et al. judged this profile to be consistent with the criterion-averaging analysis of Hirshman (1995), discussed earlier.

It might be proposed that the results of Experiment 3 reflect a shift in the distribution of the delayed distractors combined with just a *single criterion* (Hicks & Marsh, 1998; Roediger & McDermott, 1999; Shiffrin et al., 1995; Stretch & Wixted, 1998b; Wixted & Stretch, 2000). Even with just one criterion, false alarms would be greater in the delay than in the immediate condition if the distribution of the delay distractors exceeded that of the immediate distractors in either mean familiarity or variance. In both instances, a greater proportion of the distribution would exceed the criterion in the delay than in the immediate condition. However, both of these alternatives seem implausible, because they would require the assumption that some distractors increased in strength over the course of the retention interval of the delay condition. To the contrary, it might reasonably be posited that (1) distractors from delayed categories have the same familiarity as distractors from immediate categories or (2) distractors from delayed categories have slightly lower familiarity than do distractors from immediate categories, if unmentioned exemplars accrue modest activation during study that subsequently declines (e.g., Gillund & Shiffrin, 1984; Shiffrin et al., 1995).

Despite the apparent implausibility of the latter alternatives, we undertook to more definitively discount them. In Experiment 4, we replaced the *yes–no* recognition classifications of Experiments 1–3 with confidence ratings. Confidence ratings permit the extraction of ROC curves, functions that permit the evaluation of competing signal detection analyses. We also took the opportunity, in Experiment 4, to completely counterbalance category exemplars across experimental conditions: Each exemplar was now cycled across the four conditions representing the crossing of target–distractor and delay.

EXPERIMENT 4

Method

Participants. The participants were 60 naive individuals from the same population as that accessed for the other experiments.

Table 2
False Alarm Rate as a Function of Delay and Testing Procedure in the Data of Singer, Gagnon, and Richards (2002, Experiment 2)

	Procedure	
Delay	Uniform	Mixed
Immediate	.020	.053
Delay	.116	.101

Materials. The practice and experimental categories and their category members were identical to those in Experiment 3. The experimental stimuli took the form of four counterbalanced lists. The 10 experimental categories were randomly assigned to five new category pairs, subject to the same restrictions as those in Experiment 1 (e.g., *mammals* and *birds* not in the same pair). In List 1, one category of each pair was randomly assigned to the immediate condition, and the other to the delay condition. Furthermore, within each category, 20 exemplars were randomly selected to appear as targets, and the other 20 to appear as distractors. In the remaining three lists, exemplars were cycled across the four target–distractor × delay conditions, using a Latin-square procedure. The two practice categories and their exemplars were identical to those in Experiment 3.

Procedure. The procedure was identical to that in Experiment 3, except that, in the recognition tests, the participants were instructed to label negative judgments as *definitely no*, *no*, and *maybe no*, and analogously for positive judgments. Six keys—"x", "c", "v", ",", ".", and "/"—were labeled, from left to right, with corresponding numbers of minus and plus signs ranging from − − − to + + +. The participants registered their recognition confidence ratings using the index, middle, and ring fingers of their two hands.

Results and Discussion

Two participants never used any of the three positive responses for delayed targets and distractors, and 1 participant exhibited no ability to discriminate targets and distractors in either delay condition. The data from these 3 participants were excluded from all the analyses. Each of the four lists was viewed by no fewer than 13 of the remaining 57 participants.

The mean pleasantness rating time was 1,118 msec (SD = 491 msec). For the sake of comparability with the other experiments, we first considered the participants' simple *yes–no* judgments, disregarding the degree of confidence. The corresponding data appear in Table 1. As in Experiment 3, z_c was much lower in the delay condition than in the immediate condition [$F(1,53) = 136.53$, MS_e = 0.124]. The list main effect was significant [$F(3,53)$ = 5.49, MS_e = 0.274]. The ANOVA also revealed a delay × list interaction for the z_c scores [$F(1,53) = 3.03$, MS_e = 0.124]. However, tests of simple main effects indicated that the main effect of delay was significant for all four lists (Fs > 12.55, MS_e = 0.124. The d' scores were appreciably higher in the immediate than in the delay condition [$F(1,53) = 381.24$, MS_e = 0.22]. The d' analysis also revealed a significant effect of list [$F(3,52) = 4.55$, MS_e = 0.007] but no delay × list interaction. Unlike the d' analysis, the hit ANOVA did not reveal a main effect of list.

