Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Memory and Language 51 (2004) 159–172

Journal of Memory and Language

www.elsevier.com/locate/jml

Distractor plausibility and criterion placement in recognition

Aaron S. Benjamin* and Sameer Bawa

Department of Psychology, 603 E. Daniel St., University of Illinois at Urbana-Champaign Champaign, IL 61820, USA

Received 10 February 2004; revision received 1 April 2004
Available online 5 May 2004

Abstract

To set an optimal decision criterion on a test of recognition, a subject must estimate the degree to which they can discriminate previously studied from unstudied stimuli. To do so accurately, the subject must assess not only their mastery of the material but also the extent to which the distractors yield mnemonic evidence that makes them difficult to discriminate from studied items. In these experiments, we manipulated the degree to which the distractor set overlapped semantically or categorically with the previously learned stimuli, and examined the effects on criterion placement. When plausibility was manipulated by semantic membership, subjects who took tests with more plausible distractors set a higher criterion; for all other manipulations of plausibility, there were no between-subject effects of plausibility on criterion placement. However, over multiple test opportunities, subjects increased their criterion when the tests became more difficult, but they did not lower their criterion when the test became easier. This asymmetry obtained with picture and word categories, and suggests that online monitoring of recognition performance modulates the criterion shift.

From an information-processing perspective, the task of recognition can be conceptualized as a problem of discrimination. Is this face one that we have seen before in our class? Did this memo cross our desk today? In both cases, the stimulus that we are judging is familiar to some greater or lesser degree, either because the face is indeed one that we have seen in our class or the memo is one that did cross our desk today, because of other recent exposures (the face belongs to someone whom we have seen in another class before), or even because of similarity to other recently encountered stimuli (this memo is not terrifically different from any one of a large number of such documents that we see daily).

When we encounter a stimulus that prompts a recognition decision, *global matching* theories presume that we concatenate all available evidence about past encounters with that stimulus into a single scalar value, and transform that value into a recognition decision. The most prominent aspect of such theories of recognition is the separation of the mnemonic and decision components of the task. Other *multiprocess* theories postulate multiple sources of evidence that are not combined prior to the decision stage, and consequently may have complex decision mechanisms. The study of decision criteria are relevant to any model of recognition, but are most easily conceptualized in the context of global matching theories; thus, our discussion of decision processes will employ language more consistent with models of recognition that postulate such single-process global matching.

Because variables can influence either the discriminability of stimuli or the amount or quality of evidence that a subject demands before making a positive recognition decision (or both), the Theory of Signal Detection (TSD) has been imported into recognition research (e.g., Banks, 1970; Egan, 1958; Lockhart & Murdock, 1970) from engineering (Peterson, Birdsall, & Fox, 1954) by way of psychophysics (Green & Swets, 1965).

In TSD, mnemonic evidence from a test stimulus is evaluated relative to putative probability distributions of

* Corresponding author. Fax: 1-217-244-5876.
E-mail address: asbenjam@s.psych.uiuc.edu (A.S. Benjamin).

0749-596X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2004.04.001

evidence for unstudied and studied stimuli. Discriminability is represented by the degree of overlap of the two probability distributions, whereas the decision component of the decision is represented by the placement of a decision criterion (either in the units of the evidence variable or in terms of the likelihood ratio). Much of the early use of TSD in psychology used this technique to examine how variables affected performance when the concomitant effects of the decision component of the task are partialled out. Differences between subjects or between conditions in response criterion placement were thus treated as a nuisance variable. In fact, in the context of threshold models that approximate the continuous form of TSD-based models of performance, response biases are typically factored out by parameters termed '*corrections* for guessing' (e.g., Blackwell, 1963).

In the contexts of both psychophysics and recognition memory, important questions can be posed about how response criteria are established and what factors influence the updating or revision of such criteria. In research on recognition, these questions influence the theoretical interpretation of interesting laboratory phenomena (such as the revelation effect; Hockley & Niewiadomski, 2001; Westerman, 2000) and bear on important practical considerations in domains such as eyewitness testimony (e.g., Ebbesen & Flowe, 2002; Gronlund, in press).

Criterion setting in recognition

Those factors that can influence the decision component of recognition can be roughly divided into three types. First, there are stimulus-based factors, such as word frequency or picture clarity, that influence memorability and consequently also criterion placement. Second, encoding-based factors, such as study time or orienting task, influence learning and potentially also criterion-setting. Finally, the structure of the recognition test—for example, payoff structure, a priori distractor probability, or distractor type—can influence recognition performance and may influence criterion-setting. In each of these cases, the manipulation is expected to influence criterion placement because it can be shown that optimal criterion placement varies with that manipulation. In that sense, criterion placement can be evaluated relative to an ideal decision-maker. So an analysis of criterion-setting can focus not only on how variables influence placement, but also how they affect criterion optimality.

Stimulus-based factors

There is strong evidence that factors inherent to certain stimulus classes appear to influence criterion placement. For example, Brown, Lewis, and Monk (1977) showed that subjects confidently and often reject distractors that are idiosyncratically memorable to them, such as the names of close relatives or the names of towns in which they had lived. This fact suggests a strategic setting of a particularly stringent criterion for subjectively memorable stimuli, consistent with the strategy of proposing that "I would remember it, if I had studied it." In other words, subjects set recognition standards commensurate with the degree of learning that they would expect from an exposure during the study phase (see also Ghetti, 2003; Strack & Bless, 1994).

This explanation can be readily extended to the domain of word frequency, in which uncommon words elicit both higher hit rates (HR) and lower false-alarm rates (FAR) than do common words. This "mirror effect" (Glanzer, Adams, Iverson, & Kim, 1993) can be interpreted similarly: Uncommon events yield more accurate memory; thus a more stringent criterion should be—and is—set for such stimuli (Benjamin, 2003). This effect also obtains in a "pure-list" experiment in which word frequency is manipulated between-subjects or between-lists (Gorman, 1961; McCormack & Swenson, 1972).

The left panel of Fig. 1 depicts a TSD model that can account for such effects. In this model, the evidence variable is assumed to be normally distributed, and the two distributions (corresponding to studied and unstudied status) are assumed to have equal variance. Lower-frequency words, which are more memorable than higher-frequency words, attain higher levels of memorability after study, thus yielding a higher distribution mean. In addition, the response criterion for low-frequency words is assumed to lie to the right of the criterion for high-frequency words. Behaviorally, this is apparent in the lower FAR for low-frequency foils.[1]

This decision strategy can also be shown to approximate optimality from the perspective of the decision-maker who is trying to maximize the probability of a correct response.[2] Correct responses include "yes" responses to old items and "no" responses to new items:

[1] Another TSD model of the mirror effect assumes a lower distribution mean for new LF than new HF items; in that model, mirror effects obtain with a single criterion. The reasons for selecting one model over another are beyond the scope of this article, but it worth noting that the variable-criterion model is not a currently popular one. Ratcliff and McKoon (2000) even went so fair as to claim that "... everyone agrees that this is an unsatisfactory solution." (p. 574) Personally, I am not so convinced, but readers who are offended by the current formulation are encouraged to interpret this model as an illustration of how mirror effects *could* obtain with variable criteria.

