Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 4, 616–630

0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.4.616

List Composition and the Word-Frequency Effect for Recognition Memory

Kenneth J. Malmberg
University of Maryland and Indiana University Bloomington

Kevin Murnane
University of Maryland

The attention/likelihood theory (ALT; M. Glanzer & J. K. Adams, 1990) and the retrieving effectively from memory (REM) theory (R. M. Shiffrin & M. Steyvers, 1997) make different predictions concerning the effect of list composition on word recognition. The predictions were empirically tested for two-alternative forced-choice, yes–no, and ratings recognition tasks. In the current article, the authors found that discrimination of low-frequency words increased as the proportion of high-frequency words studied increased. The results disconfirm the ALT prediction that recognition is insensitive to list composition, and they disconfirm the predictions of the REM model described by R. M. Shiffrin and M. Steyvers (1997). The current authors discuss a slightly modified version of REM that can better predict our findings, and we discuss the challenges the present findings pose for ALT and REM.

For recognition memory, "mirror effects" are common (Glanzer & Adams, 1985): If *A* is a better-recognized stimulus class than *B*, then *A* items are more likely than *B* items to be recognized when studied (e.g., a *hit*) and less likely than *B* items to be recognized when not studied (e.g., a *false alarm*). Explaining mirror effects has become a central goal for memory researchers because they were difficult or impossible to explain by many older theories of recognition (Glanzer & Adams, 1985, 1990), and several more recent theories of memory predict mirror effects (Dennis & Humphreys, 2001; Estes, 1994; Glanzer & Adams, 1990; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). The word-frequency effect (WFE) is a mirror effect; low-frequency (LF) words are more likely than high-frequency (HF) words to be recognized when they had been studied and less likely to be recognized when they had not been studied (on average, Schulman, 1967; Shepard, 1967; but see Wixted, 1992). Because the WFE is one of the most robust mirror effects (Glanzer & Adams, 1985), predicting it has become an important goal of most recent theories of recognition memory (Dennis & Humphreys, 2001; Estes, 1994; Glanzer & Adams, 1990; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).

In this article, we directly address the question of how the proportion of HF words studied (vs. LF words), referred to here as a *list-composition* manipulation, affects recognition.[1] It is well documented that changes in list composition have no qualitative effect on the WFE: LF words are better recognized than HF words for both between- and within-list manipulations of word frequency (e.g., Gorman, 1961; Schulman, 1967; Shepard, 1967). However, it is also important to know how mirror effects are quantitatively affected by different factors (Hintzman, Caulton, & Curran, 1994). For example, a theory can predict a change in discrimination for HF and/or LF words that preserves the WFE, a WFE can be observed, and the theory can be disconfirmed because patterns of "old" responses do not conform to its predictions.

Thus, the prior findings showing no qualitative change in the WFE as a function of list composition only minimally constrain theory. In one experiment, for example, Dorfman and Glanzer (1988; also see Clark & Burchett, 1994) asked subjects to make lexical decisions to mixed lists of HF and LF words that varied in their composition. Later, a surprise yes–no recognition test was given for the words appearing during the lexical decision trials. Dorfman and Glanzer found that LF words were recognized better than HF words regardless of list composition. Given this finding, one might be tempted to conclude that list composition does not affect the WFE for recognition. However, Dorfman and Glanzer also found that WFE increased as the proportion of HF words studied increased. Thus, the WFE does not appear to be qualitatively affected by list composition, but it may be quantitatively affected.

Despite its importance to understanding recognition, little attention has been paid to how list composition quantitatively affects the WFE. Directly addressing this question is now timely because two theories of recognition memory, the retrieving effectively from memory theory (REM; Shiffrin & Steyvers, 1997) and the

Kenneth J. Malmberg, Department of Psychology, University of Maryland, and Department of Psychology, Indiana University Bloomington; Kevin Murnane, Department of Psychology, University of Maryland.

This study was supported in part by National Institute of Mental Health National Research Service Award Postdoctoral Fellowship MH12643 awarded to Kenneth J. Malmberg. We thank Rich Shiffrin for helpful conversations, and we thank the reviewers of a prior version of this article for insightful comments.

Correspondence concerning this article should be addressed to Kenneth J. Malmberg, Department of Psychology, Indiana University, Bloomington, Indiana 47405. E-mail: malmberg@indiana.edu or kmurnane@umd5.umd.edu

[1] We use the term "list composition" in this article to refer to a manipulation of the percentage of HF words studied versus the percentage of LF words studied for lists of constant length. It is much more compact than a manipulation of "the percentage of HF words studied versus the percentage of studied LF words," although we sometimes use the more elaborate description when needed for clarity. There are many types of list-composition manipulations, however, and we trust that our use of the general term in this article will not be confused with other types of list-composition manipulations (e.g., list strength, semantic associates).

attention/likelihood theory (ALT; Glanzer & Adams, 1990; Glanzer, Adams, Iverson, & Kim, 1993) make different qualitative predictions about the outcome: REM predicts a list-composition effect, but ALT does not. To understand these predictions, however, it is necessary to describe the ALT and REM accounts of the WFE in detail, and this will also allow us compute their predictions. After doing so, we present three experiments. Experiment 1 uses a two-alternative forced-choice (2AFC) recognition procedure, which allows for a relatively pure view of how list composition affects the ability to discriminate studied and unstudied words. Experiments 2A and 2B use yes–no and confidence ratings procedures in an attempt to generalize the findings from Experiment 1 to these common recognition procedures.

The Mirror-Patterned WFE for 2AFC Recognition

For 2AFC recognition, two items are presented and the subject's task is to determine which item was studied. If $P(x, y)$ is the probability of choosing word x over word y and words are HF or LF, then there are six independent types of comparisons that can be made P(HF-old, HF-new), P(HF-old, LF-new), P(LF-old, HF-new), P(LF-old, LF-new), P(LF-old, HF-old), and P(HF-new, LF-new). For standard comparisons, a target (an old item) and a foil (a new item) are presented, and the following is a mirror-patterned WFE for the four standard comparisons:

$$P(\text{HF-old, HF-new}) < P(\text{HF-old, LF-new}), \quad \text{(R1)}$$

$$P(\text{LF-old, HF-new}) < P(\text{LF-old, LF-new}),$$

$$P(\text{HF-old, HF-new}) < P(\text{LF-old, HF-new}), \text{ and}$$

$$P(\text{HF-old, LF-new}) < P(\text{LF-old, LF-new}).$$

On null-comparison test trials, two foils or two targets from different stimulus classes are presented, and this is a mirror-patterned WFE for the two null comparisons:

$$P(\text{LF-old, HF-old}) \text{ and } P(\text{HF-new, LF-new}) > .50. \quad \text{(R2)}$$

Thus, old words are chosen more often than new words, HF-new words are chosen more often than LF-new words, and LF-old words are chosen more often than HF-old words. Taken together, R1 and R2 comprise the mirror-patterned WFE for 2AFC recognition.

Assume there is a value of a random variable associated with each test item. If a 2AFC is based on comparing these values such that the item associated with the greatest value is chosen, then R1 and R2 indicate that the mean values for the random variable associated with new and old HF and LF words conform to the following ordering:

$$\mu(\text{LF-new}) < \mu(\text{HF-new}) < \mu(\text{HF-old}) < \mu(\text{LF-old}). \quad \text{(R3)}$$

For example, if the recognition decision is based on the levels of "familiarity" of the test items, then LF-new words are the least familiar and LF-old words are the most "familiar" (on average).

The REM Account of the WFE (Shiffrin & Steyvers, 1997)

REM predicts that LF targets are *more* likely than HF targets to be correctly recognized because LF targets match or activate to a greater degree than HF targets their own episodic memory traces. That is, LF targets are more familiar than HF targets on average. REM predicts that LF foils are less likely than HF foils to be incorrectly recognized because HF foils tend to spuriously match traces of other words to a greater degree than LF foils. That is, there is more "noise" in the output from memory when an HF foil is tested versus when an LF foil is tested. Therefore, HF foils are more likely than LF foils to be incorrectly called "old." Together these assumptions satisfy R3. Although these assumptions are straightforward, they can only properly be understood within the context of the REM framework.

Specifically, REM assumes that generic knowledge is stored in *lexical/semantic* images, and every known word has a corresponding lexical/semantic image consisting of item (w_i) and context (w_c) features. Lexical/semantic features represent the orthographic, phonemic, and semantic properties of a word. Features are integers that vary in their frequencies of occurrence. The features comprising lexical/semantic images are determined by drawing integers randomly from a geometric distribution. This occurs independently for each feature. Thus, the probability that feature value, j, is encountered in a given feature location of a lexical/semantic image is:

$$P(\text{V} = j) = (1 - g)^{j-1} g, \quad (1)$$

where $j = 1, 2, 3$, and so forth.

The g parameter determines the mean and variability of feature values. Figure 1 shows the geometric distributions for two different values of g, and it shows that the features will tend to be integers with relatively small values and with relatively little variability when sampled from a geometric distribution defined by a relatively high g value. Conversely, when g is relatively low, the features that represent a word are more varied and have a greater mean value. For this reason, words constructed by randomly sampling from a geometric distribution defined by a relatively high g value will tend to be more similar to each other than words constructed by sampling from a geometric distribution defined by a relatively low g value. As we will see below, this property is used to account for the mirror-patterned WFE: It is assumed that HF words share more common features than LF words, which is implemented in the Shiffrin and Steyvers (1997) model, by assuming that $g_{HF} > g_{LF}$, and hence the lexical/semantic images of different HF words are more similar than those of different LF words (on average).

When a word is studied, an *episodic* image (or trace) of the event is stored in memory, and different study events are stored in different images. Each episodic image is a vector of features, and hence the episodic representation of an n-item study list consists of n vectors.

