Strength-based criterion shifts in recognition memory

MURRAY SINGER
University of Manitoba, Winnipeg, Manitoba, Canada

In manipulations of stimulus strength between lists, a more lenient signal detection criterion is more frequently applied to a weak than to a strong stimulus class. However, with randomly intermixed weak and strong test probes, such a criterion shift often does not result. A procedure that has yielded delay-based within-list criterion shifts was applied to strength manipulations in recognition memory for categorized word lists. When participants made semantic ratings about each stimulus word, strength-based criterion shifts emerged regardless of whether words from pairs of categories were studied in separate blocks (Experiment 1) or in intermixed blocks (Experiment 2). In Experiment 3, the criterion shift persisted under the semantic-rating study task, but not under rote memorization. These findings suggest that continually adjusting the recognition decision criterion is cognitively feasible. They provide a technique for manipulating the criterion shift, and they identify competing theoretical accounts of these effects.

The ability to recognize previously encountered stimuli is a fundamental index of proper memory function. For the past half century, signal detection theory (SDT; Green & Swets, 1966) has been widely embraced as contributing to a model of recognition (Banks, 1970; Parks, 1966). According to this perspective, people evaluate studied stimuli (targets) and unstudied ones (lures) on a scale of strength. Probes are accepted if they exceed a specific degree of strength, which is treated as the decision criterion. Criteria that demand only modest probe strength may be thought of as lenient, and those that demand great strength as strict.

There is considerable interest in the variables that affect the placement of the SDT criterion. People readily comply with instructions to adopt lenient or strict criteria (Estes & Maddox, 1995; Verde & Rotello, 2007). The composition of the test set is important: Recognition lures that are closely related in meaning to the targets demand a more strict criterion than do unrelated lures (Benjamin & Bawa, 2004; S. Brown & Steyvers, 2005; Reder, 1987). Criterion placement becomes systematically more lenient as test delay increases (Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Singer, Gagnon, & Richards, 2002; Singer & Wixted, 2006). People can use feedback about response accuracy in a recognition test to adjust their decision criterion (Rhodes & Jacoby, 2007; Verde & Rotello, 2007). The application of different criteria as a function of such variables is called a criterion shift (e.g., J. Brown, Lewis, & Monk, 1977; Hirshman, 1995).

Also implicated in criterion shifts is the strength of stimulus encoding. The relationship, however, is a complex one. When strong and weak items are studied and tested in separate lists (between-list designs), criterion shifts are regularly detected (e.g., Hirshman, 1995; Stretch & Wixted, 1998b, Experiment 1). In contrast, when strongly and weakly encoded items, whether studied together or not, are intermixed in the test lists (within-list designs), a uniform criterion is often applied to all items (e.g., Stretch & Wixted, 1998b, Experiments 3–5; Verde & Rotello, 2007).

This study scrutinized a discrepancy in which within-list criterion shifts have been detected under delay manipulations (Singer et al., 2002; Singer & Wixted, 2006) but not under strength manipulations (Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998b). The discrepancy is puzzling: Because immediately tested items ought to be represented more strongly than those tested at a delay, the impact of delay ought to parallel that of strength manipulations. Accordingly, the present strategy was to determine whether the materials and study procedures of Singer and Wixted would yield criterion shifts in the context of repetition-based strengthening. As was noted earlier, repetition has frequently not promoted within-list criterion shifts (Stretch & Wixted, 1998b). The remainder of the introduction will present a synopsis of SDT, a summary of findings concerning strength-based criterion shifts, and an overview of the present approach.

Signal Detection Theory

The central principle of SDT is that memory test probes, both targets and lures, are evaluated on a strength-of-evidence scale. Each class of probes, such as (1) targets representing different experimental conditions and (2) lures, is associated with a distinct strength distribution. Figure 1 depicts the distributions of lures and just one type...
of target (e.g., Green & Swets, 1966; Macmillan & Creelman, 2005; McNicol, 1972; Parks, 1966). The strength of each test probe is compared with a decision criterion positioned on the strength continuum, which in Figure 1 is shown at position C. (Alternative criterion $C_w$ will be addressed shortly.) Probes with a strength greater than or equal to $C$ receive the reply “yes.”

According to this analysis, recognition performance is imperfect because the distributions of targets and lures typically overlap. Of course, more targets are accepted than lures. However, lures with strength in excess of $C$ are erroneously accepted, outcomes that are labeled false alarms. Targets with a strength less than $C$ are missed.

Classic SDT (Green & Swets, 1966) treats lures as having the standard normal distribution (mean = 0, variance = 1). Targets are also considered to be normally distributed. The measure $d'$ is the standardized distance between the means of the target and lure distributions.

