

Modeling the effects of verbal and nonverbal pair strength on associative recognition

JING XU

Iowa State University of Science and Technology, Ames, Iowa

AND

KENNETH J. MALMBERG

University of South Florida, Tampa, Florida

Operations that improve the accuracy of associative recognition can do so in qualitatively different ways. Increasing repetitions and study time increases hit rates but has small effects on false alarm rates, and the specific patterns of false alarms are dependent on the stimuli (e.g., pairs of words, pseudowords, faces, or Chinese characters). In contrast, manipulating the type of stimuli that make up pairs produces a robust mirror effect: The hit rate is greater, and the false alarm rate is lower, for better recognized stimuli. To explain these findings, a model of single-item recognition is extended to associative recognition. Within this dual-process framework, the present results suggest that words are encoded more extensively than nonverbal stimuli and that recognition of frequently encountered stimuli (words and faces) is more likely to be based on recollection than is recognition of uncommon stimuli (pseudowords and Chinese characters).

In an *associative recognition* (AR) task, subjects study pairs of items. Items that are studied together are *intact pairs*, items that are not studied together are *rearranged pairs*, and subjects discriminate between them at test. Positive endorsements of intact and rearranged pairs are respectively referred to as *hits* and *false alarms*. Generally speaking, accuracy improves as the difference between hit rates (HRs) and false alarm rates (FARs) increases, and thus, factors can improve AR via qualitatively different patterns of HRs and FARs: HRs can increase and FARs can decrease, HRs can increase and FARs can remain steady, or HRs can remain steady and FARs can decrease.

Different operations affect AR in different ways. For instance, varying the type of stimuli that make up to-be-remembered pairs (referred to here as *pair type*) produces a *mirror effect* (Glanzer & Adams, 1985): HRs are greater for words than for nonverbal stimuli, but FARs are less for words than for nonverbal stimuli (Criss & Shiffrin, 2004; Greene, 1996, 2004; Hockley, 1994). In contrast, a mirror effect is not always observed with increases in repetitions or study time: When strength is varied within a list, the probability of endorsing intact pairs increases with target repetitions, but the probability of endorsing rearranged pairs remains fairly steady (e.g., Kelley & Wixted, 2001); but when strength is varied between lists, a mirror effect is observed (Clark & Shiffrin, 1992). Thus, some operations have robust effects on both HRs and FARs, whereas other operations primarily affect HRs. (To the best of our

knowledge, no factors have been identified that produce steady HRs and decreasing FARs.)

It is unknown why these variables affect AR in different ways or how pair type and standard strengthening operations interact. In addition, it is unknown how generalizable the pair type effect is, because words have traditionally been used in recognition memory experiments. To address the empirical questions, we present an AR experiment in which repetitions, study time, and pair type (words, pseudowords, Chinese characters, or faces) were varied. Using an extension of a dual-process memory model (Malmberg, Holden, & Shiffrin, 2004), we hypothesized that strengthening different types of stimuli might affect the encoding of pairs, the contribution of recollection to AR, or both.

We will begin with a discussion of when mirror effects occur and do not occur and the constraints these findings place on models of recognition memory. In the next section, we will generate predictions of an extension of Malmberg, Holden, and Shiffrin's (2004) recognition model in order to understand how the pair type mirror effect and the effect of repetitions can be explained by the same model. Finally, we will report the results of an experiment in which standard strengthening operations and pair type were varied. Given these results, we then will fit the three competing models to the data in order to quantitatively choose among them.

When Mirror Effects Occur

Prior to Glanzer and Adams's (1985) discovery of the mirror effect, recognition memory was usually measured

K. J. Malmberg, malmberg@cas.usf.edu

in terms of a composite measure derived from HRs and FARs (e.g., d' , HRs – FARs, percentage correct, etc.). However, Glanzer and Adams (1985) showed that variables that produce mirror effects were difficult or impossible to explain, even though extant models were not challenged by the composite measures. As a result, researchers have placed greater importance on explaining patterns of HRs and FARs when developing models.

The major theoretical advancement made possible by Glanzer and Adams's (1985) discovery was the *likelihood* class of models, which produce mirror effects as a by-product of their Bayesian frameworks (Dennis & Humphreys, 2001; Glanzer & Adams, 1990; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). These models typically assume that single-item recognition is based on the outcome of a global-matching retrieval process, whereby a retrieval cue is compared with a large number of episodic memory traces. The result is a level of familiarity associated with the test item, and increasing the number of target presentations increases their familiarity and decreases the familiarity of foils. Hence, HRs increase and FARs decrease. For instance, increases in target repetitions increase HRs and decrease FARs for *old–new single-item recognition* (e.g., Ratcliff, Clark, & Shiffrin, 1990).