When a word is studied, its lexical/semantic image is retrieved from memory. The episodic encoding of a word is a relatively incomplete and inaccurate copy of the word's lexical/semantic image (as well as context). Thus, some of the item features in a word's lexical/semantic image are stored in the episodic image of the study event: After t time units of study, the probability that a lexical/semantic feature will be stored in an episodic image is $1 - (1 - u^*)^t$, otherwise 0 is stored (u^* is the probability of storing a feature in a unit of time). Features are correctly encoded with probability c. When incorrectly encoded, features are drawn randomly according to Equation 1 and stored.

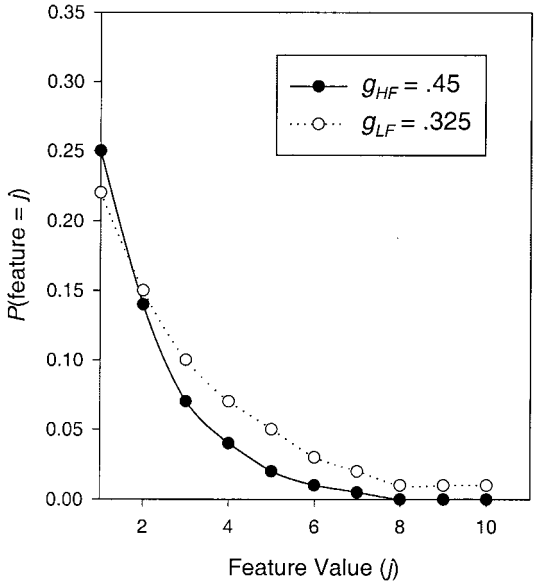

Figure 1. Probability of sampling feature j for geometric distributions defined by different base-rate parameter values (g). g_{HF} and g_{LF} refer to the values of the base-rate parameter for high-frequency and low-frequency words, respectively. When sampled from a geometric distribution defined by a relatively high g value, the features that represent a word will tend to be integers with relatively small values and with relatively little variability. Conversely, when g is relatively low, the features that represent a word are more varied and have a greater mean value. On the assumption that $g_{HF} > g_{LF}$, HF words are more similar to each other than LF words. It is this assumption that underlies the retrieving effectively from memory theory account of the word-frequency effect for recognition memory.

At test, the lexical/semantic vector of w item features corresponding to the test item serves as a retrieval cue. We assume that the cue is a perfect copy of the lexical/semantic representation of the test item (although this need not necessarily be the case, Shiffrin & Steyvers, 1997).[2] Memory is probed by comparing (or matching) the retrieval cue in parallel against the n images (I_j) in memory, and the match consists of noting which features of I_j match the corresponding features of the cue. Next, a likelihood ratio, λ_j, is computed for each I_j:

$$\lambda_j = (1 - c)^{n_{jq}} \prod_{i=1}^{\infty} \left[\frac{c + (1 - c)\, g(1 - g)^{i-1}}{g(1 - g)^{i-1}} \right]^{n_{ijm}}, \qquad (2)$$

where g is the long-run environmental base rate for the occurrence of features, λ_j is the activation of I_j in response-probing memory with the retrieval cue, and λ_j is positively related to the number of matching features and the *diagnosticity* of their values. That is, matching rare features (i.e., relatively high integer values) contributes more activation than matching common features (i.e., relatively low integer values) because rare features are less likely to match by chance, according to the Bayesian calculation (Equation 2). As we previously noted, LF words tend to consist of more unusual features than HF words in REM. Thus, LF features are said to provide relatively diagnostic information about whether an image in memory corresponds to the retrieval cue, and LF retrieval cues tend to produce higher λ_js than HF retrieval cues when a cue and an image correspond to the same word.

The recognition decision is based on the odds, Φ, the probability that the test item is old divided by the probability that the test item is new (Shiffrin & Steyvers, 1997):

$$\Phi = \frac{1}{n} \sum_{j=1}^{n} \lambda_j = \frac{1}{n} \left\{ \sum_{j=1}^{k} \lambda j_{LF} + \sum_{j=k+1}^{n} \lambda j_{HF} \right\}, \qquad (3)$$

where n is the number of items studied, k is the number of LF words studied, $n - k$ is the number of HF words studied, and λ_{jHF} and λ_{jLF} are the activations associated j^{th} HF and LF images, respectively.[3] For REM, we assume 2AFC recognition is performed by separately computing the odds for both test items, and the item producing the greatest odds is chosen (Equation 3).

We now have all the information necessary to understand why REM predicts a mirror-patterned WFE: LF words consist of more uncommon features than HF words (i.e., $g_{HF} > g_{LF}$). Therefore, a much greater λ_j is usually produced when an LF retrieval cue is matched against a similar trace in memory—like the one that was stored if the test word was actually studied—than when an HF retrieval cue is used to probe memory. This produces the advantage for LF-old words in standard-comparison test trials (R1) and in old-item null-comparison test trials (R2).

For choices involving a new item, the mirror-patterned WFE is predicted because the common features that make up HF retrieval cues tend to match the images of other words better than the uncommon features that make up LF retrieval cues. The additional spurious matches for HF foils produce greater λ_js, which are summed to produce the odds. It is easy to see from Equation 3 that HF foils tend to be called "old" more often because the average nontarget λ_j is greater for HF words than for LF words. This produces the advantage for LF-new words in standard-comparison test trials (R1) and in the new-item null-comparison test trials (R2).

The Effect of List Composition on 2AFC Recognition in REM

The question we ask is what effect does a change in the proportion of HF words studied have on recognition? The left panels of Figure 2 show that REM (Shiffrin & Steyvers, 1997) predicts a list-composition effect: The probability of choosing HF words, especially foils, is predicted to increase as the percentage of HF words studied increases. That is, P(HF-new, LF-new) and P(HF-old, LF-new) increase and P(LF-old, HF-new) and P(HF-old, HF-new) decrease as the proportion of HF words studied increases.

The same mechanism that produces the list-composition effect produces the LF advantage for foils: access to memory is global (Equation 3; for a review of global matching models, see Clark & Gronlund, 1996), and HF and LF words are represented by features

[2] Shiffrin and Steyvers (1997) discuss the necessity for context to be stored and used in the retrieval cue at test. In our modeling, however, these assumptions were not necessary because we assumed no extralist images were stored for the sake of simplicity.

[3] Technically, this is incorrect; n refers to the number of items in a set activated above a threshold by a context cue. In this model, we assume all, and only, list images are in the activated set.

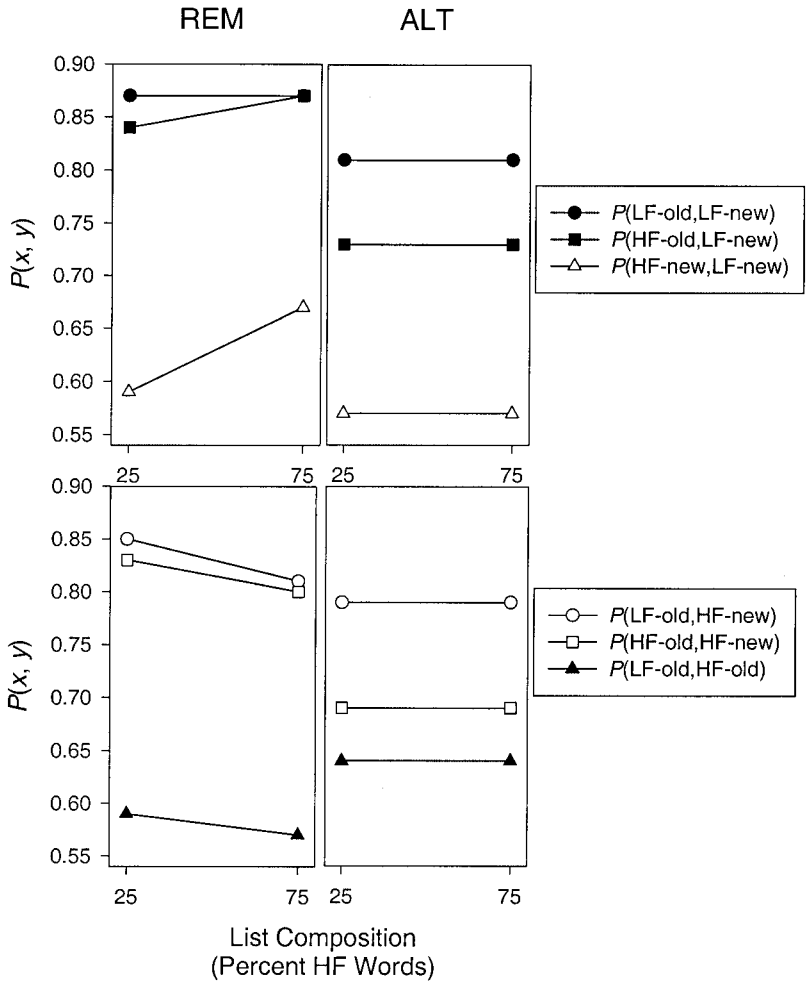

Figure 2. Retrieving effectively from memory (REM) and attention/likelihood theory (ALT) predictions for two-alternative forced-choice (2AFC) recognition as a function of list composition. Monte Carlo simulations were used to generate REM predictions (Shiffrin & Steyvers, 1997). For each simulation, 1,000 simulated subjects were run with the following parameter values: $w = 20$, $g_{HF} = .45$, $g_{LF} = .325$, $g = .4$, $t = 10$, $c = .7$, $u^* = .04$. For 2AFC recognition, six different types of pairs were tested (LF-old/LF-new, LF-old/HF-new, HF-old/LF-new, HF-old/HF-new, HF-new/LF-new, and LF-old/HF-old). The item from each pair that elicited the greatest odds was chosen to be the recognized item. The ALT predictions are computed from $p(\text{new}) = .10$, $N = 1{,}000$, $n(\text{LF}) = 60$, $n(\text{HF}) = 40$, $p(\text{LF, old}) = .154$, $p(\text{HF, old}) = .136$, $p_{25}(.,\text{old}) = .1495$, and $p_{75}(.,\text{old}) = .1405$. The results of many simulations using other sensible parameter values were very similar to those presented here for both models. $P(x, y)$ = mean probability of choosing x over y; LF = low frequency; HF = high frequency.

drawn from different geometric distributions (Equation 1). HF members of the global set of nontarget images produce relatively high activations (Equation 2) in response to HF retrieval cues because different HF words tend to share features. It is easy to see from Equation 3 that adding more HF images to memory will tend to increase the odds for HF words. Therefore, the means of the HF old- and new-item odds distributions are positively related to the percentage of HF items studied (Shiffrin & Steyvers, 1997), which means that the familiarity of HF words increases relative to the familiarity of LF words as the number of HF images in the global set increases. The top-left panel of Figure 2 also shows that increasing the proportion of HF words studied is expected to have relatively little effect on the familiarity of LF words—that is, *P*(LF-old, LF-new). This is simply because LF words tend to consist of relatively uncommon features, and therefore produce relatively few spurious matches regardless of the list composition.