[2] This analysis is summarized from the recent volume by Wickens (2001). For a fuller treatment of this topic, the reader is encouraged to read that book.

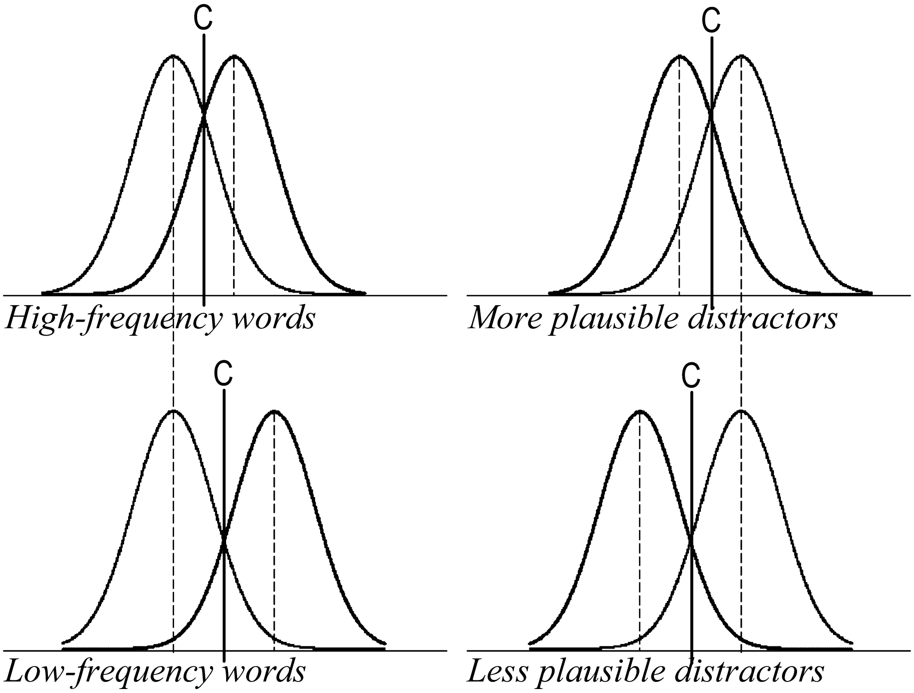

Fig. 1. Left panel: Signal-detection model for the mirror effect for word frequency. New words are equally familiar regardless of frequency. Low-frequency words (bottom panel) gain more mnemonic evidence as a function of study than do high-frequency words (top panel), and elicit a more stringent recognition criterion. Right panel: Signal-detection model for the recognition of previously studied words on a test with more plausible distractors (top panel) and a test with less plausible distractors (bottom panel). Recognition criterion is more lax in the condition in which discriminability is greater, unlike in the left panel, in which the criterion is more conservative in the condition with greater discriminability.

$$p_c = [p(\text{old})][1 - F_{\text{old}}(C_N)] + [1 - p(\text{old})][F_{\text{new}}(C_N)], \quad (1)$$

in which C_N is the response criterion, $F(x)$ represents the cumulative probability function and p(old) denotes the a priori probability of an old item on the test. Taking the derivative of this function with respect to C_N and solving for 0 yields an extremum at

$$\beta = \frac{1 - p(\text{old})}{p(\text{old})}, \quad (2)$$

in which β is the likelihood ratio at the response criterion. Assuming normal probability distributions with equal variance, the optimal standardized criterion is thus

$$C_N = \frac{d'}{2} - \frac{\log \frac{p(\text{old})}{1-p(\text{old})}}{d'} \quad (3)$$

in which d' is the standardized distance between the distribution means. Thus, when old and new items are equally probable, the optimal criterion lies exactly halfway between the means of the two distributions. For stimuli that elicit greater discriminability, then, the criterion should lie to the right of the criterion for stimuli that are less discriminable. Subjects are apparently able to use stimulus-based information to set their criterion judiciously.

Note, however, that such behavior requires no strategic adjustment on the part of subjects during the experiment itself. Memorable or LF words yield higher HR because of distinctiveness; they yield lower FAR because of preconceived notions about the relation of memorability and frequency to retention. To address the question of whether subjects can adjust a recognition criterion dynamically, within the context of a single experimental session, for example, it is necessary to experimentally manipulate memorability during study. Studies that do so are discussed in the next section.

Encoding-based factors

Although subjects are able to use information inherent to a to-be-tested word to set their criterion appropriately, the question remains of just how sophisticated such a strategy is. Whereas memorability information is inherent to a to-be-tested word in paradigms investigating recognition for word frequency, in other paradigms memorability is conferred by the nature or extent of the study activity rather than the stimulus itself. Are subjects able to optimize their criteria in situations in which memorability is experimentally manipulated?

Hirshman (1995) showed that criterion placement can vary with the overall memorability for a study list. Subjects studied one list of words at a rate of 400 ms/word and another list that contained some words presented for 400 ms and others for 2000 ms durations. He showed that subjects used a more stringent criterion in the latter condition, and attributed that effect to a strategy on the part of the subjects of estimating the approximate range in familiarity values across all items, and placing a criterion at some fixed interval in that range. This result does suggest that subjects incorporate assessments of encoding operations into decision standards on recognition, but only in a rather crude manner: subjects use a rough estimate of the range of evidence values to establish a single criterion value for all test stimuli.

Whether subjects can, on a single test, employ multiple criteria derived from encoding-based factors is less clear. Strength-based mirror effects, in which HR and FAR move in opposite directions with an experimental manipulation of learning, rather than a stimulus memorability variable, are common in paradigms in which strength is manipulated on a between-subjects or between-lists basis (Stretch & Wixted, 1998). The more appropriate analogue to the word-frequency case outlined earlier is a *within-list* manipulation of strength, however, and in this domain the results are less clear. Stretch and Wixted manipulated degree of learning via a study repetition manipulation and printed repeated words in red and unrepeated words in blue. Although the expected difference in HR between red and blue words obtained on a later test, no difference in FAR to new red and new blue words was evident. In a more telling demonstration, Morrell, Gaitan, and Wixted (2002) employed the same manipulation with two different categories of words (professions and locations). Again, well encoded words elicited a higher HR, but there was no difference in FAR between new professions and new locations.