An equally influential discovery is the finding that mirror effects are not ubiquitous: For instance, HRs increase and FARs remain steady when targets and foils cannot be discriminated on the basis of purely random factors. Such findings include those from the process dissociation procedure (see Jacoby, 1998, for a review), which requires the discrimination of items that appeared in similar contexts, from exemplar discrimination, which requires the discrimination of the occurrence of exemplars from the same semantic category (Morrell, Gaitan, & Wixted, 2002), and from plurality discrimination (Hintzman & Curran, 1994; Hintzman, Curran, & Oppy, 1992; Malmberg, Holden, & Shiffrin, 2004; Sheffert & Shiffrin, 2003), which requires the discrimination of target words (e.g., *barn*) from words of the opposite plurality (e.g., *barns*). Thus, repetitions always improve single-item recognition accuracy, but the pattern of HRs and FARs that gives rise to the improvement is variable and depends on the similarity of the targets and the foils.

We are specifically interested in how different factors affect the accuracy of AR. For AR, we note that the items making up rearranged pairs are, in fact, studied. Hence, intact and rearranged pairs are systematically related, much as items of different plurality are systematically related. Not surprisingly, therefore, the effect of repetitions on AR is similar to the effect of repetitions on a plurality discrimination task. For instance, Kelley and Wixted (2001) found that increasing pair repetitions increased HRs but had little or no effect on FARs. A stronger manipulation was used by Cleary, Curran, and Greene (2001), but still FARs decreased only slightly. Thus, when targets and foils are similar, the use of standard strengthening operations has a large positive effect on HRs and little or no effect on FARs, regardless of whether the memory test is single-item recognition or AR.¹

This finding stands in marked contrast to the robust mirror effect that is observed when targets and foils are chosen randomly (Glanzer & Adams, 1985), and the effect of repetitions on AR is inconsistent with many models of recognition (Kelly & Wixted, 2001). For instance, *threshold models* assume that AR is based on recalling that a pair of items was or was not studied together (e.g., Diller, Nobel, & Shiffrin, 2001; Yonelinas, 1997). Accordingly, increasing encoding strength should improve the ability to recall that intact pairs were studied and rearranged pairs were not, and recall-only models predict that FARs should decrease as the number of target presentations increases.

Alternatively, *compound cue* models of AR are single-process continuous-state models. They assume that pairs are represented by a concatenation of two memory traces and that memory is probed with a compound cue consisting of the representations of the items making up a test pair (e.g., Gronlund & Ratcliff, 1989). The result of the memory probe is a continuous-random variable, often referred to as *familiarity*. Increasing repetitions or study time increases the number of item features stored, and the familiarity of intact pairs is positively related to the number of features stored during study. Because familiarity is a positive function of the similarity of a retrieval cue and the contents of memory (e.g., Gronlund & Ratcliff, 1989), increasing the encoding strength of target items will increase the familiarity of rearranged pairs, and hence, FARs should increase as the number of target presentations increases (Hintzman, 1986; Shiffrin, Huber, & Marinelli, 1995). This prediction is disconfirmed by the null effect of repetitions on FARs.

Other single-process continuous-state models of AR predict the null effect of repetitions on FARs by assuming that associative information is stored during study, in addition to the representations of the individual items (Criss & Shiffrin, 2004; Doshier, 1984; Murdock, 1997). We refer to these as *independent cue* models, because the associative information and the item information are statistically independent. Strong versions of the independent cue model assume that the recognition decision is based only on the familiarity of the association, which is generated by a global-matching comparison (Murdock, 1997). Since the familiarity of the association is independent of the familiarity of the items, repetitions increase the familiarity of intact pairs but have no effect on the familiarity of the rearranged pairs. Thus, HRs are predicted to increase and FARs are predicted to remain steady as the number of target presentations increases, a prediction confirmed by Kelley and Wixted's (2001) findings.

Weaker versions of the independent cue models assume that both item and associative information are independently stored during study and are jointly used to probe memory at test (Hockley, 1992). Thus, increasing target presentations has no effect on the familiarity of the association, but it does increase the familiarity of the items. As a result, FARs are predicted to increase with increases in target presentations, and this class of models is disconfirmed (Kelley & Wixted, 2001).