The ALT Account of the WFE

According to the ALT (Glanzer & Adams, 1990; Glanzer et al., 1993), a trace is a set of features, some features are "marked" when a word is studied, and marking is positively related to the amount of attention an item attracts at study. ALT predicts a WFE because LF words are assumed to attract more attention than HF words, and because LF targets are expected to have more marked features than HF targets.

Specifically, memory representations in ALT consist of N features. A proportion of the features in each trace, p(new), are already marked prior to study. That is, p(new) indexes the amount of "noise" in a trace. When studied, a subset, $n(i)$, of a word's features are randomly sampled from its memory trace, where i indexes experimental conditions (e.g., HF vs. LF), and they are marked if they were not already. The proportion of features sampled is $\alpha(i) = n(i)/N$, and therefore the proportion of marked features after an item has been studied is:

$$p(i, \text{old}) = p(\text{new}) + \alpha(i) \cdot [1 - p(\text{new})], \tag{4}$$

where $\alpha(i)$ reflects the amount of attention devoted to studying a word. Hence, from Equation 4 it is easy to see that the proportion of marked features after study is positively related to $\alpha(i)$.

At test, $n(i)$ features are randomly sampled from the test item's trace, and x of the features are found to be marked. The recognition decision is based on the log-likelihood ratio associated with the number of marked features sampled for items of type i (LF vs. HF) and type j (old vs. new):

$$\ln L(x|i, j) = n(i) \cdot \ln\left[\frac{q(i, \text{old})}{q(\text{new})}\right] + x \cdot \ln\left[\frac{p(i, \text{old})\ q(\text{new})}{p(\text{new})\ q(i, \text{old})}\right], \tag{5}$$

where $q(i, j) = 1 - p(i, j)$. The $\ln L(x)$s are distributed according to the binomial distribution described by $n(i)$ and $p(i, j)$:

$$p(x|i, j) = \left(\frac{n(i)}{x}\right) \cdot p(i, j)^x \cdot q(i, j)^{n(i)-x}, \tag{6}$$

$p(x \mid i, j)$ should not be confused with $P(x, y)$ used in R1–R3. $p(x \mid i, j)$ is the probability of randomly sampling x-marked features from a memory trace given the test item is i (e.g., HF vs. LF) and j(old vs. new). $P(x, y)$, however, is the probability of choosing item x over item y in a 2AFC recognition task.

In ALT for 2AFC recognition, a random sample of $n(i)$ features is taken from the test item's trace (x will be marked), the log likelihoods are computed using Equation 5, and the item with the greater log likelihood is chosen. ALT predictions for the different $P(x, y)$s can be computed using Equation 6.

A fundamental property of ALT is that changes in $n(i)$, $p(i, \text{old})$, and p(new) produce opposite changes in the tendencies to call targets and foils "old." Hilford, Glanzer, and Kim (1997) state: "Any manipulation that affects the recognition of old items will also affect the recognition of new items" (p. 593). That is, any manipulation that increases the tendency to call a class of targets "old" decreases the tendency to call foils from that class "old." This property of ALT is sometimes called *concentering* because it reflects a convergence on the decision axis of the means of underlying old and new $\ln L(x)$ distributions.

In ALT, differences in the amount of attentional resources given to studying different classes of stimuli account for mirror effects in general and the WFE specifically: ALT assumes that LF words attract more attention than HF words. According to the concentering principle, the distributions of the classes of stimuli that attract relatively high amounts of attention are more spread out on the decision axis—and therefore overlap less—relative to those of classes of stimuli that attract less attention (e.g., see Figure 3). Thus, the LF distributions overlap less than the HF distributions, and LF words are therefore better recognized than HF words.

At test, the proportion of marked features that are sampled is greater, on average, for LF targets than for HF targets because more features are sampled for LF words and because LF words have more marked features (on average). Therefore, the mean of the LF target distribution is greater than the mean of the HF target distribution (R3). This produces the advantage for LF words on standard comparisons (R1) and on null-comparison test trials involving old words (R2).

For choices involving a new item, the mirror-patterned WFE is predicted because LF words are expected to have a greater number of marked features if they were studied than HF words. It should be noted that the same proportion of marked features are contained in the traces of LF and HF foils—p(new). Therefore, the mean of the HF foil distribution is greater than the mean of the LF foil distribution over the same range of possible $\ln L(x)$ values (R3), only because of the system's expectation that more marked features will be sampled from LF traces. Thus, more marked features need to be sampled from LF foil traces than HF foil traces to generate an "old" response. This produces the advantage for LF foils in standard-comparison test trials (R1) and in the new-item null-comparison test trials (R2).

The Effect of List Composition on 2AFC Recognition in ALT

The most basic version of ALT predicts no list-composition effect because only information about the test item enters into the computation of likelihood ratios (Equation 5). That is, ALT is a local-access theory because it assumes that only the contents of single trace contributes to the recognition decision. Alternatively, Glanzer et al. (1993) suggested that the average expectation of the number of marked features, p(.,old), may be used to compute the likelihood ratios for all items. On this assumption, access to memory is still locally accessed, but metalevel inferences based on the compositions of the study lists are also used in likelihood computations. If so, p(.,old) will vary with list composition. However, the right-hand panels of Figure 2 show that ALT predicts that changes in p(.,old) have no effect on 2AFC recognition.

Glanzer et al. (1993) explained the predicted null effect of list composition this way: "The reason for the stability in the predicted data values, despite changes in the log likelihood means, is that changes in means are accompanied by changes in variance" (p. 564). To see why, consider that in ALT, recognition performance is a positive function of the difference between the means of two $\ln L(x)$ distributions, one corresponding to each word in a 2AFC recognition task. Figure 3 shows the $\ln L(x)$ distributions corresponding to a WFE for two levels of list composition (25% and 75% HF). These distributions were derived on the assumption that the average $p(i, \text{old})$ was used to compute the likelihood ratios for all types of items. Thus, for p(new) = .10, N = 1,000, n(LF) = 60, and n(HF) = 40, Equation 4 produces p(LF, old) = .154 and p(HF, old) = .136. If p(.,old) is the average $p(i, \text{old})$, then p_{25}(.,old) = .1495 and p_{75}(.,old) = .1405 . p_{25}(.,old) > p_{75}(.,old) because the system expects more HF words on the 75% list than on the 25% list.

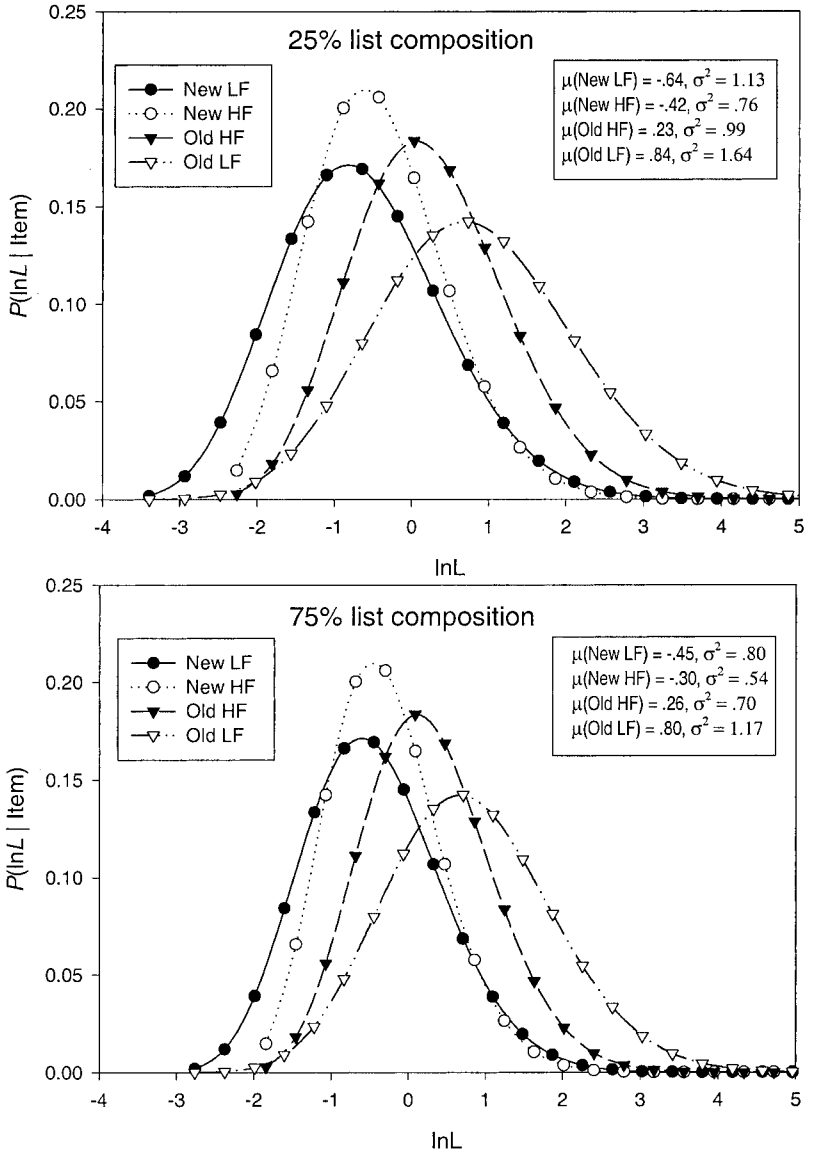

Figure 3. Attention/likelihood theory performance for two-alternative forced-choice recognition. LF = low frequency; HF = high frequency.