ROC curves were derived from the confidence ratings in the following manner. It was assumed that the delay manipulation did not affect the distribution of the immediate and delay distractors but that the criteria shifted across conditions. That assumption sanctioned one ROC analysis for the immediate data (targets and distractors) and another one for the delay data (targets and distractors).

Figure 2 presents the resulting ROC curves, obtained by pooling the data over participants. The ROC points were obtained by treating the boundaries between each pair of adjacent response categories as signal detection decision criteria. Consider, for example, the adjacent categories *maybe yes* and *yes*. The hit rate referred to responses to targets that exceeded their boundary (viz., *yes* or *definitely yes*), and the false alarm rate was based on responses to distractors in those two categories. The resulting ROCs were curvilinear in character, an outcome that is consistent with standard signal detection analyses (e.g., Yonelinas, 1994). Transformed functions, obtained by plotting the data on z score coordinates, had slopes of .73 and .77 in the immediate and the delay conditions, respectively. These values are close to the value of .80 that is indicative of an unequal-variance detection model (Ratcliff et al., 1992).

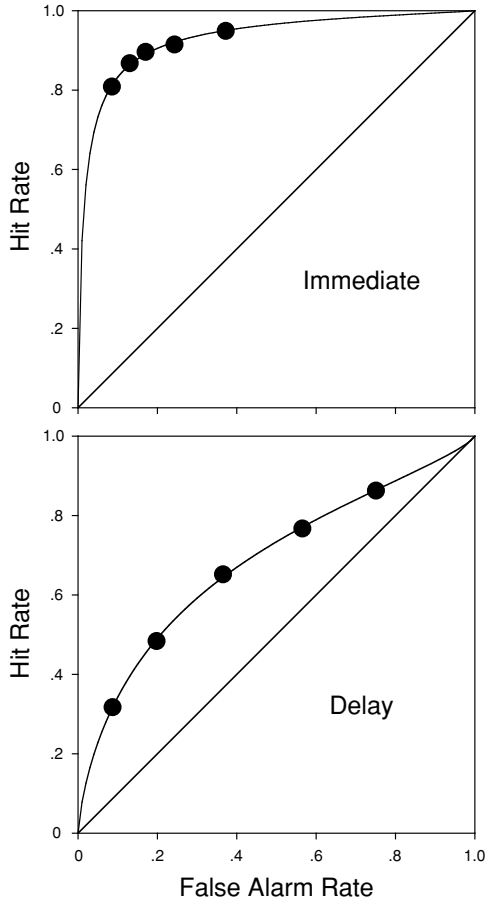

Figure 2. Receiver-operating characteristic (ROC) curves for the immediate and delay conditions in Experiment 4. The slopes were .73 and .77 in the immediate and the delay conditions, respectively; the strength values (d_e) were 2.19 and 0.68 in the immediate and delay conditions, respectively.

Table 3 presents the z_c criterion scores of the immediate and delay conditions. The boundary between the *definitely yes* and *yes* categories is denoted by the yes+++ criterion, and so forth for the other four criteria. Several features of these data are noteworthy. First, the yes+ criterion is simply the usual signal detection criterion distinguishing *yes* and *no* replies. As was mentioned earlier, an ANOVA revealed that this criterion differed significantly between

the immediate and the delay conditions. Second, the immediate and delay criteria were numerically identical for the most strict boundary (yes+++), but their difference increased systematically, proceeding from the more strict to the less strict criteria. That is, for all but the yes+++ criterion, the data consistently revealed criterion shifts.

The fanning pattern of the criteria in Table 3 is highly similar to that detected by Stretch and Wixted (1998a) for a strength manipulation. However, that manipulation was made between lists, whereas the present one resulted from the within-list manipulation of the delay variable.