However, when perceptual (Zaki & Nosofsky, 2001) or semantic (Robinson & Roediger, 1998; Shiffrin, Huber, & Marinelli, 1995; Strack & Bless, 1994) category size is manipulated within-lists by varying the number of studied items from a category, the FAR is typically higher to foils from categories for which more items were studied. One interpretation of such an effect is that as cue overload (e.g., Watkins & Watkins, 1975) increases with the number of studied items per category cue, memory for individual exemplars decreases and the criterion is naturally lowered to match the attenuated discriminability of items from such categories.

Another interpretation of these category-size results is simply that mnemonic evidence for a test item accrues as a function of its similarity to other studied items; thus foils from categories that were well represented in the study list yield an overall higher global match to the study episode and consequently elicit a higher FAR (e.g., Hintzman, 1988; Murdock, 1993; Nosofsky, 1989). However, in a paradigm in which the distractors are highly similar to words from the previous study list and in which FAR is consequently much higher than the typical recognition experiment (Roediger & McDermott, 1995), within-list manipulations of strength analogous to those employed by Morrell et al. (2002) elicited a lower FAR to strong than weak lists on a single test (Benjamin, 2001). This result is inconsistent with a constant–criterion global matching view and suggests that subjects can adjust their recognition criterion on an item-by-item basis by using information about the degree of learning for categorically related material.

Manipulations of encoding have yielded mixed evidence for the ability of subjects to place their recognition decision criterion appropriately. Over multiple study-test trials, subjects shift their criteria in a manner consistent with the overall degree of learning from the previous study episode (Hirshman, 1995). However, when items of differing strengths are mixed within a test block, some data suggest no criterion shift (Morrell et al., 2002; Stretch & Wixted, 1998) whereas other show such a shift (Benjamin, 2001).

Test-based factors

The few examinations of the role of test-based factors in criterion setting come from experiments in which either payoffs were manipulated or the proportions of old and new items were varied. Although early studies appeared to reveal that subjects are insensitive to manipulations of prior probabilities (e.g., Healy & Jones, 1975), later experiments with greater power revealed reliable but often smaller than optimal effects of both prior probabilities (e.g., Ratcliff, Sheu, & Gronlund, 1992) and payoffs (e.g., Healy & Kubovy, 1978).

However, a series of experiments by Wallace (1980, 1982; Wallace, Sawyer, & Robertson, 1978) revealed that HR were approximately the same between standard recognition tests in which both targets and distractors were presented and distractor-free tests in which every test item had been previously studied. An appropriately lower criterion in the distractor-free condition would predict higher HR, thus one interpretation of such results is that subjects are insensitive to even the most dramatic variation in prior probabilities.

A final source of evidence bearing on this issue comes from studies of text retrieval in which multiple stories were read at different times by subjects, and recognition tests for sentences were administered at varying intervals following the stories (Singer, Gagnon, & Richards, 2002). They found that, on the same test, subjects applied a more conservative criterion for sentences from stories read immediately prior to the test than for sentences read at an earlier time.

The data bearing on how subjects assimilate information about the makeup of a test into response criteria obviously are quite mixed. In addition, a vast majority of the data come from experiments in which prior probabilities were varied. Such experiments leave open the possibility that criterion differences, when evident, are a result of deliberate attempts to match response proportions to the estimated proportion of different types of items, as is apparent in certain psychophysical tasks (Thomas & Legge, 1970). This is not a fault of these experiments, but, because this probability-matching view is not easily applied to many of the experiments described in the previous sections in which item proportions were held constant, and because it mispredicts the effects of degree of learning on criterion shift in the list-strength paradigm (Hirshman, 1995), it is desirable to examine the role of test-based factors in a paradigm in which test item proportions are invariant between conditions and in experiments more obviously analogous to the cases in which stimulus- and encoding-based factors drive criterion shift.

In the experiments reported here, we examine criterion shift in a paradigm in which neither probabilities nor payoffs were varied, which makes our procedure more similar to the majority of memory experiments that examine criterion shift as a function of a manipulation of learning. However, unlike those experiments, we manipulate test characteristics, rather than study variables. By doing so, we rule out the possibility that any strategic aspects of the recognition decision are set before the experiment (as might be the case with varying stimulus characteristics, such as word frequency) or are preplanned during the study phase (as might be the case with manipulations of learning). In each of our experiments, the study phase and materials are identical across conditions, but the tests vary in makeup: some tests contain foils that are highly plausible as prior study items (i.e., members of the same categories) and other contain more implausible foils (i.e., members of different categories).

A TSD conceptualization of such an experiment is represented in the right panel of Fig. 1. Here learning is not varied, so the distribution of evidence values for studied items is in the same location on the evidence axis regardless of condition. However, because the less plausible (LP) and more plausible (MP) distractors differ in the degree to which they yield mnemonic evidence for prior study, the distribution for LP unstudied items overlaps less with the studied item distribution than does the distribution for MP unstudied items. From the decision-maker's perspective, this difference is similar to the case portrayed in the left panel of the figure, in which learning or memorability is varied. In both cases, the manipulation causes the overlap to vary, which will affect discriminability.

However, whereas manipulations at study that increase discriminability also increase the value of the optimal criterion placement, test manipulations that increase discriminability should decrease the point as which the optimal criterion should be placed. Because neither study status nor payoffs are varied in these experiments, the optimal criterion again lies at the intersection of the distributions (as per Eq. (3)). Most importantly, because the procedure is identical for all subjects until the test itself, any differences in actual criterion placement can be attributed to strategies employed during the test itself.

Experiment 1

In this first experiment, we examine criterion placement in the context of semantically categorized lists of words. After studying multiple sub-lists of categorized words, each subject took two recognition tests. The distractors on one test were drawn from the same categories represented in the study list and the distractors on the other test were drawn from a different, and unstudied, set of categories. Using a two-test procedure allows the decomposition of criterion placement effects into between-subject differences on different first tests and appropriate (or inappropriate) shifts from that test to a second test with a different makeup. However, because our manipulation is one of test composition, there is a notable difference between these experiments and the typical word-frequency experiment that should be noted. In those experiments, the different test items are typically intermixed within a single test. In these experiments, we have chosen to vary distractor type between lists (but within subjects) to provide the greatest opportunity for criterion differences to reveal themselves. Thus, the current data do not bear on the question of whether subjects can revise their criterion on an item-by-item basis, which is one of the controversial assumptions in the interpretation of the word-frequency mirror effect presented earlier.