Alternatively, Kelley and Wixted (2001) described the *some-or-none* model, in which item and associative information are distinctly represented. On some test trials, only item information is available. On the remaining trials, item information is linearly combined with associative information to form the evidence on which the AR decision is based. The some-or-none model is similar to the Hockley (1992) independent cue model, because repetitions increase item strength and, hence, repetitions increase the familiarity of intact and rearranged pairs. However, repetitions affect the associative strength of intact and rearranged pairs in opposite ways. Repetitions increase and decrease the associative familiarity of intact and rearranged pairs, respectively, according to the some-or-none model. For Kelley and Wixted's (2001) data set, the strengthening of item and that of associative information offset each other for rearranged pairs, and thus, repetitions have little or no effect on FARs.

To summarize, the repetition mirror effect for single-item recognition holds only when targets and foils are chosen randomly. Other recognition tasks, such as AR and plurality discrimination, require the discrimination of targets and systematically related foils. In these cases, improved recognition is the result of increases in HRs only. The challenge is to explain all of these in a common theoretical framework.

Modeling Associative Recognition

All of the models that have been described thus far are limited only to AR. Here, we adopt a more expansive, global memory approach to modeling AR, which entails accounting for the performance of a variety of memory tasks within a common theoretical framework.

We noted above that AR and plurality discrimination share many task features. Specifically, foils are relatively familiar, as compared with new items (although not as familiar as targets), and thus, familiarity is a relatively ineffective basis for discriminating them from targets. Therefore, it is not surprising that there is growing evidence that AR and plurality discrimination are affected similarly by several variables. For instance, repetitions and study time produce the same pattern of HRs and FARs (Hintzman et al., 1992; Kelley & Wixted, 2001; Malmberg, Holden, & Shiffrin, 2004). In addition, the accuracy of both AR and plurality discrimination is impaired under speeded response conditions, because FARs increase as repetitions increase; this pattern of performance is similar to that obtained from older adults (Healy, Light, & Chung, 2005; Hintzman & Curran, 1994; Light, Patterson, Chung, & Healy, 2004; Malmberg & Xu, 2007; Naveh-Benjamin, Craik, & Ben-Shaul, 2002).

Because AR and plurality discrimination have much in common, it would be useful to have a single framework that describes performance in both tasks. One alternative is to adapt an independent cue model or the some-or-none model. Potentially, the independent cue and some-or-none models could account for the effects of time pressure and age on AR by assuming that these variables make it difficult to form associative retrieval cues. However, the independent cue models and the some-or-none model

need to be augmented in order to account for plurality discrimination, because they revert to a standard single-item global-matching model when applied to the plurality discrimination task. They assume that single-item recognition involves the matching of item information and does not involve the matching of associative information (Murdock, 1997), and hence, they incorrectly predict that FARs should increase with repetitions for a single-item recognition task (Hintzman & Curran, 1994; Hintzman et al., 1992; Malmberg, Holden, & Shiffrin, 2004; Morrell et al., 2002; Sheffert & Shiffrin, 2003).

Another approach is to extend a model of single-item recognition to AR. For instance, the retrieving effectively from memory (REM) models have been shown to account for the effect of repetitions on single-item recognition, both when targets and foils are chosen randomly and when they are not (Malmberg, Holden, & Shiffrin, 2004; Malmberg, Zeelenberg, & Shiffrin, 2004; Shiffrin & Steyvers, 1997). Malmberg, Holden, and Shiffrin's dual-process REM model was designed to account for plurality discrimination, and thus, it might also be a good candidate for describing AR.

According to the REM model, a retrieval cue is first used to generate a familiarity value by comparing it with memory in a global-matching fashion (e.g., Clark & Gronlund, 1996). If the familiarity value does not exceed a criterion, the response is *new*. If the familiarity exceeds a criterion, an attempt to recall the details of a target trace is attempted. If the recollection fails, a guess is made. If the recalled details match the stimulus, the response is *old*, and recalling mismatching details allows for the correct rejection of an otherwise highly familiar stimulus. When an increase in foil familiarity due to an increase in target repetitions is offset by an increase in the ability to recall that the item was not studied, the dual-process model predicts steady FARs. To describe the effect of repetitions and study time on the accuracy of AR, additional assumptions about the nature of associative representation need to be made. In the next sections, we will describe this extension of the REM dual-process model and show that it accounts for the effect of strengthening operations on AR.

Some models of AR were designed to account for the null effect of repetitions on FARs (e.g., the some-or-none model). We noted, however, that all operations do not affect the accuracy of AR in the same way. Specifically, varying pair type produces a robust mirror effect (Criss & Shiffrin, 2004; Greene, 1996, 2004; Hockley, 1994). Thus, the second theoretical challenge is to develop an AR model that predicts that standard strengthening operations affect only HRs and that pair type manipulations affect both HRs and FARs. To do so, we will fit the different versions of the present model to a comprehensive set of new empirical findings, in order to let the data select the proper model. Before doing so, however, we first will describe our framework.