Stretch and Wixted (1998a) concluded that this pattern of results favored a criterion shift model, because it was unlikely that the distractors were affected by the strength manipulation.[1] The distractors in that case were unrelated words, so it seemed sensible to argue that the vast ocean of unrelated words that could be used as distractors would not have their properties affected by the strength of the words that appeared on their study list. In Experiment 4, however, the words were categorized. As a result, the presentation of numerous words in a taxonomic category might more plausibly have affected unpresented words in the same categories—words that subsequently functioned as distractors.

Consider, therefore, the possibility that there was no criterion shift and that the immediate and delay distractor distributions were shifted relative to one another. This circumstance would actually make it possible to perform an ROC analysis comparing one distractor distribution with the other. We performed such an analysis, treating the distractor distribution associated with the higher false alarm rate (delay) as the target distribution, with its false alarms regarded as hits. This analysis yielded a *z*-ROC slope of 1.89. This outcome would indicate that the variability of the immediate distractors greatly exceeded that of the delay distractors. There is no rationale for the possibility of such a difference, so the much simpler explanation is that the criteria shifted across conditions.

Finally, we note that in the 2-day delay experiments, the z_c ANOVAs revealed that delay interacted with either list version (Experiment 3) or list (Experiment 4). These interactions, however, diagnosed only the *amount* by which the immediate criterion exceeded the delay criterion. In both experiments, the delay effect was significant for all versions or lists.

Table 3
Confidence Criterion Scores (z_c) of Experiment 4 as a Function of Delay

	Delay		
Criterion	Immediate	Delay	Difference
Yes+++	1.39	1.39	0.00
Yes++	1.10	0.85	0.25
Yes+	0.92	0.32	0.60
No−	0.70	−0.17	0.87
No−−	0.33	−0.67	1.00

GENERAL DISCUSSION

In this study, we asked whether people can apply distinct decision criteria to immediate and delayed word probes that are randomly intermixed in a single recognition list. In an analogous study of the recognition of sentences with reference to stories (Singer et al., 2002), more conservative criteria were detected for immediate than for delayed intermixed probes at test delays of 20 min, 40 min, and 2 days. However, that study differed from many recognition investigations in its use of (1) sentence stimuli and (2) three types of targets (explicit, paraphrased, and inference), apart from the distractors. Singer et al.'s results also contrasted markedly with Stretch and Wixted's (1998b) finding that people apply the *same* criterion to highly distinct sets of strong (frequently presented) and weak words that are encountered in a single test list.

Therefore, we adapted Singer et al.'s (2002; see also Reder, 1988) story recognition procedure to the study of familiar taxonomic categories. At delays of under an hour (Experiments 1 and 2), the decision criteria for the immediate and the delayed categories were statistically indistinguishable. This was true even when the procedure highlighted the categorical distinctions and promoted semantic processing (Experiment 2) and even though the delay manipulation produced a large and significant effect on hit rates. With a 2-day delay, however, the decision criterion was placed at a much lower point on the memory strength axis for immediate items than for delayed items. This clearly demonstrated that the impact of delay on criterion adjustment applies to category words, as well as to story sentences. These results were detected against a sensible backdrop of greater d' scores for immediate targets than for delayed targets. We note that the methods in Experiment 2, on the one hand, and Experiments 3 and 4, on the other, were distinguished only by delay length, so the differences between their result patterns cannot be attributed to other variables.

The participants' recognition judgments took the form of confidence ratings in Experiment 4. ROC curves derived from these data yielded typical curvilinear functions that are consistent with the signal detection analysis. Equally important, the confidence data favored a criterion shift conclusion in two ways. First, when the ROC analysis was performed on the assumption that the weak and the strong distractor distributions were the same and that the false alarm rate differences arose due to criterion shifts (as we argue), the parameter estimates were typical in every respect. That is, the standard deviations of the target distributions were estimated to be slightly greater than those of the distractor distributions (the ROC slopes were .73 and .77 for the immediate and the delay conditions, values that are quite close to the expected value of .80), and the confidence criteria exhibited a fanning pattern when they shifted across conditions. That is, as was previously observed by Stretch and Wixted (1998a), the leftmost criterion exhibited a large shift, whereas the

rightmost criterion was essentially stationary across conditions. The middle criterion (i.e., the *yes–no* decision criterion) shifted an intermediate amount. All of these within-list results correspond to what is typically observed when participants shift their criteria *between* lists.