SDT also posits that target variance is equal to that of the lures. However, empirical findings deny the equal-variance signal detection model: They typically reveal the ratio of lure variance to target variance to be only .75–.80 (Glanzer, Kim, Hilford, & Adams, 1999; Malmberg, 2002; Ratcliff, Sheu, & Gronlund, 1992). In spite of this, the equal-variance assumptions provide a satisfactory approximation for explicating the SDT model (e.g., Hicks & Marsh, 1998, p. 1108; Hirshman, 1995, p. 307; Stretch & Wixted, 1998a; Wixted & Stretch, 2004).

Figure 1 also portrays the impact of a criterion shift. Consider again criterion $C$, plus the alternative criterion, $C_w$. A numerically higher or stricter criterion is typically associated with a strong condition ($C$) than with a weak condition ($C_w$). The proportion of the lure distribution that exceeds $C$ is smaller than the proportion that exceeds $C_w$. Therefore, fewer false alarms result with a strict than with a lenient criterion.

Criterion shifts are often associated with the recognition memory mirror effect, which refers to the frequent observation that a more memorable stimulus class exhibits both more hits and fewer false alarms than does a less memorable class (Glanzer & Adams, 1985; Glanzer, Adams, Iverson, & Kim, 1993). The mirror effect can signify an underlying criterion shift: The higher false alarm rate for the less memorable class than for the more memorable one may reflect the more lenient criterion of the former. However, mirror effects can also result from distribution shifts, rather than criterion shifts (e.g., Dobbins & Kroll, 2005, p. 1187, Model E), and criterion shifts can appear in the absence of the mirror effect (Hirshman, 1995; Hirshman & Arndt, 1997; Joordens & Hockley, 2000).

**Strength-Based Criterion Shifts**

The strength-based criterion shift refers to the application of a stricter criterion to strongly encoded stimuli than to weakly encoded ones. This shift is often interpreted with reference to the participant’s awareness of the encoding strength of a stimulus class (e.g., Hirshman, 1995; Stretch & Wixted, 1998b; Verde & Rotello, 2007). Rapidly accumulating evidence suggests, however, that the phenomenon is relatively complex: As was mentioned earlier, criterion shifts are readily detected when strong and weak test items appear between lists, but the evidence is less consistent for within-list manipulations.

**Between-list criterion shifts.** In one study, mixed study lists of strongly and weakly encoded words (exposure durations of 2.0 and 0.4 sec, respectively) were contrasted with study lists presenting weakly encoded words exclusively (pure-weak; Hirshman, 1995). In both conditions, only weak items were encountered during recognition testing. A more lenient criterion was applied in the pure-weak study list condition than in the mixed condition, even though the $d'$ scores were about equal. Hirshman concluded that people’s sensitivity to the memorability of the studied stimuli influenced them to apply a more lenient criterion in the pure-weak condition.

Strength-based criterion shifts are also promoted by repetition manipulations of stimulus strength. The participants in Stretch and Wixted (1998b, Experiment 1) encountered blocks of words presented either three times each (strong) or once (weak). Recognition tests that immediately followed these blocks revealed decision criteria that were more strict in the strong condition than in the weak condition. Likewise, the participants in Verde and Rotello (2007) viewed study lists of low-frequency words that were presented either once or four times. Recognition testing segregated the strong and weak items into separate blocks. False alarms were fewer (signifying a more strict criterion) across all test blocks when the first block presented strong items (Experiment 2) than when it presented weak items (Experiment 4). Thus, between-list criterion shifts are frequently detected when strength is manipulated in terms of study duration, repetition, and other variables (Hicks & Marsh, 1998).

**Within-list criterion shifts.** The evidence concerning within-list strength-based criterion shifts (WLSBCSs) is more equivocal. Some within-list nonshift results have been documented by the same researchers who measured between-list criterion shifts. In this regard, Stretch and Wixted (1998b, Experiments 3–5) strengthened (five repetitions)
high-frequency words but not low-frequency words (one repetition) or vice versa. Furthermore, strong words were consistently colored red, whereas weak words appeared in blue. Recognition testing presented intermixed strong and weak probes: The targets and the lures of each lexical frequency class maintained their study coloring. In spite of the saliency of the strength manipulation, no criterion shift was detected. Likewise, no shift appears for strong items from one taxonomic category versus weak items from another (Morrell et al., 2002). In the aforementioned study of Verde and Rotello (2007), although a more strict criterion was adopted when the initial test block presented strong rather than weak words, the participants did not shift their criterion when they proceeded from a strong test block to a weak one (Experiments 1–3) or vice versa (Experiment 4).