A Dual-Process Model of Associative Recognition

Associative representation and storage. In REM, generic knowledge is stored in lexical/semantic memory

traces, and events are stored in episodic traces. When words are studied, their lexical/semantic traces are retrieved into a short-term memory buffer. New episodic traces are created by copying the values from the lexical/semantic traces residing in the short-term memory buffer to new incomplete and error-prone episodic vectors.²

Specifically, lexical/semantic traces are assumed to be vectors of w features whose values, V_i are geometrically distributed integers. On a given attempt to store a feature, u^* is the probability of storing a feature in an episodic trace, and c is the probability of copying that feature correctly from a lexical/semantic trace. If an error in encoding is made, a feature value is drawn randomly from the geometric distribution. A zero value represents that no feature is stored.

The number of features stored increases with study time, but the amount of extra storage diminishes over time: $t_j = t_{j-1}(1 + e^{-b})$, where t_j is the number of attempts at storing a feature for an item residing continuously in the short-term buffer for j seconds and b is a rate parameter (Malmberg & Shiffrin, 2005). When an item is repeated, an additional t_j attempts at storing features occur. The additional features are usually accumulated in an existing trace, rather than a new trace being created (Malmberg & Shiffrin, 2005; Shiffrin & Steyvers, 1997, 1998), in order to account for the effects of a variety of strengthening operations. Here, we simply assume that this probability is 1.0. Thus, the probability of storing a feature m , given that the item it belongs to was studied r times for j seconds of study time, is

$$P(m) = 1 - (1 - u^*)^{rt_j}. \quad (1)$$

When a pair of items is studied, a concatenation of two episodic traces is stored.

Familiarity-based retrieval. We assume that a basic difference between AR and plurality discrimination is that foils are less similar to targets in AR than in plurality discrimination. For plurality discrimination, Malmberg, Holden, and Shiffrin (2004) assumed that vectors that represent words of opposite plurality share almost all their features (e.g., 80% or 90%). Thus, a foil is highly similar to one trace in memory. For AR, we assume that the concatenation of the two lexical/semantic or temporary short-term vectors corresponding to stimulus pair serves as a retrieval cue (cf. Gronlund & Ratcliff, 1989). Thus, the compound cue potentially overlaps in only half of its features with two target traces. In practice, the degree of overlap will be somewhat smaller, due to the incomplete and error-prone nature of episodic traces.

The compound cue is matched to all of the concatenated traces stored during study. For each trace j , a likelihood ratio, λ_j , is calculated as

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

where g is the geometric distribution parameter, c is the probability of copying that feature correctly from a lexical/semantic trace, n_{jq} is the number of mismatching

features in the j th concatenated trace, and n_{ijm} is the number of features in the j th concatenated trace that match the features in the compound retrieval cue. λ_j represents the degree of match, or *activation*, of the concatenated trace j in response to probe with the retrieval cue. The more similar the compound cue is to the concatenated trace j , the greater λ_j will be. The recognition decision is based on

$$\Phi = \frac{1}{n} \sum_{j=1}^n \lambda_j.$$

If Φ exceeds a subjective criterion the response is positive.

Increasing the number of target presentations or study time increases the number of matching features between the retrieval cue and the contents of the target traces. Familiarity increases with increases in the number of target presentations or study time. HRs are therefore predicted to increase. FARs are predicted to increase because the compound retrieval cues formed from rearranged pairs overlap with target traces. Thus, the single-process familiarity model is disconfirmed by the null effect of repetitions on AR (Kelley & Wixted, 2001; Malmberg, Holden, & Shiffrin, 2004).

Recalling to reject. Malmberg, Holden, and Shiffrin (2004) suggested that familiarity is a relatively inefficient basis for discriminating between targets and similar foils. They further assumed that when targets and foils are known to be very similar, a recall mechanism will be invoked after the familiarity value exceeds a subjective criterion. The goal of invoking recall is to recover the features that distinguish targets from similar foils, so that the foils can be correctly rejected despite their high degree of familiarity.