Figure 3 shows that as the memory system expects more HF words—that is, $p(.,\text{old})$ decreases—the distance from 0 of all the distributions decreases: The means of the target distributions become less positive, and the means of the foil distributions become less negative. Thus, the differences between the two old-item and the two new-item distributions decrease. The change in $p(.,\text{old})$, however, does not affect recognition performance because performance is also a negative function of the variance of the two $\ln L(x)$ density distributions. That is, two highly variable distributions overlap to a greater extent than two distributions that are less variable. Figure 3 shows that in ALT, a decrease in $p(.,\text{old})$ decreases the variance of all the distributions, and therefore its effects on the means and the variances trade off.

The mean and variance of the $\ln L(x)$ density distributions are given by the following equations, respectively:

$$E\ \ln L(x|i,j) = n(i)\ p(i,j)\ \ln\left[\frac{p(i,\text{old})}{p(\text{new})}\right] + n(i)\ q(i,j)\ \ln\left[\frac{q(i,\text{old})}{q(\text{new})}\right] \text{ and} \qquad (7)$$

$$Var\ \ln L(x|i,j) = n(i)\ p(i,j)\ q(i,j)\left\{\ln\left[\frac{p(i,\text{old})\ q(\text{new})}{p(\text{new})\ q(i,\text{old})}\right]\right\}^2. \qquad (8)$$

Equation 8 shows that a decrease in the $p(.,\text{old})$ results in a decrease in the variance of all the distributions simply because the same $p(.,\text{old})$ is used to compute the variance of all of them, and as $p(.,\text{old})$ decreases so does

$$\left\{\ln\left[\frac{p(i,\text{ old})q(\text{new})}{p(\text{new})q(i,\text{ old})}\right]\right\}^2. \quad (9)$$

The effect of

$$\frac{p(i,\text{ old})}{p(\text{new})}$$

in Equation 7 is to position the means of the distributions on the decision scale relative to 0. Thus, decreasing

$$\frac{p(i,\text{ old})}{p(\text{new})} \quad (10)$$

decreases the distances of means of all the underlying $\ln L(x)$ distributions from 0, converging them on the decision scale. Figure 3 shows that the ordering of the means remains unchanged with changes in $p(.,\text{old})$. The order of the means is preserved because

$$\frac{p(i,\text{ old})}{p(\text{new})}$$

are the same for both HF and LF words. That is, the effect of increasing $p(.,\text{old})$ in Equation 5 is to map values of x for LF and HF words onto a single different $\ln L(x)$ scale, which preserves the order of the $\ln L(x)$s (a proof is offered in the Appendix). In combination, the convergence of the means and the decrease in the variances of the $\ln L(x)$ distributions trade off, and 2AFC recognition performance is not predicted to vary as a function of list composition.

Experiment 1: Forced-Choice Recognition

In the prior section, we showed that ALT and REM 2AFC models make different predictions concerning the effect on the WFE of studying different proportions of HF and LF words. ALT predicts no effect, and REM predicts that the probability of choosing HF words will increase as the proportion of HF words studied increases. In this experiment, we directly test these predictions by having subjects study lists that consist of either 25% HF words and 75% LF words or 75% HF words and 25% LF words.

Method

Subjects. Forty-eight students at the University of Maryland participated in the experiment in exchange for course credit.

Design and materials. Word frequency (HF vs. LF) and list composition (25% vs. 75% HF words) were manipulated within subjects. The dependent measures were the percent correct for the standard comparisons and the probabilities of selecting the HF foil and the LF target for the new- and old-item null comparisons, respectively. Twenty-four subjects were randomly assigned to each list-order condition.

LF versus HF was operationally defined as words appearing between 1 and 10 times versus greater than 50 times per million (Kučera & Francis, 1983). Word length and concreteness are sometimes controlled when word frequency is manipulated. However, this control is not crucial, as the WFE is not qualitatively affected by the implementation of these controls (Hintzman et al., 1994; Schulman, 1967).

Two 96-item study lists were randomly constructed for each subject. Each word on the study list was presented in the center of a computer monitor for 2 s of study followed by a 150-ms interstimulus interval. The study list was followed by a 30-s math task and 96 2AFC test trials ordered randomly. Fifty-six LF-old/LF-new and 56 HF-old/HF-new pairs were tested for the 25% and the 75% lists, respectively, and 8 pairs for each of the remaining five choices were tested. List order was counterbalanced.

Procedure. On each 2AFC trial, subjects were shown two words, one above the other on a computer screen. The old item appeared above the new item on half of the standard test trials, and the order was determined randomly for the null comparisons. The *1* and *2* keys were used to indicate that the top and bottom words were studied, respectively.

Results

An alpha level of .05 is the standard of significance; *t* tests are two-tailed. List order did not significantly affect any choice and did not interact with comparison type. Therefore, the data were collapsed across list order for the remaining analyses. Figure 4 plots $P(x, y)$, the probability of choosing item x over item y. The left panel of Figure 4 plots $P(x, y)$ for those choices involving an LF foil, and the right panel plots $P(x, y)$ for the remaining choices. Comparing across panels, P(LF-old, LF-new) was greater than P(LF-old item, HF-new), $t(47) = 2.23$, $SEM = .02$. P(HF-old, LF-new) was greater than P(HF-old item, HF-new), $t(47) = 5.29$, $SEM = .02$. The old and new null comparisons were significantly greater than .50, $t(47) = 7.86$ and $t(47) = 13.03$, respectively.

List composition. The left panel of Figure 4 shows that the probability of choosing an LF foil decreased significantly as the proportion of HF words studied increased, $F(1, 47) = 4.66$, $MSE = 0.01$. The probability of choosing an LF-new word decreased as the percentage of HF words studied increased for the HF-old/LF-new choice, $F(1, 47) = 5.37$, $MSE = 0.10$, and the HF-new/LF-new choice, $F(1, 47) = 20.38$, $MSE = 0.51$. The right panel of Figure 4 shows that list composition did not significantly affect any of the other choices subjects made.

Discussion

Overall, LF foils were rejected more often when the study list consisted of mostly HF words. ALT and REM do not predict this result. For REM, differences between its predictions and our findings are observed by comparing the left panels of Figure 2 with Figure 4. Figure 4 shows that P(HF-new, LF-new), P(HF-old, LF-new), and P(LF-old, LF-new) increase as the percentage of HF words studied increases, but the left panels of Figure 1 show that REM predicts that P(HF-new, LF-new) and P(HF-old, LF-new) should increase, and P(HF-old, HF-new), P(LF-old, HF-new), and P(LF-old, HF-old) should decrease. The increase in P(HF-new, LF-new) was both predicted by REM and observed. Theoretically, this could mean one of two things: either HF foils increased in familiarity or LF foils decreased in familiarity as the percentage of HF words studied increased. However, if HF foils became more familiar—as REM predicts—then P(HF-old, HF-new) and P(LF-old, HF-new) should decrease, but neither were significantly affected by list composition. Thus, REM predicted an increase in the probability of choosing HF words, but we observed a decrease in the probability of choosing LF foils.

The finding that P(HF-old, LF-new) and P(LF-old, LF-new) increased along with P(HF-new, LF-new) is consistent with the proposition that LF foils became less familiar relative to HF targets, LF targets, and HF foils as the percentage of HF words studied increased. This proposition is, however, inconsistent with the assumptions made by ALT. ALT simply predicts no list-

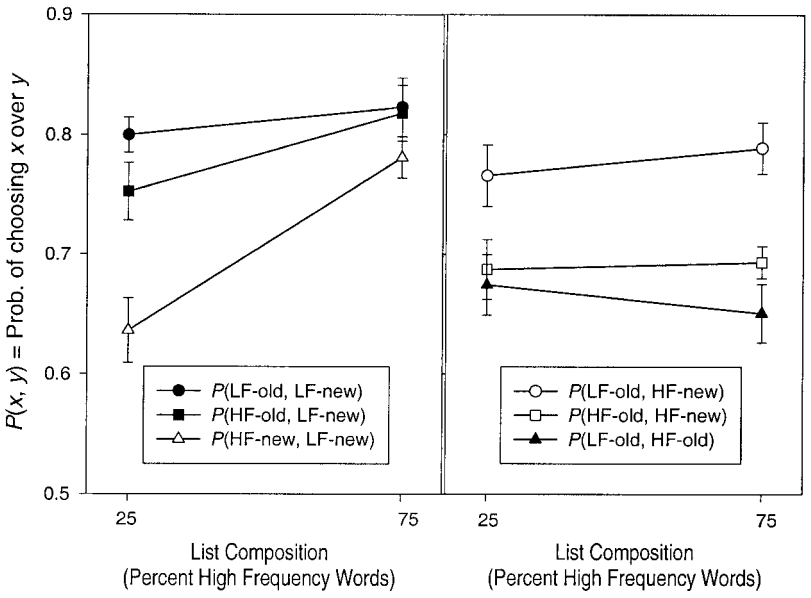

Figure 4. Two-alternative forced-choice performance as a function of pair type and list composition. Error bars indicate standard errors. Prob. = probability; LF = low frequency; HF = high frequency.

composition effect, and one was observed (see Figures 2 and 3). Of importance, the nature of the list-composition effect is also troubling for ALT: Not only does ALT not predict a list-composition effect, but if it did, it would be demonstrated by a change in all $P(x, y)$s involving LF items according to the concentering principle. However, only the $P(x, y)$s involving LF foils were affected by list composition, and this violates the concentering principle.