The criterion index discussed in Introduction (C_N) measures the standardized distance from the mean of the Gaussian evidence distribution for unstudied items to the criterion, and is thus defined as

$$C_N = -Z(\mathrm{FAR}), \tag{4}$$

in which $Z(x)$ is the inverse of the Gaussian distribution cumulated to probability x. This measure is appropriate in situations in which the distribution of evidence values for the unstudied items is presumed to be identical between conditions (see Banks, 1970; who called this measure C_j), but is not for our studies in which the placement of this distribution should vary. In our experiments, we explicitly manipulate the relative position of the new item distribution by varying the nature of the distractor set, but the placement of the evidence distribution is constant between conditions because we do not manipulate learning. Thus, we assess criterion relative to the mean of the old, rather than the new, distribution

$$C_O = d' - C_N, \tag{5}$$

in which C_O represents the standardized distance between the mean of the studied evidence distribution and the criterion. However, the bulk of evidence addressing the shapes of the underlying probability distributions in recognition suggests that the variances are not equal, thus rendering d' a biased measure of discriminability. However, simple algebraic manipulation of Eq. (5) reveals that an equivalent formulation is

$$C_O = Z(\text{HR}) \tag{6}$$

which, by virtue of using only one of the dependent measures from the experiment, is uncontaminated by (possibly) incorrect assessments of the relationship between the distributions. In other words, the quality of this estimate is independent of the relative variances of the distributions.[3]

There was one additional difference between our tests and the typical recognition test. Because we wanted to maximize the opportunity for the manipulation to affect criterion placement, subjects were given a short interval ($2\frac{1}{2}$ min) prior to each test to peruse the entire list of test items in Experiments 1 and 2 (but not Experiment 3). If subjects use estimates of evidence range (Parducci, 1984) or distribution means (Hintzman, 1994) to inform their criterion placement, this "preview" period should enhance the opportunity to use such a strategy.

Method

Subjects

Thirty-six undergraduate students from the University of Illinois participated to receive course credit. The age of the subjects ranged from 18 to 22 years.

Design

Each subject participated in a single study phase and two recognition tests, utilizing a 2 (word type: studied words, distractors) × 2 (plausibility of distractors: MP v. LP) within-subjects design. The order of the tests was counterbalanced across subjects, and thus constituted a between-subjects factor.

Materials

The studied words were nouns obtained from word lists normed by Rosch (1975). The lists were slightly amended to yield a study list containing 10 semantic categories, each composed of 15 words. The study words were obtained by choosing 10 category members. For example, for the insect category, the words *wasp*, *mosquito*, *ladybug*, *spider*, *dragonfly*, *grasshopper*, *termite*, *ant*, *cockroach*, and *cricket* were included. Three different randomized orders of the study categories were created; one-third of the subjects received each study version. Microsoft PowerPoint was used to present the words.

Each test consisted of 100 words, 50 of which were previously studied and 50 of which were distractors. No category names appeared during study or test. Five words from each previously studied 10-word categorical list were randomly chosen to appear on each of the two recognition tests, and no previously studied word appeared on both tests. The assignment of previously studied words to test was counterbalanced across subjects. The distractors for the MP test were comprised of five additional words from each of the category sub-lists. The distractors for the LP test were words that were semantically unrelated to any of the 10 presented category lists. The first and the second half of both tests contained exactly 25 studied words and 25 distractors. Two different randomized test versions were created for each condition.

The pairing of the test and study phase versions were counterbalanced in such a manner that both versions of the MP test were paired with both versions of the LP test. Each pair of tests was paired equally often with each study version, creating 12 different study phase-test/test combinations. Each test was printed on a single sheet of paper.

Procedure

Subjects were tested individually in a small, well-lit room. Prior to the study phase, subjects read instructions on the computer screen informing them that they were about to be presented a long series of words, and that they were to try to remember the words as well as possible. Subjects began the study phase by clicking the mouse button.

During the study phase, each word was presented for 4 s, with a 1s ITI between words. Between the last word of a given sub-list and the first word of the next sub-list, there was a 2s ITI. At the conclusion of the study phase, the computer instructed the subjects to alert the experimenter.

After a short distraction phase (approximately 7 min) consisting of basic arithmetic problems, subjects were given the recognition test. Through oral instructions, subjects were informed that prior to beginning the test,

[3] Strictly speaking, this is only true if the criterion in question is on an evidence scale, rather than a likelihood scale. However, the evidence that criteria are often invariant on an evidence scale (Morrell et al., 2002) and that the predictions of a likelihood scale theory are not borne out in detail (Stretch & Wixted, 1998) is strongly supportive of this assumption.

Table 1
Proportions of endorsed words and signal-detection parameter estimates as a function of prior study, distractor test type, and test order (Experiments 1, 2, and 3)[a]

Experiment and test order	Measure or parameter and test type							
	Hit rates		False-alarm rates		est d'		est C_O	
	MP	LP	MP	LP	MP	LP	MP	LP
Experiment 1								
MP–LP	**.76**	.74	**.14**	.01	**1.92**	2.91	**.76**	.71
LP–MP	.74	**.88**	.17	**.02**	1.74	**3.47**	.71	**1.34**
Experiment 2								
MP–LP	**.73**	.70	**.13**	.01	**1.94**	2.76	**.69**	.60
LP–MP	.60	**.73**	.14	**.04**	1.48	**2.49**	.28	**.68**
Experiment 3								
MP–LP	**.92**	.90	**.13**	.00	**2.89**	3.70	**1.59**	1.40
LP–MP	.85	**.93**	.08	**.00**	2.72	**3.89**	1.19	**1.62**

Note that the data are ordered in columns by test type, not by test order. Performance on the first test is indicated by boldface type.
[a] Signal-detection parameters were estimated by changing all mean proportions of 0 to .01 and all mean proportions of 1 to .99. (This shift of .01 is equal to one-half the smallest possible difference in proportions on a test with 50 old and 50 new items.)

$2\frac{1}{2}$ min would be devoted to a reading period. During this time, subjects were instructed to read through the entire test. At the conclusion of the reading period, subjects were given a pencil and were asked to begin circling only the words that they thought had been presented in the study phase. There was no time limit. After the first test, they were administered the second test, which followed the same procedure (including the $2\frac{1}{2}$ min preview period).

Results

The results from all inferential statistical tests reported in this article are reliable at the $\alpha = 0.05$ level using two-tailed tests unless otherwise noted. Table 1 shows the raw proportion endorsement data from this and the following two experiments. The MP test elicited a higher FAR (M = .16) than did the LP test (M = .02) for both the MP–LP condition ($t[17] = 6.97$) and the LP–MP condition ($t[17] = 6.01$), suggesting that the manipulation of distractor plausibility successfully affected memory discriminability. Thirty-five of 36 subjects showed this effect.