Other studies, in contrast, have documented WLSBCSs. Hockley and Caron (2007) presented either words for a short duration (weak, 500 msec) and pictures for a long duration (strong, 3,000 msec) or vice versa. When the pictures and words were randomly intermixed in recognition testing, more lenient criteria were detected for the weak than for the strong stimuli, both when participants performed a follow-up remember/know judgment (Experiment 2; cf. Gardiner & Java, 1991; Tulving, 1985) and when they did not (Experiment 3). Morrell et al. (2002), in contrast, did not detect criterion shifts in the differential strengthening of words versus pictures. Also suggestive is Hockley and Niewiadomski’s (2007) report that different criteria were applied to strong and weak single-word versus word-pair test probes. The test lists were mixed with regard to stimulus type (i.e., single or pair), although, in most of the experiments, not with regard to strength.

The WLSBCS has been also detected in studies of retrieval from text. In these investigations, participants have read texts either once or multiple times and then have performed recognition tests that intermixed sentences from the repeated tests (strong probes) and ones from the unrepeated texts (weak probes). The results have revealed the mirror effect of higher hits and lower false alarms for strong than for weak test sentences (Brainerd, Reyna, & Estrada, 2006; McDermott & Chan, 2006). The false alarm differences were interpreted as reflecting criterion shifts.

As was mentioned earlier, delay manipulations are relevant in this context because item representation is likely to be stronger at a short than at a long test delay. In this regard, Singer and Wixted (2006) documented a within-list criterion shift in the comparison of items tested immediately versus at a 2-day delay (although not at shorter intervals). It should be noted that delay may differ from other strength manipulations in terms of people’s metacognitive reasoning about immediate and delayed memory tests (Hasher & Griffin, 1978; Reder, 1982).

**Conclusion.** There has been equivocal evidence as to whether strength manipulations support within-list criterion shifts. For example, whereas a repetition manipulation of strength yielded invariant decision criteria in the recognition of categorized words (Morrell et al., 2002), an analogous delay manipulation revealed a criterion shift (Singer & Wixted, 2006). Therefore, Singer and Wixted’s method for investigating the delay-based criterion shift was applied to a repetition manipulation. The considerable similarity between the method used by Morrell et al. to produce a nonshift outcome and the method in Experiment 1 suggested that Experiment 1 would not yield a criterion shift.

### EXPERIMENT 1

The goal of Experiment 1 was to determine whether distinct decision criteria would be applied to intermixed weakly and strongly encoded words that represented particular taxonomic categories. The method was analogous to the one that Singer and Wixted (2006, Experiment 3) applied to the delay-based criterion shift. During study, participants viewed lists of exemplars of different categories, with all the words from a category being encountered consecutively. Strong categories were immediately repeated a second time, whereas weak categories were presented only once. After all study was complete, the participants encountered blocks of recognition testing. Each block presented the intermixed testing of one strong category (e.g., occupations) and one weak one (e.g., cities).

**Method**

**Participants.** The participants were 118 male and female native English-speaking students of introductory psychology at the University of Manitoba. They took part in partial fulfillment of a course requirement.

**Materials.** The stimuli were based on the materials in Singer and Wixted (2006). Those authors identified 40 words that represented each of the following 10 categories: birds, body parts, chemical elements, cities, countries, diseases, male names, mammals/four-footed animals, occupations, and American states. The words were all familiar and readily associative with their categories. Multisyllable examples (e.g., blue jay) and ones that named their category (e.g., bluebird) were avoided. Place names and proper names were capitalized, as usual.

The items for cities, countries, and American states were based on population rankings determined from Internet sources. The items for the remaining categories were selected according to the frequency orderings of the norms in Battig and Montague (1969), subject to the restrictions mentioned earlier. In addition, 30 color words, selected in a similar manner, made up the stimuli for a practice category. The stimuli appear in the Appendix of Singer and Wixted (2006).

Four counterbalanced lists were derived from the experimental materials. In List 1, equal numbers of categories were randomly designated as strong (studied twice) or weak (studied once). To determine an item’s status as target or lure, the items in each category were first arranged in order of frequency (Battig & Montague, 1969) or population. Then, from each successive pair of items, one word was randomly selected as a target and the other as a lure. This procedure yielded 20 targets and 20 lures per category.

Next, a list structure of 10 category positions was created. The strong categories were randomly assigned to Study Positions 1, 3, 5, 7, and 9 (e.g., 1, occupations; 3, birds; etc.). The weak categories were randomly assigned to the remaining even-numbered study positions. As such, the categories were encountered in strong–weak pairs. However, cities, states, and countries could not appear in the same pair, nor could mammals and birds. For testing, all 40 items of each sequence of one strong category and one weak one (e.g., Categories 1 and 2) were combined to produce recognition test lists of 80 items each.