When the ROC analysis was conducted on the basis of the assumption that the criteria did not shift across conditions (contrary to what we argue) and that the false alarm rate differences arose because of differences between the weak and the strong distractor distributions, the results were atypical. In particular, the standard deviation of the strong distractor distribution was estimated to be almost twice that of the weak distractor distribution, even though the distance between the means of the two distributions was estimated to be fairly small (d_e was estimated to be 0.91). This result is theoretically possible, but such large variance differences are almost never observed.

For the present 20- and 40-min delays (Experiments 1 and 2), the false alarm differences between the immediate and the delay conditions were not statistically significant, whereas in Singer et al.'s (2002) study they were. However, both studies revealed that the difference between the delay and the immediate false alarm rates increased monotonically with delay. These trends are shown in Table 4. This pattern has also regularly been measured in text recognition studies using uniform-delay testing (Kintsch et al., 1990; Reder, 1982; Singer, 1979). These results suggest that delay exerts a regular impact on the selection of the recognition decision criterion. However, criterion placement is also regulated by the experimental stimuli and the session parameters. Therefore, the amount of delay that is needed to produce a delay criterion that is statistically more lenient than the immediate criterion will differ from experiment to experiment.

Factors Promoting Criterion Shifts

Magnitude of *d′* differences. The delay manipulation exerted a significant effect on d' in all four experiments, but z_c differed significantly only in Experiments 3 and 4. As might be expected, the latter 2-day-delay experiments yielded the largest effect on d', so one might posit that within-list criterion shifts are promoted by effects of d' that are quite substantial. However, an appreciable d' difference appears neither sufficient nor necessary for a criterion shift. With regard to *sufficiency*, highly significant d' differences resulting from within-list manipulations have, in some instances, not generated criterion shifts (e.g., the present Experiments 1 and 2; Morrell et al., 2002; Stretch & Wixted, 1998a). In this regard, the d' difference of Morrell et al. (2002, Experiment 3) reflected a hit rate difference with an effect size of 0.91 but was accompanied by no criterion shift.

Nor are d' differences *necessary* for a criterion shift. It was discussed earlier, for example, that Hirshman (1995) detected a criterion shift, in the form of a lower criterion for weak items in a pure list of weak items than in a mixed list of weak and strong items, despite the fact that d' for the weak items was approximately equal in the two conditions. In another experiment of Hirshman (Hirshman & Arndt, 1997, Experiment 1), the participants rated the concreteness of high- and low-frequency words within a list. In a subsequent recognition test, the hit rates were approximately equal in the high- and low-word-frequency conditions. However, false alarm rates were higher in the high-frequency condition, signifying a criterion shift. These findings tend to deny that there is a simple relation between sensitivity (e.g., d') and the appearance of a criterion shift.

Criterion shifts: Delay versus strength. It will be important, in the future, to determine why delay manipulations might promote criterion shifts, whereas strength manipulations do not. In this regard, both delay manipulations (Reder & Wible, 1984) and strength manipulations affect the current activation of the stimulus. Delay, however, may carry additional effects. First, the similarity between encoding and retrieval contexts diminishes with delay (Gillund & Shiffrin, 1984, p. 28; Mensink & Raaijmakers, 1988). Contextual similarity, as captured by principles such as encoding specificity (Tulving & Thomson, 1973), is a central factor in memory performance. Second, delay manipulations may be particularly subject to metacognitive reasoning (e.g., Hasher & Griffin, 1978). Thus, people might judge that they cannot retrieve stimuli after a lengthy poststudy interval, an observation that might constrain them to modify their decision criterion.

It is noteworthy that other sets of memory variables likewise exhibit subtly different effects. For example, memory is superior both for strong items, as compared with weak ones, and for items in short lists, as compared with those in long lists (Gillund & Shiffrin, 1984). However, there are qualitative differences between the effects of list length and item strength. In one study, Shiffrin et al. (1995) examined the joint impact of length and strength on item recognition. False alarms increased systematically with list *length*, operationalized as the number of words in a category embedded in a list, but hits were invariant. Conversely, hits increased with item *strength*, and false alarms were invariant. The present study indicates that, analogously to the strength–length contrast, delay and strength may differ qualitatively in their effects.