Figure 1 illustrates the decision-making steps in Malmberg, Holden, and Shiffrin's (2004) model. A familiarity value is computed in the same manner as that described above. If it does not exceed a criterion, the response is *new*. If it does exceed a criterion, the outcome of the recollection process is examined. If recollection is successful and the recovered details match the stimulus, the response is *old*. If the details do not match the stimu-

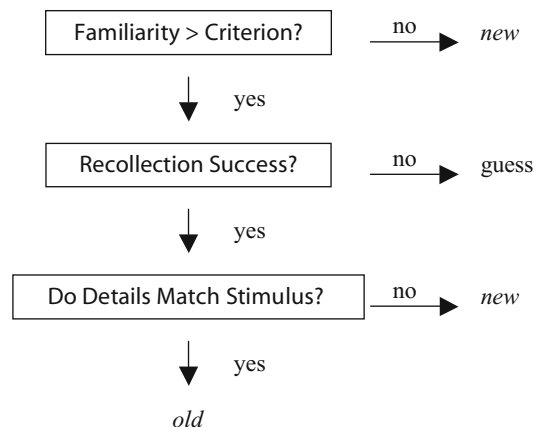


Figure 1. Flowchart for the REM dual-process model.

lus, the response is *new*. If recollection fails, the recognition decision is based on a guess. The probability of guessing *old* is γ .

We assume that recall conforms to the standard set of assumptions in the SAM (Raaijmakers & Shiffrin, 1980) and REM (Malmberg & Shiffrin, 2005) models. After the probe of memory, a trace is sampled with replacement. The greater the similarity between trace i and the cue, the more likely it is that trace i will be sampled. The greater the similarity between trace k and the cue, the less likely it is that trace i will be sampled, where traces $k \neq i$. Once a trace is sampled, an attempt is made to recover its contents. If recovery is successful and the contents match the stimulus, the response is *old*. If recovery is successful and the contents do not match the stimulus, the response is *new*. Otherwise, the subject guesses. We refer to this as a *recall-to-reject* strategy (see also Atkinson & Juola, 1974; Rotello, & Heit, 1999, 2000; Rotello, Macmillan, & Van Tassel, 2000).

The contribution of recollection to recognition performance is partly determined by how well traces are encoded, by interference, and by control processes that implement different retrieval strategies. We further assume that the recall-to-reject strategy requires a degree of cognitive effort or control beyond that which familiarity-based retrieval imposes. Hence, strategic, situational, and subjective factors can affect the tendency to rely on recollection as a source of evidence on which to make the recognition decision. For instance, instructions to respond very quickly are likely to reduce the emphasis on a recall-to-reject strategy (cf. Hintzman & Curran, 1994; Light et al., 2004; Naveh-Benjamin et al., 2002).

The relatively complex control and retrieval assumptions are expressed by a simple function (Malmberg, Holden, & Shiffrin, 2004). For pairs studied r times for j seconds, the probability that recollection is the basis for the *old-new* decision is

$$\hat{q} = ac [1 - (1 - u^*)^{rj}]. \quad (3)$$

The a parameter varies between 0 and 1. It scales the contribution of recollection to performance independently of how well items are encoded. When $a = 0$, $\hat{q} = 0$, and the dual-process model reverts to the single-process familiarity-based model (Shiffrin & Steyvers, 1997). When $a = 1$, $\hat{q} = c [1 - (1 - u^*)^{rj}]$, and the contribution of recollection to performance is bounded only by how well pairs were encoded at study, which is the probability that a feature was stored correctly, given that the pair was studied r times for j seconds.

Equation 3 is, of course, an overly simplistic characterization of the recall-to-reject strategy. For instance, factors that might influence a are interference, situations or strategies that emphasize speed over accuracy, and age, among others. A more complete model would implement the retrieval and control processes. Nevertheless, the simplified model suffices for the present purpose of describing the effects of different variables on the accuracy of AR.

Simulations. The accuracy model is influenced by two factors: how well pairs are encoded (c and u^*) and the tendency to utilize or the effectiveness of a recall-to-reject

strategy (a). For moderate values of a , the model predicts that increasing presentations or study time will increase HRs and have little or no effect on FARs. However, the effect of pair type produces a robust mirror effect: HRs are higher and FARs are lower for words than for nonverbal stimuli (Criss & Shiffrin, 2004; Greene, 1996, 2004; Hockley, 1994). Thus, pair type deficits could be due to encoding, recollection, or both factors, but the mirror pattern puts strong constraints on the model. We will demonstrate these constraints with a series of simulations.

To compare the effects of encoding on HRs versus FARs, we conducted a Monte Carlo simulation in which the number of target presentations was varied at three levels (one, two, and six) and study duration was varied at two levels (1.5 and 4.0 sec). All the parameters were fixed, except for u^* (i.e., the probability of storing features in a unit of time), to values that had been used previously to characterize a variety of episodic memory tasks (e.g., Criss & Shiffrin, 2004; Malmberg, Holden, & Shiffrin, 2004; Malmberg & Murnane, 2002; Malmberg, Zeelenberg, & Shiffrin, 2004; Shiffrin & Steyvers, 1997).³

Figure 2 shows that HRs increase as u^* increases, because increasing feature encoding produces more matching features and the probability increases that an intact pair's familiarity will exceed the criterion. When this occurs, a recall attempt is made, and according to Equation 3, increasing u^* increases the likelihood that this attempt will be successful (q). Even when recall fails, however, pairs will usually still be called *old* because the bias to guess *old* (γ) is assumed to be very high (.9).⁴ Thus, the target pair will usually be called *old* once it exceeds the criterion, regardless of whether the recall-to-reject strategy is successful or not.