Lists 2–4 were constructed by cycling the materials of List 1 across the four strength × target–lure conditions, according to a Latin-square scheme.
Procedure. The participants first read instructions indicating that they would make pleasantness ratings about words organized according to their categories and then perform a recognition test. The instructions acknowledged that they might find rating pleasantness peculiar for such categories as natural disasters and diseases. The instructions also specified that each recognition block would include words from just two categories.

During study, each category was signaled with a message such as OCCUPATION WORDS, plus an instruction for the participants to place their fingers on the numeric keys 1 to 4 on the top keyboard row. The participant pressed the space bar to signal readiness to view the category. The first word appeared immediately. The participant had 3 sec during which to rate the word on a 4-point pleasantness scale (4 = highly pleasant), using the numeric keys. If no response was registered within the 3-sec limit, a tone sounded for 250 msec. The 20 target words of the category appeared in random order. Each word was preceded by a 250-msec interstimulus interval.

The five strong categories appeared twice in succession. Each presentation was treated as a distinct category (e.g., new-category signal), and the targets appeared in a new random order. The five weak categories were presented only once. With 20 targets per category, the participants thus studied a total of 300 experimental items (200 strong and 100 weak), preceded by the 15 items from the practice category. At the end of study, the participants performed a paper-and-pencil anagram solution task for 90 sec.

Recognition testing began with 15 targets and 15 lures from the practice category. Testing then continued in blocks. In each block, the participants encountered randomly intermixed words representing one strong category and its immediately following weak category (e.g., occupations and cities). Each category pair was preceded by a reminder for the participant to place her or his index fingers on the “yes” and “no” keyboard buttons, which were the “.” and “x” keys, respectively. The screen also showed a message such as RECOGNITION TEST ITEMS ABOUT CITIES AND OCCUPATIONS. A press of the space bar initiated a test block. The test items representing both categories of each pair (total of 80) were randomly intermixed. There were 10 experimental categories, so there was a total of five blocks of 1 strong category plus 1 weak one. Across the five strong–weak category pairs, 400 probes were encountered.

On each trial, a fixation “x” appeared at row 10, column 1 of the screen for 500 msec, followed immediately by the test word at the same location. The “yes” and “no” keys were used to indicate whether the test word had been studied. There was no answer time limit. A 500-msec interval followed the participant’s response, and the fixation point then began the next trial. Each test block comprised 80 words.

The practice items (colors) were intermixed with those of another category during neither study nor testing. Practice was intended to familiarize the participant with the experimental procedures.

Results

The data of 5 individuals who exhibited no ability to discriminate targets and lures were disregarded. In addition, the data of two pairs of participants who were inadvertently assigned the same participant numbers were irretrievably combined and lost. All data analyses were based on the results of the remaining 113 participants.

Of central concern here were the acceptance rates for the targets and the lures—that is, the hits and false alarms, respectively. To evaluate criterion placement, the signal detection measure $z_c$ was derived from the false alarms. The $z_c$ measure identifies the position on the strength-of-evidence continuum beyond which a test probe is labeled as a target: It is the $z$ score of the criterion relative to a strength of 0. Criterion measures that evaluate the participant’s bias to answer “yes” or “no,” such as $C$ (Macmillan & Creelman, 2005), are especially suitable for evaluating manipulations such as the base rate appearance of targets and lures in different screen regions (Rhodes & Jacoby, 2007). However, they do not unequivocally indicate whether the criterion has shifted locations (Dobbins & Kroll, 2005; Singer & Wixted, 2006, p. 128). Assessment of criterion shift with respect to either $z_c$ or raw false alarm rates is commonplace in this domain (Hockley & Caron, 2007; Morrell et al., 2002; Singer & Wixted, 2006; Verde & Rotello, 2007). The strength parameter $d'$ was also computed.

An ANOVA was applied to all the measures, with strength (unrepeated, repeated) functioning as a within-subjects variable and list as a between-subjects variable. Effects involving the counterbalancing list variable will not be reported. Throughout, hits and false alarm statistical outcomes were usually congruent with the corresponding signal detection measures. They will be reported only in incongruent cases. Finally, the statistical criterion of $\alpha = .05$ will be used throughout, unless otherwise indicated.

The mean pleasantness rating time was 995 msec ($SD = 519$ msec). The acceptance rates and the derived signal detection measures are shown in Table 1. In ANOVAs of the signal detection measures, strength was a within-subjects variable, and list was a between-subjects variable. Most importantly, the $z_c$ criterion was lower for weak than for strong items [$F(1,109) = 69.83, MS_e = 0.040$]. In addition, $d'$ was greater in the strong than in the weak condition [$F(1,109) = 100.03, MS_e = 0.076$].