Likelihood Ratio Models

Some psychological models (e.g., Glanzer, Adams, Iverson, & Kim, 1993; Shiffrin & Steyvers, 1997) posit a likelihood ratio decision criterion, rather than the strength-of-evidence criterion embraced in our analyses. Likelihood ratio criteria (e.g., beta, log beta) reflect the odds that a probe has originated from the target distribution. However, they do not address the position of the

Table 4
Mean Difference Between False Alarm Rates (%) in the Immediate and Delay Conditions as a Function of Delay

	Delay		
Study	20 min	40 min	2 days
Present	−1.8	2.0	15.4
Singer, Gagnon, & Richards (2002)	1.2	2.1	4.8

criterion on the strength axis, an issue of central concern in this study. Consider the arrangement in Figure 3. Distributions are shown for immediate targets, delay targets, and distractors. Furthermore, the delay criterion is lower than the immediate criterion, reflecting the results of Experiments 3 and 4. Each criterion is placed at the point of intersection between its corresponding target distribution and the distractor distribution. Therefore, a likelihood ratio criterion would be equal for the immediate and the delay conditions. This outcome might suggest the *absence* of a criterion shift, but in doing so it would disregard the distinct placement of the criteria on the familiarity scale.

Two features of Stretch and Wixted's (1998b) findings also weighed against the use of a likelihood ratio criterion. First, in that study, likelihood ratio calculations suggested that the manipulation of stimulus strength exerted opposite effects on response bias between lists (Experiment 1), as opposed to within lists (Experiment 4). Second, in simulation analyses, the only likelihood ratio model that fit Stretch and Wixted's (1998b) data presumed that their participants ignored the distinctive coloring of the strong and the weak stimuli. It is possible that the participants ignored the information provided by the color cue; but it seems odd that they would, given that (1) every participant was able to report which color indicated the strong condition and (2) likelihood ratio models assume that participants compute likelihood ratios on an item-by-item basis. Stretch and Wixted (1998b) proposed that neither of these implications was plausible, a conclusion that tended to deny the appropriateness of the likelihood ratio analyses.

Processes of Shifting the Criterion

Proposals of criterion shifts, particularly within lists, raise the question of the feasibility, in information-processing terms, of continually adjusting the recognition decision criterion. A variety of observations bear on this issue. In principle, recognition criteria must be positioned on the basis of (1) a small number of test items or even (2) people's assumptions about the test distributions. It is not practicable, in either laboratory or real-world recognition, to examine an entire test set before setting the decision criterion. This supposition is consistent with the signal detection notion that the criterion is placed in a favorable, if not optimal, location (Stretch & Wixted, 1998b).

Of course, basing criterion location on a small number of items does not entail continually adjusting it. However, theorists have proposed that criterion location is updated with reference to a very small number of recent distractors (Gillund & Shiffrin, 1984, pp. 54–55) or just a single, preceding distractor item (McNamara & Diwadkar, 1996). Proponents of within-list criterion shifts take into consideration the fact that continual adjustment of the criterion is posited to be quite demanding of cognitive resources (Hockley & Niewiadomski, 2001). These claims are consistent with the growing body of evidence of within-list criterion shifts (Hockley & Niewiadomski, 2001; Reder, 1987, 1988; Singer et al., 2002).

A Methodological Implication

Finally, it is noteworthy that frequent criterion adjustments, whether controlled or automatic, can hinder the measurement of the impact of experimental manipulations upon probe familiarity. As was discussed earlier, hit/false-alarm profiles are ordinarily compatible with a variety of signal detection interpretations, and criterion shifts can only further complicate data interpretation. To address this circumstance, Shiffrin et al. (1995) implemented a procedure with which to preclude their participants' adjustment of a decision criterion. They constructed lists that presented items from different categories. However, the categories were difficult for the participants to detect because they were (1) many in number and (2) very subtle (e.g., words that varied from the prototype word *sip* either in their first or last letter, but not in both). This manipulation successfully prevented the participants from customizing the criterion to the category and facilitated the investigators' study of list length and list strength effects. The study of *text retrieval* is complicated by the multilevel character of the text representation (van Dijk & Kintsch, 1983), the differential loss of those levels with increased delay (Kintsch et al., 1990), and the possibility of criterion shifts. Therefore, preventing criterion shifts in the manner of Shiffrin et al. may provide the opportunity to appreciably advance our understanding of text retrieval.