In contrast to HRs, increasing u^* simultaneously exerts upward and downward pressure on FARs. The upward pressure on familiarity increases the probability that a rearranged pair's familiarity will exceed the criterion, but the increased effectiveness of the recall-to-reject strategy provides opposing evidence that the pair was not studied. Thus, FARs remain fairly steady when encoding strength is increased, increasing a small amount and then decreasing a small amount.⁵ Figure 2 further shows that for all intents and purposes, increasing repetitions or study time has the same effect as increasing u^* . Thus, only after many repetitions and/or with relatively long study times does the negative pressure on FARs dominate, and a very small (perhaps undetectable) mirror effect is predicted.

To compare the effects of varying recollection on HRs versus FARs, a similar simulation was conducted in which only a was varied. Note that when $a = 0$, the dual-process model is equivalent to the single-process global-matching model, and hence, it is not considered. Figure 3 shows that when $a = .2$, FARs are relatively high and that they increase steadily as the number of target presentations and study time increase. When $a = 1.0$, the FARs are relatively low, and they tend to initially increase and then decrease as the number of target presentations and study time increase. This might correspond to a subject who attempts to utilize a recall-to-reject strategy on each test trial. Note, however, that in contrast to changing the strength of encoding (u^*),

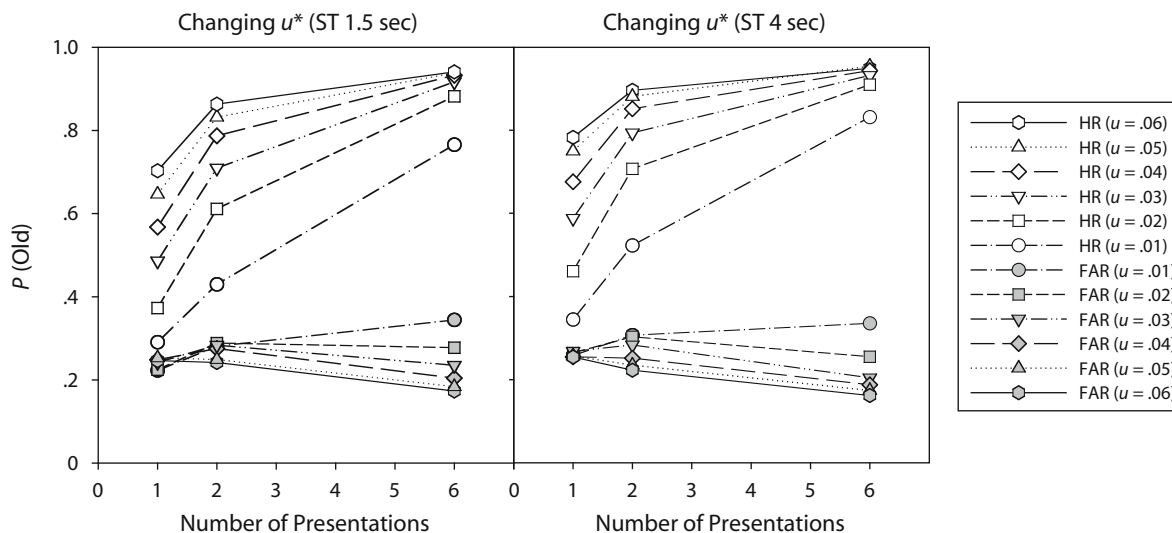


Figure 2. The effect of varying encoding strength (u^*) on associative recognition. In this simulation, $a = 1.0$, and the probability of encoding a feature on a given attempt is varied from .01 to .06. Two study times (STs) are simulated (1.5 and 4.0 sec). The other parameters used in this simulation are $w = 20$, $t_1 = 7$, $c = .7$, $g = .4$, $b = 1.0$, $a = 1.0$, *old-new* criterion = 1.0, and $\gamma = 0.9$. These parameters were selected because they have been used in a variety of other applications of REM (e.g., Shiffrin & Steyvers, 1997). HR, hit rate; FAR, false alarm rate.

changing the contribution of recollection to AR has little effect on HRs, because subjects are biased to guess *old* when recollection fails. Thus, variations in the a parameter primarily affect FARs and have little or no effect on HRs.