Discussion

The data revealed a clear within-list criterion shift: A more lenient criterion was applied to the weak items than to the strong items. This outcome was surprising, considering that other studies of word recognition have revealed uniform, rather than shifted, criteria using strength manip-

<table>
<thead>
<tr>
<th>Measure</th>
<th>Hits</th>
<th>False Alarms</th>
<th>$d'$</th>
<th>$z_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strength</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td>Strong</td>
<td>.857</td>
<td>.01</td>
<td>.198</td>
</tr>
<tr>
<td></td>
<td>Weak</td>
<td>.838</td>
<td>.03</td>
<td>.259</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>Strong</td>
<td>.827</td>
<td>.02</td>
<td>.206</td>
</tr>
<tr>
<td></td>
<td>Weak</td>
<td>.782</td>
<td>.02</td>
<td>.259</td>
</tr>
</tbody>
</table>
ulations of high- versus low-frequency words (Stretch & Wixted, 1998b) and categorically organized words (Morrell et al., 2002). Close scrutiny, however, reveals numerous methodological differences among ostensibly similar studies in this domain. Word stimuli may be random (Stretch & Wixted, 1998b) or categorized (Morrell et al., 2002). Stimuli are alternatively strengthened in terms of repetition (Stretch & Wixted, 1998b), duration of study (Hockley & Caron, 2007), list length (Ohrt & Gronlund, 1999), and other manipulations. In the case of categorized word stimuli, all the items stemming from a category may be studied consecutively, or they may be intermixed with items from another category (Morrell et al., 2002). Participants may be instructed simply to study the stimuli for a memory test or else to make judgments about the semantic or lexical status of each word (Hirshman & Arndt, 1997; Joordens & Hockley, 2000).

Among those methodological differences, the procedure in Experiment 1 differed most conspicuously from that in Morrell et al. (2002) in two ways: in the consecutive (or blocked) study of items from each category and in the participants’ semantic orienting task. Experiments 2 and 3, respectively, inspected the contribution of those procedures to differences between the present results and those in Morrell et al.

**EXPERIMENT 2**

Study of the exemplars of each category was blocked in Experiment 1, whereas items from a category pair were randomly intermixed, or unblocked, in the study of Morrell et al. (2002). In Experiment 1, the reappearance of a category was highly conspicuous: Immediately after viewing exemplars of, for example, birds, the participant studied them all again. Arguably, repetition of a given exemplar would be less conspicuous in unblocked study. Therefore, Experiment 2 replaced the blocked study of Experiment 1 with unblocked study. The purpose of the experiment was to determine whether, under conditions of unblocked study, the criterion shift in Experiment 1 would be abolished.

**Method**

The participants were 72 naive individuals from the same pool as before. The only difference between Experiment 2 and Experiment 1 was that, during study, the items of the strong–weak category pairs were randomly intermixed. For example, in studying the strong category *occupations* and the weak category *cities*, the participants encountered 60 randomly ordered stimuli: the 20 occupation terms, occurring twice each, and the 20 cities, appearing once each.

**Results and Discussion**

The mean pleasantness rating time was 1,012 msec ($SD = 519$ msec). Seven participants exhibited little or no discrimination in the recognition test. The recognition results of the remaining 65 participants appear in Table 1. As in Experiment 1, there was a criterion shift: The criterion measure $z_c$ was lower in the weak condition than in the strong condition [$F(1,61) = 26.67, MSe = 0.046]$. In addition, $d'^\prime$ was greater in the strong condition than in the weak condition [$F(1,61) = 4.96, MSe = 0.087$].

Experiment 2 replaced the blocked study of Experiment 1 with the unblocked study procedure in Morrell et al. (2002, Experiment 1). In spite of this, a within-list criterion shift was detected, extending the variety of methods documented to yield this outcome. Experiment 3 addressed impact on recognition memory of a semantic orienting task (Experiments 1 and 2), as opposed to rote study (Morrell et al., 2002, Experiment 1).

**EXPERIMENT 3**

The instructions in Experiments 1 and 2 directed the participants to judge the pleasantness of each stimulus word, whereas those in Morrell et al. (2002) promoted a rote strategy. It is well known that the semantic processing of words and text yields qualitatively different memory representations than does surface (or otherwise shallow) processing (Craik & Lockhart, 1972). According to one pertinent account in this domain, semantic processing of a stimulus results in both surface detail and gist representations of the stimulus, whereas shallow orientation tasks foster mainly surface representations of the sort that are reflected in tests of implicit memory (Jacoby, 1983). This proposal applies both to the learning of word lists (Brainerd, Reyna, & Mojardin, 1999; see also Anderson, 1983) and to discourse comprehension (van Dijk & Kintsch, 1983). Gist representations are more available to awareness than are lower level perceptual representations. To the extent that criterion adjustment is a resource-demanding process (e.g., Dobbins & Kroll, 2005; Rhodes & Jacoby, 2007), gist representations would support the criterion shift more readily than would surface representations.