REFERENCES

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

Battig., W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, **80**(3, Pt. 2).

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

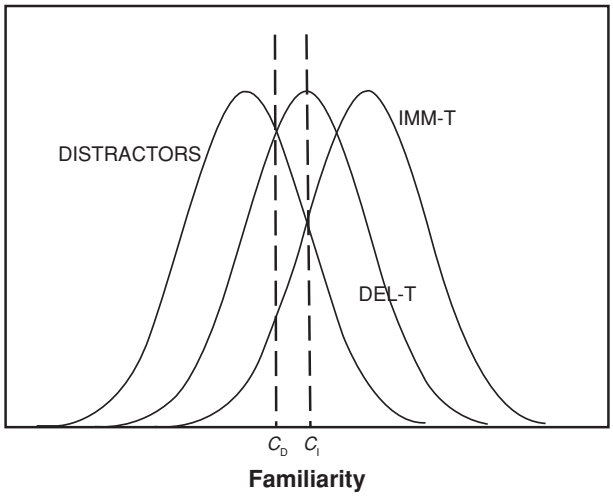

Figure 3. Hypothetical arrangement of immediate and delay criteria (C_I and C_D, respectively), plus distributions of distractors, immediate targets (IMM-T), and delay targets (DEL-T).

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.
Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, **100**, 546-567.
Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 500-513.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Hasher, L., & Griffin, M. (1978). Reconstructive and reproductive processes in memory. *Journal of Experimental Psychology: Human Learning & Memory*, **4**, 318-330.
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.
Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list strength effect. *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.
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.
Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory & Language*, **29**, 133-159.
Lorch, R. F., Jr. (1981). Effects of relation strength and semantic overlap on retrieval and comparison processes during sentence verification. *Journal of Verbal Learning & Verbal Behavior*, **20**, 593-611.
Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
McNamara, T. P., & Diwadkar, V. A. (1996). The context of memory retrieval. *Journal of Memory & Language*, **35**, 877-892.
McNicol, D. (1972). *A primer of signal detection theory*. Sydney: Allen & Unwin.
Mensink, G.-J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, **95**, 434-455.
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.
Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review*, **73**, 44-58.
Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, **99**, 518-535.
Reder, L. M. (1982). Plausibility judgements versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, **89**, 250-280.
Reder, L. M. (1987). Strategy-selection in question answering. *Cognitive Psychology*, **19**, 90-134.
Reder, L. M. (1988). Strategic control of retrieval strategies. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 22, pp. 227-259). San Diego: Academic Press.
Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 45-77). Mahwah, NJ: Erlbaum.
Reder, L. M., & Wible, C. (1984). Strategy use in question-answering: Memory strength and task constraints on fan effects. *Memory & Cognition*, **12**, 411-419.
Roediger, H. L., III, & McDermott, K. B. (1999). False alarms about false memories. *Psychological Review*, **106**, 406-410.
Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 267-287.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**, 145-166.
Singer, M. (1979). Temporal locus of inference in the comprehension of brief passages: Recognizing and verifying implications about instruments. *Perceptual & Motor Skills*, **49**, 539-550.
Singer, M. (1991). Question-answering strategies and conceptual knowledge. *Bulletin of the Psychonomic Society*, **29**, 143-146.
Singer, M., Gagnon, N., & Richards, E. (2002). Question answering strategy: The effect of mixing test delays. *Canadian Journal of Experimental Psychology*, **56**, 41-57.
Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1397-1410.
Stretch, V., & Wixted, J. T. (1998b). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1379-1396.
Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, **91**, 68-111.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 359-380.
van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
Wixted, J. T., & Gaitan, S. C. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior*, **30**, 289-305.
Wixted, J. T., & Stretch, V. (2000). The case against a criterion shift account of false memory. *Psychological Review*, **107**, 368-376.
Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, **11**, 616-641.
Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1341-1354.