The important comparison, of course, involves hit rates. There was an interaction between test order and distractor plausibility ($F[1,34] = 9.61$) such that HR did not differ between tests in the MP–LP condition ($t[17] = 1.16$) but did in the LP–MP condition ($t[17] = 4.23$). Five of 18 subjects in the MP–LP condition had a higher HR on the LP test, as predicted by an ideal-decision making model (Eq. (3)); two had equivalent HR, and the remaining 11 had a higher HR on the MP test. In the LP–MP condition, 15 of 18 subjects had a higher HR on the LP test. In addition, the between-subject comparison of HR on the first test also yielded a reliable difference ($t[34] = 2.88$).

Discussion

The data from this experiment were reparameterized as d' and C_O, and are shown in the right half of Table 1. The depiction conveys two effects that were apparent in the raw endorsement data: First, on the first test, subjects who took a test with more plausible distractors set a higher decision criterion than did subjects who took a test with less plausible distractors.[4] This result is consistent with the direction of optimal placement (see Eq. (3)). We address the strength and reliability of this result in the following experiments.

Second, subjects shifted their criterion to a more conservative position when progressing from the LP test to the MP test, but maintained a constant placement when moving from the MP to the LP test. This asymmetry was unexpected and poses a theoretical challenge that we will take up later in this article. Here it is worth noting that there are at least some conditions under which subjects do strategically adjust their recognition criterion in response to assessments of relative test difficulty. Most investigations of criterion placement have employed manipulations of learning, which

[4] Throughout this paper, inferential statistics are not provided for comparisons using TSD-derived measures. Because the magnitude of such measures is strongly tied to fundamentally arbitrary decisions about the treatment of ceiling and floor performance effects, the variability structure central to inferential comparisons between means is unstable across different choices about that treatment. In every case in which we describe an "effect" in such measures, it is tied to appropriate testing of the untransformed recognition data. These measures are reported because they clearly illustrate the ordinal effects of our manipulations on criterion estimates.

leave open the possibility that criteria are responsive to assessments during the learning phase or immediately thereafter. In addition, the a priori probabilities of old and new items were kept constant throughout this experiment, so we were additionally able to rule out strategic response matching explanations as a source of the criterion shift.

Experiment 2

This experiment replicates and generalizes the findings from Experiment 1. We sought to replicate the interaction of test type with test order using a new set of stimuli. In particular, we wanted to determine category membership on a somewhat more subtle basis than in Experiment 1, in which the test items could quite clearly be attributed to previously studied categories, or not. One possibility in Experiment 1 is that the first test elicited a between-subjects difference in criterion placement by virtue of the LP distractors being obvious noncandidates for membership in the studied categories, and that on the succeeding tests, all subjects simply reverted to some common, "all-purpose" criterion value. For Experiment 2, we wanted to minimize the opportunity for a between-subjects difference on the first test, and see if the asymmetry still affected the transition from the first to the second test.

To do so, we used stimuli that were related by virtue of being high associates to a common word, but not necessarily to one another. These stimuli were drawn from Stadler, Roediger, and McDermott (1999). The experimental methodology was very similar to that used in Experiment 1.

Method

Subjects

Thirty-five undergraduate students from the University of Illinois participated to receive course credit. The age of the subjects ranged from 18 to 22 years.

Design

Each subject participated in a study phase and two recognition tests, utilizing a 2 (word type: studied words, distractors) × 2 (plausibility of distractors: MP v. LP) within-subjects design. Again, counterbalancing test order yielded an additional between-subjects factor.

Materials

The studied words were nouns, verbs, and adjectives obtained from word lists normed by Stadler et al. (1999). The original lists were comprised of 15 words each. All of the words in a given list were semantically related to a single target word that was neither studied nor tested in the current experiment. For the purposes of this study, 10 lists were chosen. The study words were obtained by choosing the first 10 words from each of the original 15-word lists, for a total of 100 words. Three different randomly generated orders of the study categories were generated; one-third of the subjects received each version. Microsoft PowerPoint was used to present the words.

Each test consisted of 100 words, 50 of which were previously studied and 50 of which were distractors. Five words from each previously studied 10-word associative list appeared on the first recognition test, and the other five appeared on the second. The assignment of previously studied words to test was counterbalanced across subjects. The distractors for the MP test were comprised of the last 5 words of each of the original 15-word associative lists. The distractors for the LP test were words that were semantically unrelated to any of the 10 presented category lists (by our assessment). Each half of both tests contained exactly 25 studied words and 25 distractors. The test words were randomly ordered, and two different versions of each test were created.

The pairing of the test and study phase versions was counterbalanced in such a manner that both versions of the MP test were paired with both versions of the LP test. Each pair of tests was also grouped with each study phase, creating 12 different study phase-test/test combinations. Each test was printed on a single sheet of paper.

Procedure

Subjects were tested individually in the same conditions as described for Experiment 1. The words from a given category were presented in descending order of relatedness to the critical lure, according to normative data obtained from Stadler et al. (1999). Each word was presented for 4 s, with a 1s ITI between words. Between the last word of a given list and the first word of the next list, there was a 2s ITI. At the conclusion of the study phase, the computer instructed the subjects to alert the experimenter. The distraction and test phases proceeded exactly as described in the Procedure section for Experiment 1.

Results

Again, the effect of distractor plausibility was apparent in a larger mean FAR to MP ($M = .13$) than LP distractors ($M = .03$) for both the MP–LP ($t[17] = 6.20$) and the LP–MP group ($t[16] = 4.64$). Of the 35 subjects, 32 showed this effect. Critically, the interaction between test order and distractor plausibility on HR was again reliable ($F[1, 33] = 9.62$), and revealed the same form as in Experiment 1. For the MP–LP group, mean HR on the first (MP) test was not reliably different from the second (LP) test ($t[17] = 1.60$). Five out of 18 subjects had a higher HR on the LP test. For the LP–MP group, mean HR on the two tests did differ ($t[16] = 5.31$). All 17 subjects showed the effect. Unlike in Experiment 1, the mean HR between the two first tests (a between-subject comparison) did not differ ($t[33] = .10$).

Discussion

This experiment replicated the effect that subjects are willing to shift their criterion to a more stringent position when moving from an easy to a hard test, but not to shift it to a more lax position when moving from a hard to an easy test. This asymmetry obtained despite the fact that hit rates on the first test were not different, and suggests that the phenomenon represents more than a single test condition eliciting the effect. In fact, whereas the "odd" datum in Experiment 1was a high HR on the first test with less plausible distractors, in this experiment it was a low HR on the second test with more plausible distractors.