Experiments 2A and 2B

Neither ALT nor REM fared well in predicting the results of Experiment 1. The results from Experiment 1, which used a 2AFC procedure, provide evidence within the framework of the signal detection theory that LF words are better recognized as the proportion of HF words studied increases because LF foils become relatively less familiar. The goal of Experiment 2 is to generalize the list-composition findings from 2AFC recognition to yes–no and ratings tasks in Experiments 2A and 2B, respectively. The yes–no task requires the subject to respond positively to studied items, and the ratings task requires the subject to assess their confidence that an item was studied. The following ordering is a mirror-patterned WFE for yes–no—P(old)—recognition and mean ratings: LF-new < HF-new < HF-old < LF-old (cf. R3). Both ALT and REM predict the yes–no and ratings WFEs, but they make different predictions concerning the effect of list composition.

In REM, comparing the odds (Equation 3) associated with a test item with a criterion performs yes–no recognition: If the odds exceed the criterion, then an "old" response is made. REM predicts an LF hit-rate (HR) advantage because it assumes that LF words consist of more uncommon features than HF words (i.e., $g_{HF} > g_{LF}$), and matching uncommon features contributes more to λ_j than matching common features (Equation 2). The false-alarm rate (FAR) effect is predicted because the common features that make up HF retrieval cues tend to match the images of other words better than the uncommon features that make up LF retrieval cues. The additional spurious matches produce the LF FAR advantage (as explained above) and also produce a list-composition effect on the means of the HF odds distributions, which are predicted to increase with increases in the percentage of HF words studied.

For yes–no recognition, the issue of a possible shift in the criterion for responding "old" needs to be addressed. A criterion shift between list-composition conditions could produce a list-composition effect for yes–no recognition independently of one that may be caused by a change in sensitivity, especially if differences in the list structures are detected (e.g., Higham, Brooks, & Lee, 1997). Because neither ALT nor REM make predictions about the location of the criterion in such situations, we consider three possible outcomes for the present experiments: As the percentage of HF words studied increases, the criterion is not affected by list composition, the criterion becomes stricter, and the criterion becomes less strict.

The left panels of Figure 5 show the REM predictions for yes–no recognition when different criterion shifts are assumed. The middle-left panel of Figure 5 assumes no criterion shift as a function of list composition. Under this condition, REM predicts that the HF HR and FAR will increase as the percentage of HF words studied increases. The top-left panel of Figure 5 assumes a less strict criterion in the 75% condition: REM predicts that the HR and FAR for both HF and LF words will increase as percentage of HF words studied increases. The bottom-left panel assumes a stricter criterion in the 75% condition: REM predicts that the HR and FAR for LF words will decrease as percentage of HF words studied increases. The exact predictions for HF words, however, depend on the magnitude of the shift. The bottom-left panel of Figure 4 shows the predicted pattern of HRs and FARs if the increase in the criterion exactly offsets the increase in the famil-

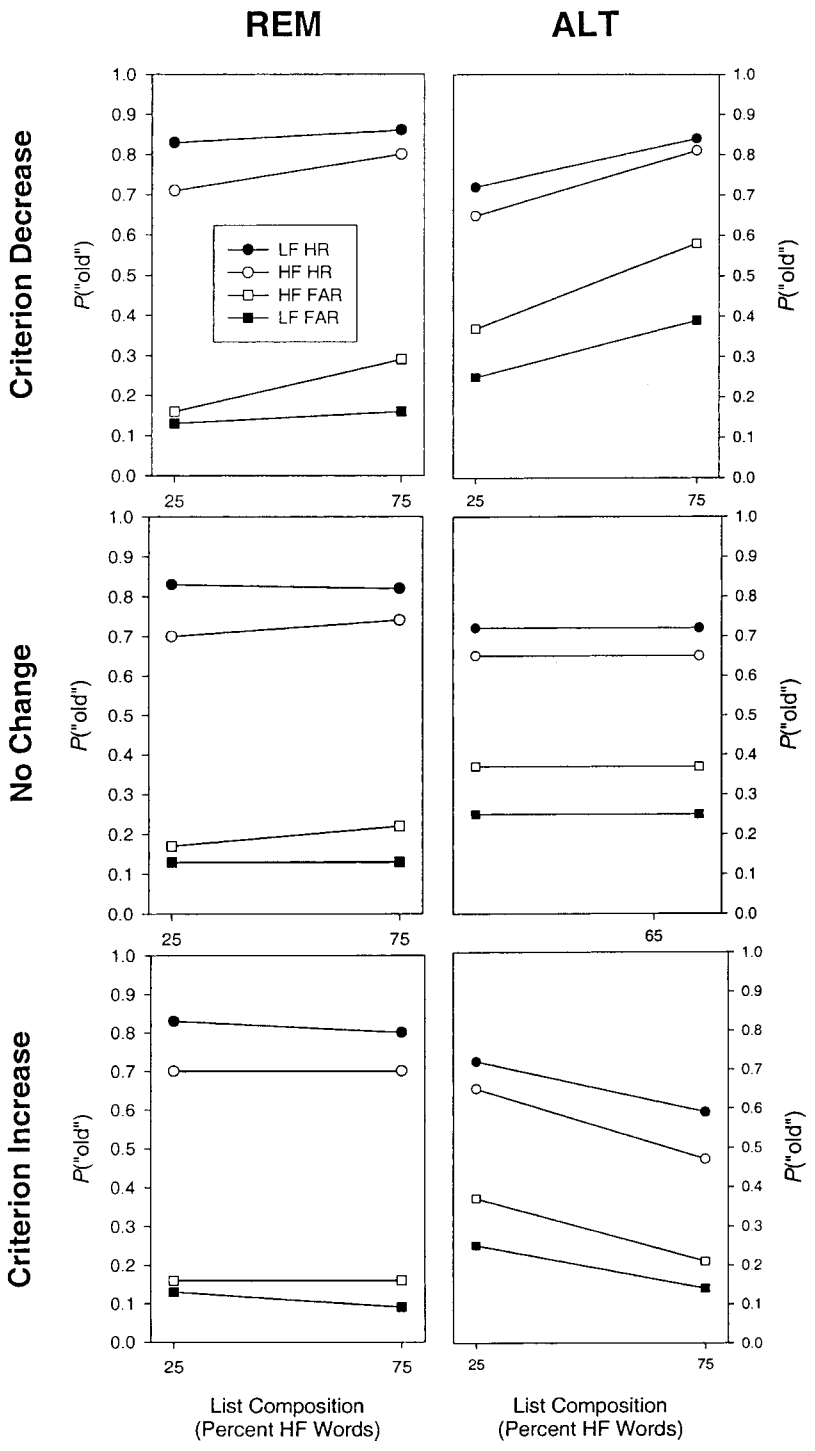

Figure 5. Retrieving effectively from memory (REM) theory and attention/likelihood theory (ALT) predictions for yes–no recognition as a function of list composition. The same parameters were used for the ALT and REM yes–no models as were used in Figure 1 for the two-alternative forced-choice recognition task. For REM, the criteria used were 1.0 in each of the 25% conditions and 0.8, 1.0, and 1.2 in the 75% conditions, assuming a decrease, no change, and an increase in the criterion, respectively. For ALT, the criteria used were 0.0 in each of the 25% conditions and −.33, 0.0, and .33 in the 75% conditions, assuming a decrease, no change, and an increase in the criterion, respectively. LF = low frequency; HR = hit rate; HF = high frequency; FAR = false-alarm rate.

iarity of HF words. In this case, REM predicts no effect of list composition for HF words. A smaller increase, however, would result in a small increase in HF HRs and FARs, and a larger increase would result in a small decrease in HF HRs and FARs.

In ALT, yes–no recognition involves comparing the $\ln L(x)$ associated with a test item to a criterion. If the $\ln L(x)$ is greater than the criterion, then the word is called "old." ALT predicts an LF HR advantage because $\ln L(x)$ tends to be greater for LF targets than for HF targets because $n(i)$ is greater for LF targets. The LF FAR advantage is predicted because LF items are expected to have more marked features if they were studied (as previously explained).

ALT once again predicts no list-composition effect based on the counteracting changes in the means and variances of the log-likelihood distributions produced by changes in $p(.,\text{old})$. On this matter, Glanzer et al. (1993) wrote: "The differences, however, have no effect on the hits and false alarms" (p. 564). Thus, any list-composition effect must be a criterion shift according to ALT. The right panels of Figure 5 show ALT's yes–no recognition performance when different criterion shifts are assumed. The middle panel assumes no criterion shift as a function of list composition: ALT predicts no list-composition effect. The right panel assumes a less strict criterion in the 75% condition: ALT predicts an increase in HRs and FARs for both HF and LF words. The bottom panel assumes a stricter criterion in the 75% condition: ALT predicts a decrease in HRs and FARs for both HF and LF words.