NOTE

1. Furthermore, Stretch and Wixted (1998a) proposed that the fanning patterns were qualitatively more consistent with a likelihood-ratio model than with two competing analyses. More generally, however, and in agreement with the present analysis, Wixted and his colleagues favored a strength-of-evidence decision criterion over a likelihood-ratio criterion (Stretch & Wixted, 1998b; Wixted & Gaitan, 2002).

(*Continued on next page*)

APPENDIX
Stimuli for Experiments 1–3

Occupations	Birds	Mammals	Countries
1. doctor	1. robin	1. dog	1. China
2. lawyer	2. sparrow	2. cat	2. India
3. teacher	3. cardinal	3. horse	3. Indonesia
4. dentist	4. eagle	4. cow	4. Brazil
5. engineer	5. crow	5. lion	5. Russia
6. professor	6. canary	6. tiger	6. Pakistan
7. carpenter	7. parakeet	7. elephant	7. Bangladesh
8. salesperson	8. hawk	8. pig	8. Japan
9. nurse	9. wren	9. bear	9. Nigeria
10. psychologist	10. oriole	10. mouse	10. Mexico
11. plumber	11. parrot	11. rat	11. Germany
12. accountant	12. pigeon	12. deer	12. Philippines
13. clerk	13. starling	13. sheep	13. Vietnam
14. farmer	14. woodpecker	14. giraffe	14. Egypt
15. laborer	15. vulture	15. goat	15. Iran
16. chemist	16. swallow	16. zebra	16. Turkey
17. merchant	17. chicken	17. squirrel	17. Ethiopia
18. banker	18. dove	18. wolf	18. Thailand
19. physicist	19. duck	19. donkey	19. France
20. fireman	20. owl	20. rabbit	20. Italy
21. manager	21. thrush	21. leopard	21. Ukraine
22. electrician	22. falcon	22. fox	22. Burma
23. judge	23. jay	23. buffalo	23. Spain
24. mechanic	24. pheasant	24. moose	24. Colombia
25. secretary	25. finch	25. rhinoceros	25. Poland
26. bricklayer	26. ostrich	26. camel	26. Argentina
27. mathematician	27. flamingo	27. antelope	27. Tanzania
28. architect	28. lark	28. hippopotamus	28. Sudan
29. pharmacist	29. peacock	29. monkey	29. Algeria
30. minister	30. penguin	30. raccoon	30. Kenya
31. writer	31. raven	31. llama	31. Morocco
32. janitor	32. swan	32. skunk	32. Peru
33. artist	33. crane	33. cheetah	33. Afghanistan
34. baker	34. goose	34. jaguar	34. Nepal
35. psychiatrist	35. chickadee	35. beaver	35. Venezuela
36. grocer	36. pelican	36. gazelle	36. Uganda
37. sailor	37. stork	37. elk	37. Romania
38. barber	38. warbler	38. chipmunk	38. Iraq
39. cook	39. quail	39. coyote	39. Malaysia
40. pilot	40. nightingale	40. hamster	40. Ghana

Male Names	Cities	Body Parts	American States
1. John	1. Bombay	1. leg	1. California
2. Bob	2. Seoul	2. arm	2. Texas
3. Bill	3. Jakarta	3. head	3. Florida
4. Jim	4. Manila	4. eye	4. Illinois
5. Tom	5. Istanbul	5. foot	5. Pennsylvania
6. Joe	6. Shanghai	6. nose	6. Ohio
7. Dick	7. Moscow	7. finger	7. Michigan
8. Mike	8. Tokyo	8. ear	8. Georgia
9. George	9. Tehran	9. hand	9. Virginia
10. Jack	10. Lima	10. toe	10. Massachusetts
11. Harry	11. London	11. mouth	11. Indiana
12. Steve	12. Beijing	12. stomach	12. Washington
13. Larry	13. Bogota	13. hair	13. Tennessee
14. Frank	14. Calcutta	14. neck	14. Missouri
15. Paul	15. Santiago	15. heart	15. Wisconsin
16. Sam	16. Baghdad	16. knee	16. Maryland
17. Dave	17. Sydney	17. chest	17. Arizona
18. Fred	18. Melbourne	18. liver	18. Minnesota
19. Mark	19. Berlin	19. brain	19. Louisiana
20. Charles	20. Rome	20. lung	20. Alabama