The between-subjects effect of distractor plausibility on first-test criterion placement did not replicate in this experiment. This null result[5] is, however, consistent with the findings of Morrell et al. (2002), who showed remarkably invariant measures of criterion with manipulations of learning. Whether this result reflects true equivalence or simply a much smaller effect as in Experiment 1, the conclusion is the same: whereas subjects in Experiment 1 accurately utilized semantic membership information from the test makeup to set relative decision criteria, subjects in this experiment, in which membership was determined associatively and was thus quite subtle (and perhaps more idiosyncratic), plausibility did not play a major role in determining relative criterion placement.

It also appears, in both Experiments 1 and 2, as though criteria become generally more conservative on the second test, regardless of the change in test makeup. This effect is revealed by the fact that a minority of subjects showed an increase in HR across tests, even in the MP–LP condition, in which that effect was predicted by the ideal decision-making model (Eq. (3)). This effect can be partially explained by the decrease in discriminability from the first to the second test, as is apparent in the memory discriminability parameters (est d') shown in Table 1. If subjects do not appropriately incorporate assessments of forgetting and proactive interference from the first test to the second, then a constant criterion in evidence space would yield a lower HR because the two old item distributions do not coincide in that space.

However, this effect should operate equivalently in both order conditions and thus cannot explain the interaction between test type and the order variable. If nothing more was in operation across tests than a constant criterion shift (regardless of test type) or forgetting, then we would not see the interaction that has appeared twice. Nonetheless, we sought one additional replication of the effect in Experiment 3 under conditions in which forgetting should be less apparent than in the first two experiments.

One interesting aspect of these data is that, although the form of the criterion shift is equivalent to that seen in Experiment 1, the pattern that emerges on the first test is different. In Experiment 1, a between-groups difference was apparent on the first test and disappeared as a result of the asymmetric criterion shift to the second test. In Experiment 2, in which the categories were less well defined, from the perspective of the subjects, no difference was apparent between groups on the first test, but the asymmetric shift yielded a difference on the second test. We have already commented on the fickle nature of criterion shifts on a single test (e.g., Morrell et al., 2002; cf. Benjamin, 2001); it is reassuring to see that the current finding replicates regardless of whether such a difference appears on the first test or not.

Experiment 3

In this experiment, we seek once again to replicate the interaction between test order and test type, and, in doing so, generalize our findings to another set of stimuli. We chose categorically organized pictures in part because they differed dramatically from the word stimuli used in the first two experiments, and partly because pictures are known to elicit such robust recognition memory performance that forgetting should be minimal (e.g., Standing, 1973). This fact tests the limits of the finding from the first two experiments, because such high levels of performance are often thought to take place under conditions in which familiarity is not a primary basis for the recognition decision, and also because criterion shifts play a relatively minor role in overt performance when discriminability is high.

We made two additional changes in this experiment. First, subjects were occasionally run in small groups (but in individual cubicles). Second, because the effects from the first two experiments were robust, we eliminated the preview period prior to the recognition tests, and administered the tests on an item-by-item basis. This change also stacks the deck against the detection of a criterion shift. If subjects were only able to adjust their decision standards because of the extra time that they were forced to devote towards an evaluation of test composition, then that shift should not be evident in this experiment. This change also makes it possible to examine whether the preview plays a role in the unstable effects of plausibility on criterion placement that we have observed across subjects.

[5] The power to detect an effect of equivalent size as in Experiment 1 was ~0.80 for this test. The power to detect an effect of 1/2 the magnitude as in Experiment 1 was ~.35.

Method

Subjects

Thirty-two undergraduate students from the University of Illinois participated to receive course credit. The age of the subjects ranged from 18 to 22 years old.

Design

Each subject participated in a study phase and two recognition tests, utilizing a 2 (picture type: studied pictures, distractors) × 2 (plausibility of distractors: MP v. LP) within-subjects design. The order of the tests was counterbalanced across subjects, yielding a between-subjects factor.

Materials

The studied pictures were of nouns obtained from various web sites. These pictures represented semantic categories but were not normed. Pictures were chosen based on their clearly depicting a chosen member of a semantic category. For the purposes of this study, 10 categories were chosen. Each categorical list was comprised of 15 pictures each. The study pictures were obtained by randomly choosing 10 pictures from each of the original 15-picture lists. For example, for the category *insects*, pictures of the following insects were shown: *butterfly*, *scorpion*, *spider*, *ant*, *bee*, *ladybug*, *stick*, *bug*, *beetle*, *cockroach*, and *fly*. Two different and randomly generated orders of the study categories were created; half of the subjects received each study version. SuperLab LT was used to present the pictures as well as administer the distractor task and the recognition tests.

Each test consisted of 100 pictures, 50 of which were previously studied and 50 of which were distractors. Five pictures from each previously studied 10-picture categorical list appeared on the first test; the other five appeared on the second recognition test. The assignment of study items to test was counterbalanced across subjects, and no studied picture appeared on both tests. The distractors for the MP test were comprised of five pictures from each of the original 15-picture categorical lists. The distractors for the LP test were pictures that were semantically unrelated to any of the 10 presented category sub-lists (again, by our assessment). Each half of both tests contained exactly 25 studied pictures and 25 distractors. The test pictures were randomly ordered, and two different versions of each test were created. The pairing of the test and study phase versions was counterbalanced such that both versions of the MP test were paired with both versions of the LP test. Each pair of tests was also grouped with both study phases, creating eight different study phase-test/test combinations.

Procedure

Subjects were tested individually or in small groups of two or three. Prior to the study phase, subjects read instructions on the computer screen informing them that they were about to be presented a long series of pictures, and that they were to try to remember the pictures as best they could. Subjects began the study phase by pressing the ‘Y’ key on the keyboard.

During the study phase, the pictures from a given category were presented in one of the previously randomized orders. Each picture was presented automatically, one at a time, for 4 s, with a 1s ITI between pictures. Between the last picture of a given list and the first picture of the next list, there was a 2s ITI.

At the conclusion of the study phase, instructions were given for the distraction phase. The distraction phase (lasting approximately 5 min) consisted of basic arithmetic problems, after which subjects were given instructions for the first recognition test. Subjects were told that they would be shown a series of pictures, some of which had been previously studied and others of which would be novel. Subjects were instructed to press the ‘Y’ key if they thought that a given picture had been presented earlier and to press the ‘N’ key if they thought the picture was new. There was no time limit. The next test item appeared automatically only after a response was given. After completing the first recognition task, subjects were given a second recognition test. If subjects received the MP distractors test first, then they received the LP distractors test second, and vice versa. Half of the subjects received the MP distractors test first (MP–LP order), while the other half of the subjects received the LP distractors test first (LP–MP order). The instructions for the second test were identical to the instructions for the first test.