There are two reasons why a simple yes–no recognition task is used in Experiment 2A and a rating task in Experiment 2B. Empirically, the sensitivity and bias can be measured using the ratings data from each subject in Experiment 2B to construct a receiver operating characteristic (ROC) in *z*-transformed space, and its slope can be used to calculate a measure of sensitivity and the criterion location. These measures will be helpful when interpreting the pattern of HRs and FARs that we observe, and taking into account the slopes of the *z*-ROCs is important because they have been shown to be different for HF and LF words (Glanzer et al., 1993). The second reason for using both a yes–no and a ratings task is because ratings do not always correspond directly to the yes–no decision (e.g., Van Zandt, 2000). By using both procedures, the generality of the list-composition effect can be determined better.

Method

Subjects. One hundred forty-five students enrolled in introductory psychology courses at the University of Maryland participated in exchange for course credit. Eighty subjects performed the yes–no task, and 65 subjects performed the ratings task. For the yes–no task, 40 subjects were randomly assigned to either the 25% or the 75% condition. For the ratings task, 32 and 33 subjects were randomly assigned to the 25% and 75% list-composition conditions, respectively.

Design and materials. The basic design and materials used in Experiment 1 were used here with the following exceptions. List composition (25% vs. 75% HF words) was manipulated between subjects: For each subject, a 100-item study list and a 200-item test list were constructed. The study list was formed by randomly selecting $n = 25$ or $n = 75$ HF words and $100 - n$ LF words. The test list consisted of the studied items and the same number of HF and LF foils. For each yes–no test trial, a word was presented on the computer screen, and subjects were instructed to answer "yes" if it was a studied word or "no" using the *d* and *k* keys, respectively. For each ratings trial, keys *1*, *2*, and *3* indicated that an item was not studied (*1* = low confidence, *2* = moderate confidence, and *3* = high confidence), and keys *7*, *8*, and *9* indicated that an item was studied (*7* = low confidence, *8* = moderate confidence, and *9* = high confidence).

Results

Word frequency. The mean HRs and FARs for Experiment 2A are presented in the left panel of Figure 6. HRs are greater for LF words than for HF words, $F(1, 78) = 45.24$, $MSE = 0.34$, and the FARs are less for LF words than for HF words, $F(1, 78) = 136.72$, $MSE = 1.07$. The confidence ratings from Experiment 2B were converted to HRs and FARs by collapsing over the "old" ratings. The middle panel of Figure 6 shows that HRs are greater for LF words than for HF words, $F(1, 63) = 15.90$, $MSE = 0.15$, and FARs are less for LF words than for HF words, $F(1, 63) = 100.80$, $MSE = 0.97$.

List composition. The left panel of Figure 6 shows that in Experiment 2A the FARs for the LF words decreased significantly as the proportion of HF words studied increased, $F(1, 78) = 4.48$, $MSE = 0.036$. The interaction of list composition and word

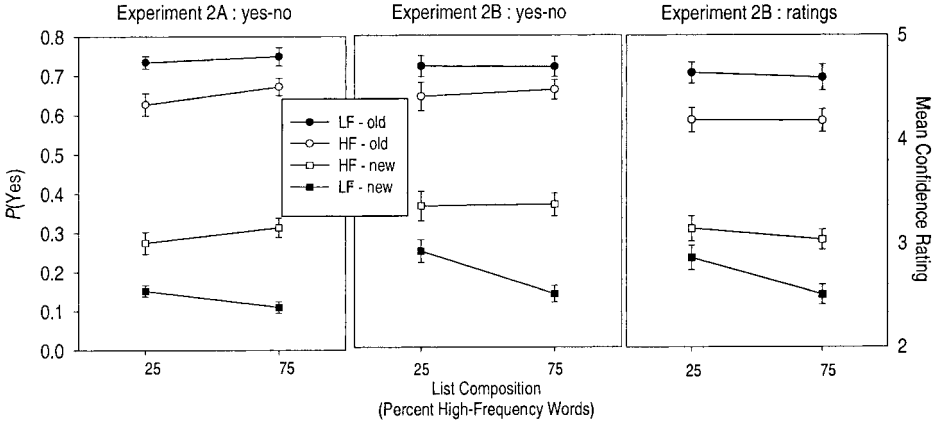

Figure 6. Yes–no and ratings performance as a function of pair type and list composition. Error bars indicate standard errors. LF = low frequency; HF = high frequency.

frequency was significant for FARs, $F(1, 78) = 8.29$, $MSE = 0.06$, but not for HRs, $F(1, 78) = 1.19$. List composition did not significantly affect HRs ($F < 1$) or FARs, $F(1, 78) = 1.00$.

A similar pattern of data was observed in Experiment 2B. The middle panel of Figure 6 shows that the LF FAR decreased as the proportion of HF words studied increased, $F(1, 63) = 9.47$, $MSE = 0.14$. The HF FAR ($F < 1$) was not significantly affected by list composition. The interaction of list composition and word frequency was significant for FARs, $F(1, 63) = 10.76$, $MSE = 0.10$, but not for HRs ($F < 1$). List composition did not significantly affect the HRs ($F < 1$) or FARs, $F(1, 63) = 1.96$.

The average ratings, shown in the right panel of Figure 6 (1 and 6 are the lowest and highest levels of confidence), are consistent with the HRs and FARs. As the proportion of HF words studied increased, the ratings for LF foils decreased significantly, $F(1, 63) = 5.48$, $MSE = 2.07$. List composition did not significantly affect the ratings for targets or HF foils (all Fs < 2.75). Ratings were greater for LF than for HF targets, $F(1, 63) = 45.08$, $MSE = 6.15$, and smaller for LF than for HF foils, $F(1, 63) = 36.29$, $MSE = 5.32$.

The ratings were used to compute individual-subject ratings z-ROCs, and their slopes, m, were used to estimate the criterion location, $C = [-1/(m + 1)][z(\text{HR}) + z(\text{FAR})]$, where C estimates the location the criterion relative to the mean of a foil distribution (MacMillan & Creelman, 1991). C for LF words was greater in the 75% than in the 25% list-composition condition, $t(63) = 2.53$, $p \leq .015$, but list composition did not affect C for HF words ($p = .995$). Thus, the yes–no criterion was positively related to the proportion of HF words studied relative to the LF-new distribution.

The slopes of z-ROCs were used to compute a measure of sensitivity for each subject, d_a (MacMillan & Creelman, 1991). The mean d_a for HF words was .94 and .92, and the mean d_a for LF words was 1.69 and 2.30 for the 25% and 75% conditions, respectively (*SEM*s = between .10 and .22 inclusive). The interaction between list composition and word frequency was significant, $F(1, 63) = 8.92$, $MSE = 2.85$, $p \leq .005$. Recognition of HF words was not affected by list composition, $t(63) = 0.11$, but LF words were better recognized as the proportion of HF words studied increased, $t(63) = 2.17$.

Discussion

Recognition of LF words but not HF words was affected by the composition of the study list. The FAR decreased and d_a increased for LF words as the proportion of HF words studied increased. This outcome is consistent with the findings from Experiment 1, which indicated that LF foils become relatively less familiar as the proportion of HF words studied increases. In addition, subjects adopted a stricter criterion for the 75% versus the 25% list composition. The bottom panels of Figure 5 show the ALT and REM predictions for yes–no recognition when the criterion increases: ALT predicts a decrease in both the HF and LF HRs and FARs, and REM predicts a decrease in the LF HR and the LF FAR. The observed effect was on the LF FAR only. Thus, ALT fails to qualitatively predict three out of the four observed trends in HRs and FARs, and REM fails to predict one out of four.

General Discussion

Our main finding is that LF words are better recognized as the proportion of HF words studied increases. We also found that the criterion location for yes–no recognition is a positive function of the percentage of HF words studied. In combination, these two findings more strongly constrain models designed to account for WFE. The goal for this section is to describe some of the implications of our findings for ALT and REM.

Our findings are problematic for both the ALT and REM models of the WFE. To predict our findings, a model needs to predict that the difference between the mean of the LF-new distribution and the means of the LF-old and both HF distributions increase as the number of HF words studied increases. Is either the ALT or REM theoretical framework amenable to solving this problem? We can think of a number of plausible assumptions that might produce this effect in REM, and here we describe one explicitly, as we momentarily defer our analysis of ALT. We will not consider here every possible REM model. Rather, the goal of this modeling exercise is to determine only if it is possible for the REM theory to predict our major finding: LF words are better recognized as the percentage of HF words studied increases. For the modified REM model, the representational and global-matching schemes of REM described in prior sections (and in Shiffrin & Steyvers, 1997) are assumed. With these assumptions, the modified REM model predicts a WFE (Shiffrin & Steyvers, 1997).

One way in REM to predict that the difference between the LF odds distributions increases as the number of HF words studied increases is to assume that LF words are encoded better when they are studied on lists dominated by HF words, perhaps because their unusual features stand out more or are more salient when presented in the context of HF words. Consider that the c parameter in REM is the probability of correctly storing a feature, and words are encoded well when c is relatively high. If c increases for LF words as the proportion of HF words studied increases, then LF images will tend to be stored better when lists are dominated by HF words. In this model, the λ_js for LF targets increase (on average) because c increases for LF words, and this produces the required increase in performance for LF words as the proportion of HF words studied increases. The model also predicts that the odds for HF words are positively related to the proportion of HF words studied (see The Effect of List Composition on 2AFC Recognition in REM section; also see Shiffrin & Steyvers, 1997). Thus, the mean of the LF-old distribution and the means of the HF distributions increase relative to LF new-item distribution, and this pattern of movement is what is required to predict a decrease in the probability of choosing LF foils for 2AFC recognition.

If the criterion location varies, then the effects of list composition on the means of the distributions may be offset by some stricter criteria. Such a shift will decrease the FAR for LF words (see the bottom panels of Figure 5). Our findings indicate that the criterion location used in yes–no recognition is positively related to proportion HF words studied. Thus, REM should be able to predict the yes–no data, and to the extent that the difference between the ratings and the yes–no task is the number of criteria used to make decisions, list composition will similarly affect performance of both tasks.