EXPERIMENT

The Interactions Between Strengthening Operations and Pair Type

The model predicts the effects of repetitions and study time on AR on the assumption that these factors primarily influence how well pairs are encoded. Both repetitions and study time increase the number of features stored, and the corresponding increase in pair familiarity increases HRs; but for foils, the increase in pair familiarity is offset by an increase in the contribution of recollection. Thus, standard strengthening operations produce little or no effect on FARs.

The accuracy model also suggests that the pair type mirror effect is due to both encoding and retrieval benefits associated with the pair types that are better recognized. On the encoding side, it might be, for instance, that common stimuli (e.g., words) are more “fluently” processed than nonverbal stimuli and, therefore, more features are stored in a given amount of time (Jacoby & Dallas, 1981), or more information might be known about common stimuli, and hence, more features are available for encoding. In either case, there would be an encoding advantage for common stimuli in the accuracy model, and therefore, they produce greater HRs. In addition, nonverbal or novel stimuli do not have labels associated with them; hence, recovery of episodic details might be impaired, or recollection of the details might be slower or more difficult (Curran, 1999; Curran, Schacter, Norman, & Galluccio,

1997; Gardiner & Java, 1990). Hence, the recall-to-reject strategy is emphasized less for nonverbal stimuli, and the rearranged nonverbal pairs will produce higher FARs than will the rearranged verbal pairs.

There are, however, alternative explanations for the recognition advantage of words over nonwords. One is that novelty is the critical factor determining the extent to which recollection contributes to performance. To test this hypothesis, we include in the design pairs of novel faces and pairs of Chinese characters. If novelty determines the contribution of recollection to AR, performance should be similar for pseudowords, faces, and Chinese characters and should be uniformly worse than performance for words.

A different hypothesis is that neither lexicality nor novelty is what determines the contribution of recollection to AR; the primary determinant is the frequency with which different stimuli are encountered in everyday life. That is, one might expect that differences in expertise in processing different types of stimuli might be the key factor in determining the contribution of recollection to AR. If so, we should observe similar patterns of performance for words and faces (patterns that reflect relatively high levels of recollection) and similar patterns of performance for pseudowords and Chinese characters (patterns that reflect relatively low levels of recollection).

Method

Although the effects of traditional strengthening operations and pair type have been documented, the interaction between these variables is unknown. In this experiment, pairs of words, pseudowords, faces, or Chinese characters were studied once, twice, or six times, with each presentation lasting for 1.5 or 4 sec. The subjects were instructed to discriminate intact pairs from rearranged pairs at test. We will use these data to conduct a series of modeling analyses with which to interpret, within the present framework, the effects of study time, repetitions, and pair type on AR.