Results

Again, the LP test elicited a lower FAR ($M = .003$) than did the MP test ($M = .10$) for both the MP–LP ($t[15] = 4.24$) and LP–MP ($t[15] = 6.09$) conditions. Thirty of 32 subjects had a higher FAR on the MP test; the other two made no false alarms during the entire experiment.

The critical interaction between test order and test type on HR replicated as well, $F(1, 30) = 4.38$. The difference between mean HRs in the MP–LP condition was not reliable, but was close to being so ($t[15] = 1.98$) in the opposite direction as predicted by the ideal decision-making model. Four of 16 subjects had a higher HR on the LP test, two had equal HR, and the remaining 11 subjects had a higher HR on the MP test. In the LP–MP condition, the HR was reliably higher on the LP test ($t[15] = 3.56$); 12 of 16 subjects showed this effect. Of the remaining four, three had equivalent HR on the two tests. As in Experiment 2, the between-subjects comparison of HR on the first test did not yield a significant difference ($t[30] = .10$).

Memory discriminability, as measured by d', did not differ across the two times for either the MP ($t[30] = 0.54$) or the LP ($t[30] = 1.09$) test.

Discussion

In this experiment, we replicated the interaction between test type and test order under conditions that were dramatically different from the first two experiments. First, picture stimuli were used, and these stimuli elicited superior performance to that seen in the first two experiments, especially in the MP condition. Despite the fact that performance was so high, and forgetting minimal, the general shift towards increasing conservative responding across tests replicated. This reveals that such a shift does not reflect a constant criterion with shifting underlying evidence distributions, nor a response to an accurate self-assessment of forgetting across the tests. Other possibilities will be discussed in General Discussion.

In addition, this experiment eliminated the preview period prior to each recognition test, so any difference in criterion placement between conditions were computed and administered "on-the-fly" by subjects, during the test itself. Also, we "replicated" the null effect[6] of plausibility on the between-subjects first-test effect from Experiment 2.

General discussion

These three experiments revealed several important aspects about criterion shifts based on manipulations of distractor plausibility. First, when the categories are sufficiently transparent to the subject, criterion differences can be observed between subjects on a single test (Experiment 1), a finding at odds with the suggestion of a pure "anchor-and-adjust" mechanism for criterion placement. Subjects bring some absolute standards to bear in setting a recognition criterion (cf. Laming, 1984), but the evidence here clearly shows that they modulate that criterion based on their assessment of how the composition of the test will affect discriminability. However, this result did not replicate when plausibility was manipulated through more delicate means, as in Experiments 2 and 3. Other concurrent work in our laboratory on the nature of "Remember" and "Know" judgments in recognition used the semantic plausibility method of Experiment 1, but without a preview period (as in current Experiment 3), and revealed a between-subjects effect of plausibility on criterion placement (Benjamin, 2004). Thus it seems as though the effects we see here are attributable to the nature of the stimuli, rather than the effects of test previewing: Experiment 2 revealed a null effect with preview, and Benjamin (2004) revealed an effect without preview.

Second, subjects made more conservative judgments on the second than the first test, as reflected by the general decrease in endorsements across tests. This even occurred in a paradigm in which forgetting across the tests was minimal (Experiment 3), suggesting that such a shift did not reflect accurate self-assessments of retention. More important, the replication of this finding in Experiment 3 rules out the possibility that our apparent criterion shift is in fact criterion maintenance, coupled with a violation of the assumption that the old item distribution mean remained constant across tests. It is more likely that the general shift reveals a generally accurate belief about the effects of time on memory accuracy, unmodulated by a specific assessment of current performance. This hypothesis makes a strong and testable prediction about shifts of equivalent magnitude across a delay, regardless of the type of stimuli or learning and consequent degree of forgetting.

The focus of this paper was, however, the relationship between test type and criterion shifts across tests, which leads us to the third and most important finding. Regardless of whether the previously discussed between-subjects difference is obtained, there is an asymmetry in criterion-shifting across multiple tests with differentially plausible distractors: when moving from a test with less plausible (i.e., less difficult) distractors to one with more plausible (i.e., more difficult) distractors, subjects alter their decision standards accordingly. However, they do not do so when the first test is more difficult and the second less difficult. This finding was replicated in all three experiments.

This asymmetry is reminiscent of a similar result in the skill learning literature, in which transfer to a difficult same/different shape discrimination task was shown to be superior following training on difficult, rather then easy, discriminations (Doane, Alderton, Sohn, & Pellegino, 1996). Their interpretation of that result was that subjects in the easy discrimination condition failed to optimize an ordering of feature comparisons for the shapes, and thus were at a disadvantage on the difficult transfer items on which a majority of such comparisons were nondiagnostic (see also Fisher & Tanner, 1992). An analogous explanation can be applied to the current results. Subjects who begin by making easy discriminations (on the LP test) can implement a strategy that heavily weights membership in a studied category and correspondingly devalues specific memory for the individual item. When those subjects then confront the difficult test, on which all items yield approximately equivalent evidence for category membership, other evidential

[6] The power to detect an effect of equivalent size as in Experiment 1 was ~0.99, and the power to detect an effect of 1/2 the size was ~0.82. However, both of these estimates are qualified by the very high hit rates, which likely collapsed the functional range of the scale measurably. The power to detect an effect of 0.03 was ~0.35.

bases—such as memory for the item—play a larger role in the decision. Because subjects know their item memory to be more imperfect than their memory for the categories, they then employ a more stringent criterion.

However, when subjects begin the task by making difficult discriminations in the MP condition, they develop a strategy that "transfers" quite well to the easy discriminations in the LP condition. Using individual item memory as a basis for the decision works regardless of the distractor set, and thus there is no impetus for the subject to shift that strategy. It is suboptimal in the sense that P_C could be higher if subjects were to adopt a more liberal strategy, but its continued application does not result in poor performance, unlike in the other group.

Note that this explanation differs somewhat from the canonical criterion-shift explanation, which assumes that the evidence axis is equivalent between conditions, and thus that subjects simply demand more or less of the same type of evidence. Such an account is also plausible here. For example, memory evidence might reflect a summed total of a series of feature matches between the test item and recent memory. These features could include both individuating and categorical semantic knowledge; thus new items that share the latter with members of the previously studied list yield higher evidence scores than distractors that do not share such features. The overall criterion on this evidence axis is raised when the tests become more difficult and the subject detects that a preserved criterion would yield a high FAR. However, when the tests become easier, no "red flag" in performance (i.e., high FAR or low HR) is evident to cue subjects to revise criterion placement.

The former explanation is consistent with the contribution of multiple sources to the recognition decision, all of which are projected onto a single decision axis (Banks, 1999). Such an explanation thus illustrates the separability of single-process recognition theories from the application of signal-detection theory to situations in which multiple contributions to the recognition decision are evident.