Accordingly, pleasantness rating versus rote study was manipulated as a between-subjects variable in Experiment 3. A second priority of Experiment 3 was to precisely replicate the method in Morrell et al. (2002). Several methodological details, apart from those already discussed, distinguished Experiments 1 and 2 from Experiment 1 in Morrell et al. The procedure in Experiment 3 eliminated those differences.

**Method**

**Participants.** Eighty-five individuals who had not participated in Experiments 1 and 2 were sampled from the same pool as before.

**Materials.** In Experiment 1 in Morrell et al. (2002), the materials were derived from pools of 40 occupation words and 40 countries. Accordingly, the present stimuli comprised the occupation and country words in Experiments 1 and 2. No other categories were examined.

**Procedure.** The participants were randomly assigned to perform either the (rote) “study” or the pleasantness-rating orienting task. In both task conditions, items from the two categories were unblocked (randomly intermixed) during study.

To further harmonize the procedure with that in Morrell et al. (2002), learning comprised one presentation of the weak items and five repetitions of the strong items (rather than the two repetitions in Experiments 1 and 2). With this arrangement, the participants encountered a total of 120 words. Following Morrell et al., there was no practice category, and no anagram task intervened between study and test.

Apart from the latter differences, the procedure for the rating condition was identical to that in Experiment 2. Consider next the rote
study condition. Across five experiments in this study and Singer and Wixted (2006), the mean time to rate the pleasantness of the stimulus word, using a 3-sec time limit, was 1,069 msec. In the interest of equating stimulus exposure time under the two orienting tasks, presentation time during the learning phase in the rote study condition was set to 1,100 msec. All other features of the study task procedure were identical to those in Experiment 2. As in Experiment 2, counterbalancing was achieved by cycling categories and items across the four strength \( \times \) target–lure conditions, yielding four stimulus lists. Recognition testing followed the exact procedure in Experiment 2. The test items were the 20 targets and 20 lures for each of the two categories, for a total of 80.

Results

Eight participants exhibited no ability to distinguish targets from lures. All the data analyses are based on the results of the remaining 77 participants.

The mean pleasantness rating time was 1,056 msec \((SD = 459\) msec). This closely approximated the 1,100-msec fixed learning time per item in the study condition.

The main results of Experiment 3 are presented in Table 2. On the \( z_c \) measure, there was a strength \( \times \) task interaction \([F(1,69) = 4.66, MS_e = 0.20]\). Tests of simple main effects indicated that, in the rating condition, \( z_c \) was lower in the weak than in the strong condition \([F(1,69) = 4.54, MS_e = 0.20]\). In the study condition, in contrast, the strength effect was not significant \((F < 1)\). There was also a main effect of task: \( z_c \) was significantly lower in the study task than in the rating task. The main effect of strength was not significant \((F < 1)\).

The \( d' \) measure exhibited the usual effect of strength \([F(1,69) = 88.31, MS_e = 0.27]\). In addition, \( d' \) was greater in the rating condition than in the study condition \([F(1,69) = 11.26, MS_e = 0.95]\). No other \( d' \) effects approached significance. An ANOVA of the hits likewise revealed main effects of strength and task, but a strength \( \times \) task interaction also emerged \([F(1,69) = 11.33, MS_e = 8.17]\). Tests of simple main effects revealed that, in the weak condition, the hit rate was greater for the rating task than for the study task \([F(1,69) = 15.08, MS_e = 8.27]\); but in the strong condition, the rating versus study hit rates did not differ significantly \((F < 1)\). The latter outcome may reflect that, at over 90%, the strong condition hit rates were approaching ceiling.

Discussion

The results of Experiment 3 confirmed the conjecture that the criterion shift is at least partly dependent on the learning orienting task. When the participants rated the stimuli, a criterion shift emerged, in the form of a \( z_c \) simple main effect of strength. Under an instruction simply to study the stimuli, in contrast, there was no such effect. This outcome is consistent with observations of the within-list strength-based criterion shift when participants have been instructed to judge, during study, the concreteness (Hirshman & Arndt, 1997) and pleasantness (Shiffrin, Huber, & Marinelli, 1995) of word stimuli. Likewise, Joordens and Hockley (2000) detected the effect in many conditions in which their participants made lexical decisions about word stimuli.