The recognition task varied between subjects, and the different groups performed at different overall levels of performance.

Table 1
Data and Fits for Experiments 1 and 2B

	Data[a]		New REM model		Old REM model		ALT	
Condition	25%	75%	25%	75%	25%	75%	25%	75%
				2AFC				
P(LF-old, LF-new)	.80	.82	.79	.83	.79	.79	.81	.81
P(HF-old, LF-new)	.75	.81	.75	.82	.75	.81	.73	.73
P(HF-new, LF-new)	.63	.78	.56	.68	.56	.67	.57	.57
P(HF-old, HF-new)	.69	.70	.71	.71	.71	.71	.69	.69
P(LF-old, HF-new)	.76	.79	.77	.76	.77	.72	.79	.79
P(LF-old, HF-old)	.68	.65	.61	.62	.61	.55	.64	.64
				Yes–no				
LF HR	.71	.71	.78	.78	.78	.72	.72	.65
HF HR	.63	.65	.65	.65	.65	.64	.65	.47
HF FAR	.36	.36	.32	.32	.32	.33	.37	.21
LF FAR	.24	.15	.28	.19	.28	.19	.25	.14
G^2			$G^2(9) = 0.193$		$G^2(10) = 0.236$		$G^2(13) = 0.340$	

Note. For each retrieving effectively from memory (REM) simulation, 500 simulated subjects were run with the following parameter: $w = 20$, $g_{HF} = .46$, $g_{LF} = .30$, $g = .41$, $t = 10$, $c = .7$, $u^* = .025$, criterion = .75, and criterion shift = .15. $d = .068$ in the new REM model and $d = 0$ in the old REM model. The attention/likelihood theory (ALT) predictions were derived from Equations 5 and 6 on the assumptions that p(.,old) decreases as the percentage of HF words studied increases. Because ALT does not predict a list-composition effect, we needed to choose either the 25% or 75% data to fit, and we chose to fit the data for the 25% condition. ALT's predictions were derived from p(new) = .10, n(HF) = 40, n(LF) = 60, $N = 1{,}000$, criterion = 0.0, and criterion shift = .33. 2AFC = two-alternative forced-choice; LF = low frequency; HF = high frequency; HR = hit rate; FAR = false-alarm rate.

[a] The data in the top half of the table are from Experiment 1; the data in the bottom half of the table are from Experiment 2B.

Therefore, we chose the yes–no data from Experiment 2B (rather than Experiment 2A) to predict because the overall level of performance was more similar to that for 2AFC recognition than if we chose to fit the yes–no data from Experiment 1.[4] Thus, no attempt was made to find a "best fitting" set of parameters for any model. Rather, we only wanted see whether the different models could qualitatively predict the major trends in the data from both tasks. Therefore, for the ALT and the modified REM model, we first found a set of parameters that provided a decent prediction for 2AFC recognition, and then we used those parameters to generate predictions for yes–no recognition. We also assumed, on the basis of our finding from Experiment 2B, that the yes–no criterion was positively related to the percentage of HF words studied.

To generate some predictions from the modified REM model, let $c_{HF}(i)$ and $c_{LF}(i)$ equal the probabilities of correctly encoding a feature from an HF and an LF word, respectively, given a study list consisting of i% HF words. Assume:

$$c_{LF}(75) = c_{LF}(25) + d \text{ and}$$

$$c_{HF}(75) = c_{HF}(25),$$

where $d > 0$, $c_{LF}(25) = c_{HF}(25)$, and the value of c used in Equation 3 is the mean of c_{HF} and c_{LF} —that is, $c(75) = (2 * c[25]) + d)/2$. Thus, d is the increase in the probability of correctly encoding an LF feature as the percentage of HF words studied increases. To generate a prediction for the REM model described by Shiffrin and Steyvers (1997), we set $d = 0$. All the other parameters were fixed to the parameters used to generate predictions for the modified REM model. The ALT and REM parameter values we used are listed in Table 1, and they correspond very closely with those used by Shiffrin & Steyvers (1997) and by Glanzer et al. (1993) to account for a variety of other findings (listed in Figure 1).

Performance of the modified REM model, the Shiffrin and Steyvers' (1997) REM model, and ALT are shown in Figure 7 and listed in Table 1. We want to determine two things: Do these models predict the 2AFC performance (top two rows of Figure 7), and can the same memory system predict the yes–no performance (bottom row of Figure 7)? A maximum-likelihood analysis of these predictions indicates that all three models can predict the data reasonably well—new REM: $G^2(9) = 0.193$, old REM: $G^2(10) = 0.236$, ALT: $G^2(13) = 0.350$, all ps $> .95$.

[4] The inferences drawn from the analyses of yes–no data from Experiments 2A and 2B are qualitatively similar. However, the decrease in the LF FAR in Experiment 2A was smaller than in Experiment 2B, and HRs and the HF FARs were slightly greater. The model predicts this if the criterion shift in Experiment 2A was not as great as the criterion shift in Experiment 2B. We fit the data from Experiment 2A with the same parameters used to fit the data from Experiment 2B, and a good qualitative fit was obtained, but the quantitative fit was not so good because the subjects' performance was better in Experiment 2A. With slightly different parameters, however, the model also does a good job for Experiment 2A.

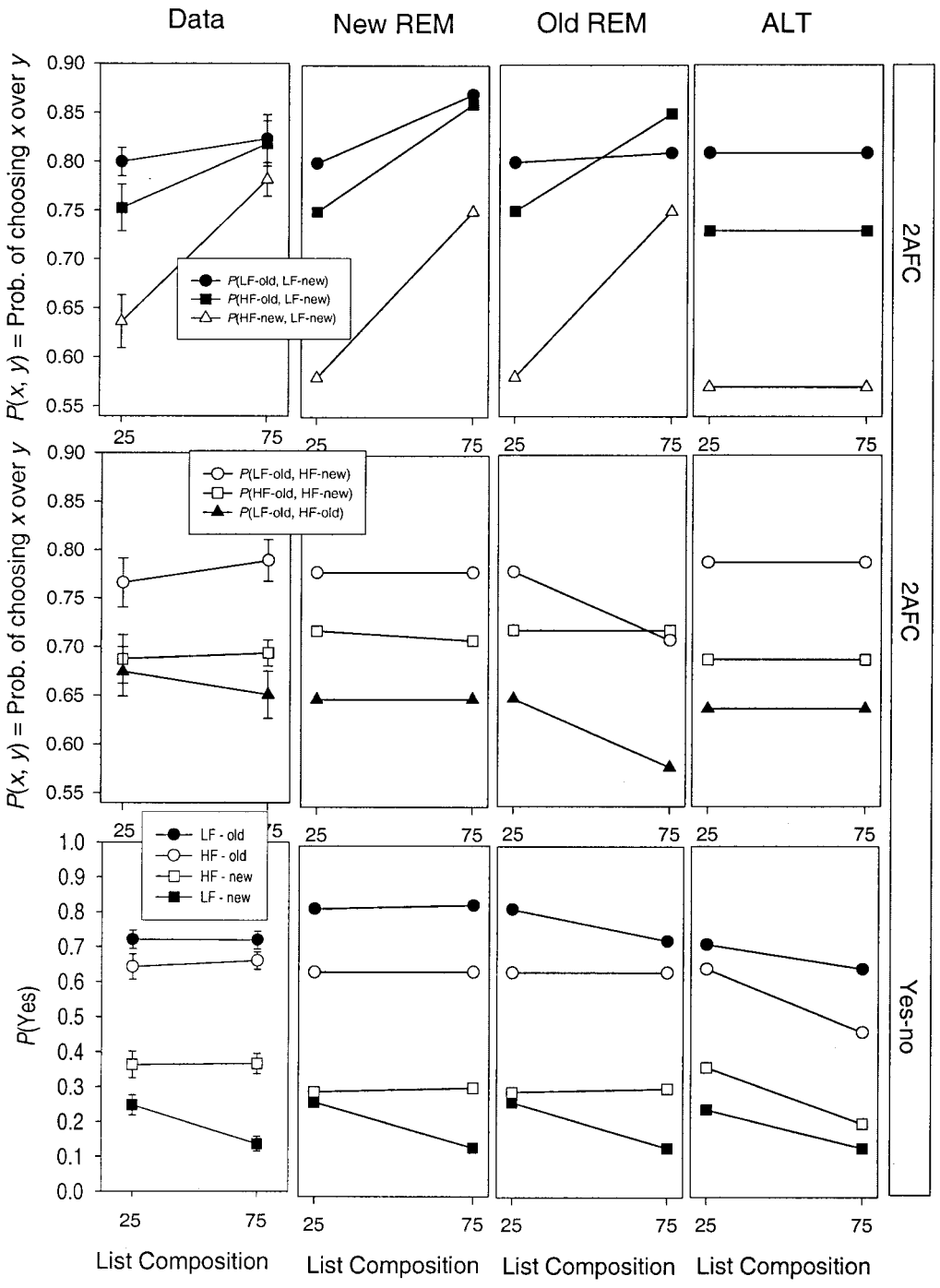

Figure 7. Performance of the revised retrieving effectively from memory (REM) model, REM (Shiffrin & Steyvers, 1997), and attention/likelihood theory (ALT). Error bars indicate standard errors. The parameters for these predictions are listed in Table 1. Prob. = probability; 2AFC = two-alternative forced-choice; LF = low frequency; HF = high frequency.

This is not too surprising because each model was designed to account for mirror effects in general and the WFE in particular, and hence each model captures much of the data.