APPENDIX (Continued)

Male Names	Cities	Body Parts	American States
21. Jerry	21. Osaka	21. tooth	21. Colorado
22. Ed	22. Toronto	22. elbow	22. Kentucky
23. Don	23. Nairobi	23. shoulder	23. Oklahoma
24. Bruce	24. Havana	24. face	24. Oregon
25. Gary	25. Budapest	25. tongue	25. Connecticut
26. Carl	26. Hamburg	26. ankle	26. Iowa
27. Henry	27. Johannesburg	27. throat	27. Mississippi
28. Ron	28. Warsaw	28. back	28. Kansas
29. Ken	29. Quito	29. intestine	29. Arkansas
30. Al	30. Vienna	30. hip	30. Utah
31. Jeff	31. Brisbane	31. lip	31. Nevada
32. Ralph	32. Barcelona	32. wrist	32. Nebraska
33. Dan	33. Montevideo	33. kidney	33. Idaho
34. Peter	34. Perth	34. pancreas	34. Maine
35. Ted	35. Phoenix	35. thigh	35. Hawaii
36. Tony	36. Milan	36. bone	36. Montana
37. Ray	37. Stockholm	37. muscle	37. Delaware
38. Brian	38. Amman	38. waist	38. Alaska
39. Tim	39. Dallas	39. thumb	39. Vermont
40. Wayne	40. Pretoria	40. chin	40. Wyoming

Diseases	Chemical Elements	Colors	Female Names
1. cancer	1. oxygen	1. blue	1. Mary
2. tuberculosis	2. hydrogen	2. red	2. Ann
3. measles	3. nitrogen	3. green	3. Jane
4. polio	4. sodium	4. yellow	4. Judy
5. mumps	5. iron	5. orange	5. Carol
6. smallpox	6. helium	6. black	6. Barbara
7. leukemia	7. silver	7. purple	7. Cathy
8. mononucleosis	8. potassium	8. white	8. Linda
9. malaria	9. copper	9. pink	9. Joan
10. syphilis	10. carbon	10. brown	10. Nancy
11. pneumonia	11. sulphur	11. violet	11. Betty
12. flu	12. chlorine	12. gray	12. Jeanne
13. leprosy	13. zinc	13. turquoise	13. Susan
14. diphtheria	14. magnesium	14. gold	14. Karen
15. diabetes	15. aluminum	15. indigo	15. Pat
16. arthritis	16. fluorine	16. maroon	16. Joyce
17. cholera	17. phosphorus	17. chartreuse	17. Dianne
18. hepatitis	18. calcium	18. tan	18. Sally
19. rickets	19. uranium	19. lavender	19. Sharon
20. rabies	20. lead	20. beige	20. Alice
21. tetanus	21. argon	21. amber	21. Lynne
22. dysentery	22. neon	22. aqua	22. Ellen
23. encephalitis	23. mercury	23. magenta	23. Helen
24. scurvy	24. boron	24. olive	24. Ruth
25. typhus	25. lithium	25. rose	25. Margaret
26. asthma	26. manganese	26. mauve	26. Janet
27. bronchitis	27. iodine	27. scarlet	27. Pam
28. epilepsy	28. tin	28. fuchsia	28. Carolyn
29. diarrhea	29. bromine	29. azure	29. Gail
30. meningitis	30. radium	30. crimson	30. Jill
31. rheumatism	31. krypton		
32. anthrax	32. cobalt		
33. gangrene	33. barium		
34. glaucoma	34. nickel		
35. jaundice	35. platinum		
36. acne	36. beryllium		
37. angina	37. xenon		
38. anorexia	38. plutonium		
39. earache	39. radon		
40. eczema	40. silicon		

(Manuscript received August 4, 2003;
revision accepted for publication January 25, 2005.)