Also consistent with the enhanced mnemonic impact of a semantic orienting task was the main effect of task on the \( d' \) measure. Although the strength \( \times \) task interaction was, as was discussed earlier, significant in the hits analysis, it is more important that there was no strength \( \times \) task interaction in the \( d' \) ANOVA: Rather, strong \( d' \) clearly exceeded weak \( d' \) in both tasks.

It is noteworthy that Experiment 3 simultaneously replicated both (1) the criterion shifts in Experiments 1 and 2, in which the pleasantness-rating task was used, and (2) the nonshift result emerging from the rote study orientation in Morrell et al. (2002, Experiment 1). As was treated in the Method section, Experiment 3 adopted several features of Morrell et al.’s materials and procedure. This discounts factors such as number of studied category pairs, role of a practice category, and presence of an intervening anagram task as likely determinants of the criterion shift.

These results raise the question of how the representations that result from different study orienting tasks contribute to criterion shifts. This matter will be addressed in the General Discussion section.

GENERAL DISCUSSION

There is accelerating scrutiny of the placement of the signal detection decision criterion of recognition memory. One reason for this is that the evaluation of recognition probes on a strength-of-evidence continuum is postulated both by single-process recognition theories (e.g., Banks, 1970; Dunn, 2004; Gillund & Shiffrin, 1984; Parks, 1966; Shiffrin & Steyvers, 1997; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008) and dual-process theories (DeCarlo, 2002; Diana, Reder, Arndt, & Park, 2006; Wixted, 2007; Yonelinas, 1994). In this context, strength-based criterion shifts have been of particular interest. Intuition

<table>
<thead>
<tr>
<th>Measure</th>
<th>Hits</th>
<th>False Alarms</th>
<th>( d' )</th>
<th>( z_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( M )</td>
<td>( SE )</td>
<td>( M )</td>
<td>( SE )</td>
</tr>
<tr>
<td>Rating</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strong</td>
<td>.904</td>
<td>.01</td>
<td>.142</td>
<td>.01</td>
</tr>
<tr>
<td>Weak</td>
<td>.841</td>
<td>.03</td>
<td>.076</td>
<td>.01</td>
</tr>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strong</td>
<td>.934</td>
<td>.02</td>
<td>.177</td>
<td>.02</td>
</tr>
<tr>
<td>Weak</td>
<td>.713</td>
<td>.02</td>
<td>.155</td>
<td>.02</td>
</tr>
</tbody>
</table>

**Table 2**

Experimental Measures As a Function of Strength and Task (With Standard Errors) in Experiment 3
suggestions that people ought to be able to capitalize on conspicuous differences among the strengths of item classes to assign different criteria to those classes. Whereas that has been the typical finding in between-list testing (Hirshman, 1995; Stretch & Wixted, 1998b), within-list experiments have produced both shift and nonshift results (Benjamin & Bawa, 2004; Brainerd et al., 2006; Hockley & Caron, 2007; Morrell et al., 2002; Stretch & Wixted, 1998b; Verde & Rotello, 2007).

This study was especially motivated by the observation of within-list criterion shifts in manipulations of delay (e.g., Singer et al., 2002; Singer & Wixted, 2006), but not in manipulations of strength (e.g., Morrell et al., 2002; Stretch & Wixted, 1998b). Accordingly, the materials and procedure in Singer and Wixted were adapted to a repetition-based strength manipulation. It was initially anticipated that no criterion shift would emerge. The present experiments did not fulfill that expectation; criterion shifts were detected in all instances except the rote study condition in Experiment 3.

It is noteworthy that theorists have suggested that frequent adjustment of the criterion may be curtailed by the lack of (1) cognitive resources (Hockley & Niewiadomska, 2001; Morrell et al., 2002; Stretch & Wixted, 1998b), (2) useful information (Verde & Rotello, 2007), and (3) motivation (Morrell et al., 2002; Verde & Rotello, 2007). Notwithstanding those proposals, there is computational and empirical evidence (Gillund & Shiffrin, 1984, and McNamara & Dswadkar, 1996, respectively) that criterion placement can be based on a very small number of recent lures or even the single most recent one (see Treisman & Williams, 1984, for a review). To continually shift the decision criterion, one must be able to derive information from the current probe. The impact of recent lures on criterion placement is one indication of the availability of the needed information, although it does not establish that people will use that information to shift the criterion.

Within-List Criterion Shifts:
The Role of Orienting Task

The present results raise the issue of how the learner’s orienting task might regulate criterion shifts. Experiment 3 was introduced in the context of a multirepresentation hypothesis: A semantic orienting task during learning produces both surface detail and robust gist representations, whereas rote learning yields predominantly surface representations. If gist representations, but not surface representations, were available to the conscious processing posited to enable criterion adjustment, this analysis would account for why a criterion shift would appear in the rating condition and not in the rote study condition in Experiment 3. The multirepresentation hypothesis is also consistent with the fact that the rating $d'$ exceeded the study $d'$.