When more than one model provides decent quantitative predictions, how do we choose among them? There is no clear-cut answer to that question (Shiffrin & Nobel, 1997). One factor that should be considered when evaluating models is the number of parameters each model needs to predict the data. Because the *d* parameter was added to REM, the new REM model has additional flexibility. ALT and the Shiffrin and Steyvers' (1997) REM model

have fewer parameters and also provide decent quantitative predictions.

It also seems prudent to take into account the qualitative predictions of each model: Figure 7 shows that the revised REM model mimics the inverse relationship between the probability of recognizing an LF foil and the proportion of HF words studied, whereas the REM model described by Shiffrin and Steyvers (1997) and ALT do not. ALT fails to predict the increases in *P*(LF-old, LF-new), *P*(HF-old, LF-new), and *P*(HF-new, LF-new), as well as the other smaller trends found in the middle-left panel of Figure 7. Shiffrin and Steyvers's REM model fails to predict the increase in *P*(LF-old, LF-new), and it incorrectly predicts large decreases in *P*(LF-old, HF-new) and *P*(LF-old, HF-old) because only HF words vary in familiarity with changes in list composition.

The modified REM model does a better job of predicting our major findings because LF words are better encoded on the 75% list and because HF words become more familiar. Even with the additional parameter, however, we found it difficult to find a set of parameters that would generate predictions that exactly match the data. One problem for the modified model is that it tends to underpredict the magnitude of the mirror effect for the null-comparison trials. Larger null-comparison effects are possible with different parameters, but they trade off with performance on standard comparison trials. For example, a greater *c* value will increase *P*(LF-old, HF-old), but it will also increase performance on the standard comparisons. Another problem for the modified REM model is that it has a difficult time predicting the minor trends in the middle row of Figure 7. For example, the data show that *P*(HF-old, HF-new) increases by 1%, *P*(LF-old, HF-new) increases by 3%, and *P*(LF-old, HF-old) decreases by 3%, which suggests that the increase in the mean of HF-old distribution is slightly greater than the increase in the mean of the HF-new distribution. This trend for HF words to be better recognized is relatively minor, however, and slightly reverses for d_a in Experiment 2B.

The more flexible REM theory shows some signs of being able to handle our main finding: LF words are better recognized as the percentage of HF words studied increases. Can ALT predict the data by adding the same assumption? If not, can ALT in its current form predict the data at all? The answer is "no" to both questions. To see why, consider that if we assume that α(LF) increases as the proportion of HF words studied increases, then the ALT concentering principle predicts that *P*(LF-old, LF-new), *P*(HF-old, LF-new), *P*(LF-old, HF-new), and *P*(LF-old, HF-old) will increase, and *P*(HF-new, LF-new) will decrease. However, Experiment 1 showed that only those 2AFCs involving an LF foil were affected by list composition. Thus, even if ALT were to predict a list-composition effect, it would be incorrect because our findings are inconsistent with the concentering principle of ALT. This is not the first time the concentering principle has been violated (e.g., Hirshman & Arndt, 1997; Hoshino, 1991; Murnane, Phelps, & Malmberg, 1999, among others). Thus, evidence disconfirming ALT's simple theory of recognition continues to mount.[5]

Some Final Comments on the Models

In the prior section, we described a REM model that does a better job than the REM model described by Shiffrin and Steyvers (1997) in predicting our major finding: LF words are recognized better as the percentage of HF words studied increase. However, the modified REM model does not make perfect quantitative predictions (in spite of having an additional parameter). There are, however, other plausible assumptions that might produce similar or better predictions in REM. For example, the number of attempts at storing LF features may vary (*t*), the system might expect more or fewer LF words at test by varying *g* in the activation function, or the composition of the retrieval cues at test may be affected by list composition. Fully exploring these models is beyond the scope of this article. Therefore, it remains a somewhat open question as to how problematic our findings are for the REM theory. For now we conclude that the additional modeling options combined with the performance of the modified REM model suggest that further theoretical and empirical investigations of list composition and WFEs within the REM framework are warranted.

Our findings pose a clear problem for the ALT theory. The assumption that memory is locally accessed is the main problem for ALT. Because access to memory is local, the composition of the study list is irrelevant when computing the activation of any one trace. In addition, metalevel rescaling of the decision axis based on *p*(.,old) also does not produce a list-composition effect.

---

[5] Another reason that we think that ALT cannot be modified slightly to accommodate our findings is because if ALT can handle our list-composition findings, then it cannot predict another critical list-composition finding: the *null list-strength* effect. A positive list-strength effect is observed when adding relatively well-encoded items to memory interferes with the ability to remember relatively less well-encoded items. For yes–no recognition memory, a null or "slightly negative" list strength is observed (Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990). It should be noted that the list-strength effect is a null or slightly negative list-composition effect because a list-strength manipulation is a list-composition manipulation. It should also be noted that adding LF words to memory is for all intents and purposes adding well-encoded items to memory within the ALT framework, and therefore the correspondence between a list-strength manipulation and manipulation of the percentage of HF words studied is a strong one for ALT.

Predicting both the WFE and the null list-strength effect has been a critical benchmark for the evaluation of many (if not all) new theories of recognition (e.g., Dennis & Humphreys, 2001; Estes, 1994; Glanzer et al, 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Indeed predicting the null list-strength effect inspired many of them and was deemed critical by Glanzer et al. (1993):

> Attention/likelihood theory, as presented here, does not have a problem with the absence of list strength effects. It does not predict them. To predict the presence of such effects would require an extension of the theory. The *one* way that such effects could be produced without extension of the theory would be if mixed lists produces a change in *d'* compared with pure lists . . . We have shown, however, that those different kinds of estimates produce identical *d's*. Because there is nothing in the theory that predicts a list strength effect, the findings of no list strength effect are not problematic for attention/likelihood theory as they are for global theories of memory. (p. 565, italics added)

Thus, ALT and several more recent global theories of memory have been shown to predict the null list-strength effect. The reasons why ALT predicts null list-composition effects is because access to memory at test is restricted to that of the test item, and because the use of metalevel information concerning the composition of the study lists at test—that is, *p*(.,old)—does not affect performance. Local access to memory allows ALT to predict a null list-strength effect, but it also makes it very difficult for ALT to accommodate our findings.

We found that list composition does affect recognition memory and that the patterns of data that we observe do not reflect a concentering of the underlying distributions. This is a violation of a fundamental property of the ALT activation function. Therefore, it appears that ALT needs to be modified to account for the present findings.

References

Clark, S. E., & Burchett, E. E. R. (1994). Word frequency and list composition effects in associative recognition and recall. *Memory & Cognition, 22,* 55–62.

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

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478.

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decisions and recognition memory. *Journal of Memory and Language, 27,* 633–648.

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 12,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

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, 61,* 23–29.

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

Hilford, A., Glanzer, M., & Kim, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition,* 25, 593–605.

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 275–289.

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

Hoshino, Y. (1991). A bias in favor of the positive response to high-frequency words in recognition memory. *Memory & Cognition, 19,* 607–616.

Kučera, H., & Francis, W. (1983). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

MacMillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge, England: Cambridge University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory, *Psychological Review, 105,* 724–760.

Murnane, K., Phelps, M. P., & Malmberg, K. J. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General, 128,* 403–415.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855–874.

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

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science, 9,* 47–52.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior, 6,* 156–163.

Shiffrin, R. M., & Nobel, P. A. (1997). The art of model development and testing. *Behavior Research Methods, Instruments, and Computers, 29,* 6–14.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600.

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

Appendix

To show that changes in $p(.,\text{old})$ have no effect on recognition in attention/likelihood theory (ALT), one needs to show that for any value of integer $x_i > 0$, $\ln L(x_A|i, j) > \ln L(x_B|i, j)$ for all $p(.,\text{old})$ where the class of A stimuli is better recognized than B stimuli. That is, we need to show that changes in $p(.,\text{old})$ do not affect the ordering of $\ln L(x_A|i, j)$ and $\ln L(x_B|i, j)$ on the decision axis. One should consider that

$$\ln L(x|i, j) = n(i) \cdot \ln\left[\frac{q(i, \text{old})}{q(\text{new})}\right] + x \cdot \ln\left[\frac{p(i, \text{old})\ q(\text{new})}{p(\text{new})\ q(i, \text{old})}\right]. \quad \text{(A1)}$$

Next, it should be noted that

$$\left[\frac{q(i, \text{old})}{q(\text{new})}\right] \quad \text{and} \quad \left[\frac{p(i, \text{old})\ q(\text{new})}{p(\text{new})\ q(i, \text{old})}\right]$$

are constant for any $p(.,\text{old})$ where $p(\text{new})$ is constant. Then let

$$\varphi = \left[\frac{q(i, \text{old})}{q(\text{new})}\right],$$

and let

$$\delta = \left[\frac{p(i, \text{old})\ q(\text{new})}{p(\text{new})\ q(i, \text{old})}\right],$$

and Equation A1 becomes

$$\ln L(x|i, j) = n(i) \cdot \ln(\varphi) + x \cdot \ln(\delta). \quad \text{(A2)}$$

It is trivial to show that $[n(\text{A}) \cdot \ln(\varphi)] > [n(\text{B}) \cdot \ln(\varphi)]$ for all φ when $n(\text{A}) > n(\text{B})$ and that δ does not vary for Stimuli A and B. Thus, $\ln L(x_A|i, j) > \ln L(x_B|i, j)$ when $n(\text{A}) > n(\text{B})$, which is necessarily the case in ALT (Glanzer et al., 1993). It is also straightforward to show using a similar argument for Equation 6 that for all $p(.,\text{old})$ and for any value of integer $x_i > 0$ that $p(x_A|i, j) > p(x_B|i, j)$. That is, a change in $p(.,\text{old})$ does not affect the ordering of $p(x_A|i, j)$ and $p(x_B|i, j)$.

Received October 11, 2000
Revision received December 14, 2001
Accepted December 14, 2001 ■