Metacognitive monitoring and criterion revision

It might thus be the case that the asymmetric shift is tied to the online evaluation of performance. In the LP–MP case, strategy maintenance would lead to an uncomfortably high FAR (between ~20 and 40% in the experiments reported here). In the MP–LP case, strategy maintenance leads to superior performance—not because of the particularly "wise" decision to preserve criterion placement, but rather by virtue of the greater discriminability afforded by the less plausible distractors. In other words, subjects reevaluate strategies and criterion placement only when the experimental circumstances oblige them into lower performance.

This interpretation highlights the metacognitive nature of criterion placement in recognition, as we have done in other recent work (Benjamin, 2003), and stands in contrast to views of criterion setting that deemphasize a metacognitive role in recognition performance (e.g., Estes & Maddox, 1995). In their model (see also, Estes, 1994), the bias parameter is influenced solely by the a priori probabilities of old and new items on the test, as well as the quality of the match between a test item and memory for the prior study list. Such a model can not handle our central finding that the nature of the distractor material influences criterion placement, because the proportions of items and degree of matching to memory are equivalent between conditions in our experiments. In addition, it is incompatible with our interpretation of the asymmetric criterion shift, which relies on the assumption that subjects' ongoing assessments of performance influence whether or not a shift is seen. Nonetheless, we agree strongly with the general sentiment (cf. Estes, 2002) that criterion placement has too often been an extra degree of freedom in theorizing about recognition, and not well constrained either by empirical regularities or model boundaries. A successful model of recognition performance will necessarily incorporate a mechanism that evaluates and establishes decision criterion placement. Models that do so by assuming a likelihood-ratio decision criterion (e.g., Glanzer et al., 1993; Shiffrin & Steyvers, 1997)—and thus that imply mirror effects as a obligatory consequence of learning manipulations—are ill-equipped to deal with the asymmetry in the criterion shift represented here.

Summary

The evidence here suggests that subjects can set a decision criterion based on information about the makeup of an upcoming recognition test, but only do so when that information and its implications for recognition are quite apparent. However, it appears as though subjects are more able to modulate criterion placement across multiple tests based on assessments of the makeup of the distractor set. They appropriately shifted to a more conservative criterion on a second test when it contained more plausible distractors than a prior test. There was no evidence of criterion change when the second test included less plausible distractors, suggesting that the shift was triggered at least in part by an assessment of their actual performance. These results indicate that criterion-setting in recognition is supported by assessments of discriminability between old and new items, rather than simply the memory strength of studied items or estimates of a priori probabilities of studied and unstudied items on the test. However, neither the assessments nor the consequent criterion revision are obligatory given a change in memory discriminability.

Thus, any comprehensive model of criterion-setting in recognition will necessarily include a metacognitive component that dictates when and how performance is monitored, what aspects of that output influence the decision to adjust criteria, and how and to what extent criteria are revised.

References

Banks, W. B. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*, 81–99.

Banks, W. B. (1999). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*, 267–273.

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

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*, 297–305.

Benjamin, A. S. (2004). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about apparently nonparametric measures. Manuscript under review.

Blackwell, R. H. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America, 53*, 129–160.

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

Doane, S. M., Alderton, D. L., Sohn, Y. W., & Pellegino, J. W. (1996). Acquisition and transfer of skilled performance: Are visual discrimination skills stimulus specific? *Journal of Experimental Psychology: Human Perception and Performance, 17*, 781–791.

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. sequential lineups: What do we really know?* Retrieved June 3, 2003 from http://psy.ucsd.edu/~eebbesen/SimSeq.htm.

Egan, J. P. (1958). Recognition memory and the operating characteristic. *Unites States Air Force Operational Applications Laboratory Technical Note No. 58–51.*

Estes, W. K. (1994). *Classification and cognition.* Oxford: Oxford University Press.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review, 9*, 3–25.

Estes, W. K., & Maddox, W. T. (1995). Interactions of similarity, base rate, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1075–1095.

Fisher, D. L., & Tanner, N. S. (1992). Optimal symbol set selection: A semiautomated procedure. *Human Factors, 34*, 79–95.

Ghetti, S. (2003). Memory for nonoccurrences: The role of metacognition. *Journal of Memory and Language, 48*, 722–739.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546–567.

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 62*, 532–533.

Green, D.M., & Swets, J.A. (1965). Signal detection by human observers. *United States Air Force Electronic Systems Division Technical Documentary Report No. 64–174.*

Gronlund, S.D. (in press). Sequential lineups: Shift in criterion or decision strategy? *Journal of Applied Psychology*.

Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task? *Memory & Cognition, 3*, 233–238.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6*, 554–563.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528–551.

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 201–205.

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

Hockley, W. E., & Niewiadomski, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition, 29*, 1130–1138.

Laming, D. (1984). The relativity of 'absolute' judgments. *British Journal of Mathematical and Statistical Psychology, 37*, 152–183.

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

McCormack, P. D., & Swenson, A. L. (1972). Recognition memory for common and rare words. *Journal of Experimental Psychology, 95*, 72–77.

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

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

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

Parducci, A. (1984). Perceptual and judgmental relativity. In V. Sarris & A. Parducci (Eds.), *Perspectives in psychological experimentation* (pp. 135–149). Hillsdale, NJ: Erlbaum.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, 4*, 171–212.

Ratcliff, R., & McKoon, G. (2000). Memory models. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 571–581). Oxford: Oxford University Press.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518–535.

Robinson, K. J., & Roediger, H. L. (1998). Associative processes in false recall and false recognition. *Psychological Science, 8*, 231–237.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.

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

Stadler, M. A., Roediger, H. L., III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*, 494–500.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology A, 25*, 207–222.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33*, 203–217.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1379–1396.

Thomas, E. A. C., & Legge, D. (1970). Probability matching as a basis for detection and recognition decisions. *Psychological Review, 77*, 65–72.

Wallace, W. P. (1980). On the use of distractors for testing recognition memory. *Psychological Bulletin, 88*, 696–704.

Wallace, W. P. (1982). Distractor-free tests of memory. *American Journal of Psychology, 95*, 421–440.

Wallace, W. P., Sawyer, T. J., & Robertson, L. C. (1978). Distractors in recall, distractor-free recognition, and the word-frequency effect. *American Journal of Psychology, 91*, 295–304.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive interference as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 442–452.

Westerman, D. L. (2000). Recollection-based recognition eliminates the revelation effect in recognition memory. *Memory & Cognition, 28*, 167–175.

Wickens, T. D. (2001). *Elementary signal detection theory*. New York, NY: Oxford.

Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old–new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1022–1041.