Critiques of the multirepresentation view are also available. First, the rote study instructions in Experiment 3 may be argued not to prohibit semantic analysis. Against that possibility, however, is appreciable evidence that deep (i.e., semantic) and shallow orienting instructions promote qualitatively different processing (Craik & Lockhart, 1972). Second, and perhaps more telling, is that the multirepresentation view would seem to predict a negligible difference between the strong and weak $d'$s in the rote study task, because a $d'$ difference ought to permit participants to shift the criterion. Ostensibly inconsistent with this analysis, Experiment 3 revealed a strong versus weak $d'$ difference in the rote task, and yet no criterion shift.

Alternative explanations of the present results therefore merit consideration. A second approach derives from differentiation theories of memory (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Differentiation models subscribe to global matching, entailing the single process of comparing the memory probe against the complete contents of episodic memory. Critically, differentiation posits that strength manipulations, such as repetition, reduce the noisiness of existing representations. As a result, a strong target discriminates its corresponding memory representation from those of other targets better than does a weak target (Bruno, Higham, & Tey, 2009; Criss, 2006).

However, the converse is true for strong and weak lures: Strong lures match the representation of strong targets particularly poorly. The net effect is a lure distribution shift, as depicted in Figure 2. Whereas strong targets are higher on the strength continuum than are weak targets, the opposite is true for lures (see also Bruno et al., 2009; Criss, 2006, Figure 3). One uniform criterion is applied to all probes.

The configuration of Figure 2 yields higher hits and lower false alarms in the strong than in the weak condition, thereby capturing the mirror effect. Figure 2 refers to a between-list manipulation. With the added assumption that words (either targets or lures) are preferentially matched against their own category, differentiation theory could also accommodate the present rating task data (Morrell et al., 2002; see also Dennis & Humphreys, 2001). In that event, however, differentiation would likewise predict a mirror effect for the rote study task in Experiment 3, an

![Figure 2. Distribution shift model posited by a differentiation analysis of learning repeated and unrepeated items. The weak lures exhibit a higher mean strength than do the strong lures. Targ, target; Stg, strong; Wk, weak; C, criterion.](image_url)
outcome that was not detected. That does not preclude the possibility of differentiation solutions to the data profile in Experiment 3.

According to a third analysis, people base criterion position on cue strength only when the subjective memorability of the learning context is low (Bruno et al., 2009). In that regard, Bruno et al. detected a criterion shift for words studied for 0.5 sec (low subjective memorability), but not for those studied for 3.0 sec (high subjective memorability). This account would address the results of Experiment 3 if the postlearning recognition task were less expected in the rating condition than in the study condition. That seems somewhat unlikely, because all the participants were informed about the recognition test before learning. However, rating the pleasantness of 120 stimuli might sufficiently background the prospect of a recognition test so that the participants would perceive recognition testing as a low-memorability experience.

These and other contemporary analyses of recognition memory offer principles that have the potential to address the impact of orienting task on the criterion shift. Discriminating among these accounts will require further investigation.

In conclusion, these results carry several interrelated ramifications. They provide further evidence of WLSBCSs. They identify the learning orienting task as one variable that modulates that effect, and, in so doing, they provide a procedure for manipulating the effect. This, in turn, provides the opportunity to distinguish among the theoretical accounts of the mechanisms that promote criterion shifts, such as the three accounts that were just considered.

AUTHOR NOTE

These data were presented at the 49th Annual Meeting of the Psychonomic Society, Chicago, November 2008. 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. I thank Lauren Unik and Bee Leng Oh for conducting the experimental sessions and tabulating the data. I am grateful to John Wixted and David Huber, who offered their insights about these issues during my research leave at the University of California, San Diego. I have benefited from discussions with William Hockley about the criterion shift. Correspondence should be directed to M. Singer, Department of Psychology, University of Manitoba, Winnipeg, MB, R3T 2N2 Canada (e-mail: m_singer@umanitoba.ca).

REFERENCES


Jacoby, L. L. (1983). Remembering the data: Analyzing interactive pro-


**NOTE**

1. Brainerd et al. (2006) used a complicated method. The pertinent data are those for the testing of the first narrative in the verbatim condition (Table 2). Furthermore, those data combine the results for between-list (separate testing of high [strong] and low [weak] stimuli) and within-list (mixed) testing. However, the authors reported (p. 677) that all testing groups exhibited similar result profiles.

(Manuscript received July 10, 2008; revision accepted for publication April 12, 2009.)