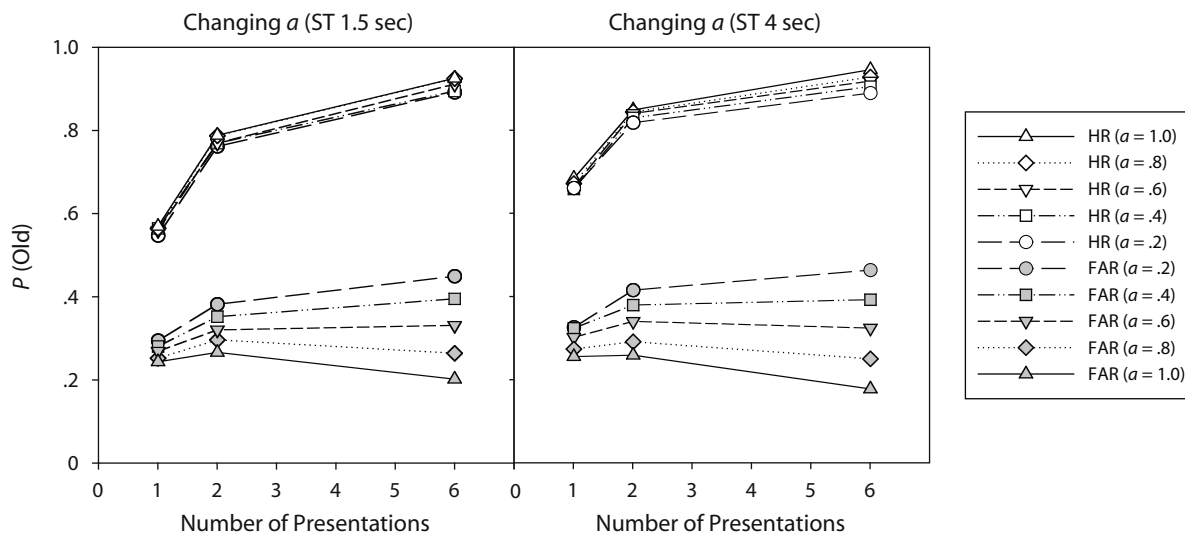


Figure 3. The effect of varying the use of a recall-to-reject strategy (a) on associative recognition. In this simulation, $u^* = .04$, and the probability of encoding a feature on a given attempt is varied from .01 to .06. Two study times (STs) are simulated (1.5 and 4.0 sec). The other parameters used in this simulation are $w = 20$, $t_1 = 7$, $c = 0.7$, $g = 0.4$, $b = 1.0$, $u^* = .04$, $old-new$ criterion = 1.0, and $\gamma = 0.9$. These parameters were selected because they have been used in a variety of other applications of REM (e.g., Shiffrin & Steyvers, 1997). HR, hit rate; FAR, false alarm rate.

Subjects. Two hundred twenty-five non-Chinese-speaking undergraduate students at Iowa State University participated in exchange for extra course credit.

Design and Materials. Repetitions and study time were manipulated within subjects and within lists, and pair type was manipulated between subjects. Four 24-pair lists were studied, for a total of 96 pairs. The pairs were constructed randomly and anew for each subject. Within each list, 8 pairs were studied once, twice, or six times, with at least 1 intervening pair between repetitions. Four pairs in each repetition condition were studied for 1.5 sec, and 4 pairs were studied for 4 sec. Each test list consisted of 12 intact and 12 rearranged pairs. Intact and rearranged pairs preserved the item orientations from study, and rearranged pairs were formed from items in the same study time and repetition condition.

The subjects studied pairs of words, pseudowords, faces, or Chinese characters. English words were selected from the Francis and Kučera (1982) database, with a normative frequency of between 20 and 50 per million. Pronounceable pseudowords were constructed by changing one letter in each of the words used in the word condition (e.g., *ansle* from *angle*). Chinese characters were selected by the first author to be relatively dissimilar in physical appearance. Black-and-white pictures of faces with standardized head orientation, level of eyes, and position of chin were selected from college yearbooks and the Olivetti Research Database of Faces (A. T. & T., Cambridge, 1994). Forty-eight participants studied pairs of English words, 65 participants studied pseudowords, 50 participants studied pairs of faces, and 65 participants studied pairs of Chinese characters.

Procedure. The subjects were informed that they were to study four lists of pairs of English words, pseudowords, Chinese characters, or faces and that, after each study list, a math task was to be performed. The 20-sec math task consisted of mentally adding a series of digits. The subjects were instructed on how the intact and rearranged pairs were constructed and were told that they were to distinguish the intact pairs from the rearranged pairs by rating, on a scale of 1 to 4, how confident they were that the pair had been studied. Responses of “1” and “2” indicated *high* and *moderate confidence*, respectively, that the pair had been studied, and responses of “4” and “3” indicated *high* and *moderate confidence*, respectively, that the pair had not been studied. To make their response, the sub-

jects were to enter the appropriate number into the computer, using the keyboard.

Results

The mean HRs and FARs and their standard errors for the four pair types, the two study times, and the three levels of repetitions are shown in Figure 4. An alpha of .05 was the standard for reliability.

Effect of pair type on AR. We first computed d_a , a summary measure of accuracy, in order to obtain a comparison of absolute levels of AR performance.⁶ The effect of pair type on d_a was reliable [$F(3,222) = 34.36$]. The mean d_a s for words, faces, pseudowords, and Chinese characters were 1.64, 0.93, 0.85, and 0.47, respectively. AR was best for words ($p < .0005$) and worst for Chinese characters ($p < .0005$). The difference in d_a for pseudowords and faces was not reliable ($p = .47$).

Pair type had reliable and similar-sized effects on HRs [$F(3,224) = 20.11$] and on FARs [$F(3,224) = 20.05$]. HRs were greatest for words ($M = .80$, $p < .0005$) and least for Chinese characters ($M = .62$, $p < .05$). The difference in HRs for pseudowords ($M = .70$) and faces ($M = .67$) was not reliable ($p = .24$). FARs were lowest for words ($M = .22$, $p < .005$) and greatest for Chinese characters ($M = .45$, $p < .0005$). In addition, the difference in FARs for pseudowords ($M = .39$) and faces ($M = .33$) was reliable ($p = .03$), but the difference in FARs for pseudowords and Chinese characters only approached reliability ($p = .06$). Thus, there were mirror effects for words versus pseudowords, faces, and Chinese characters and for pseudowords and faces versus Chinese characters.

Effects of repetitions and study time. HRs increased with the number of target presentations [$F(2,448) = 361.20$] and study time [$F(1,224) = 155.19$]. Repetitions

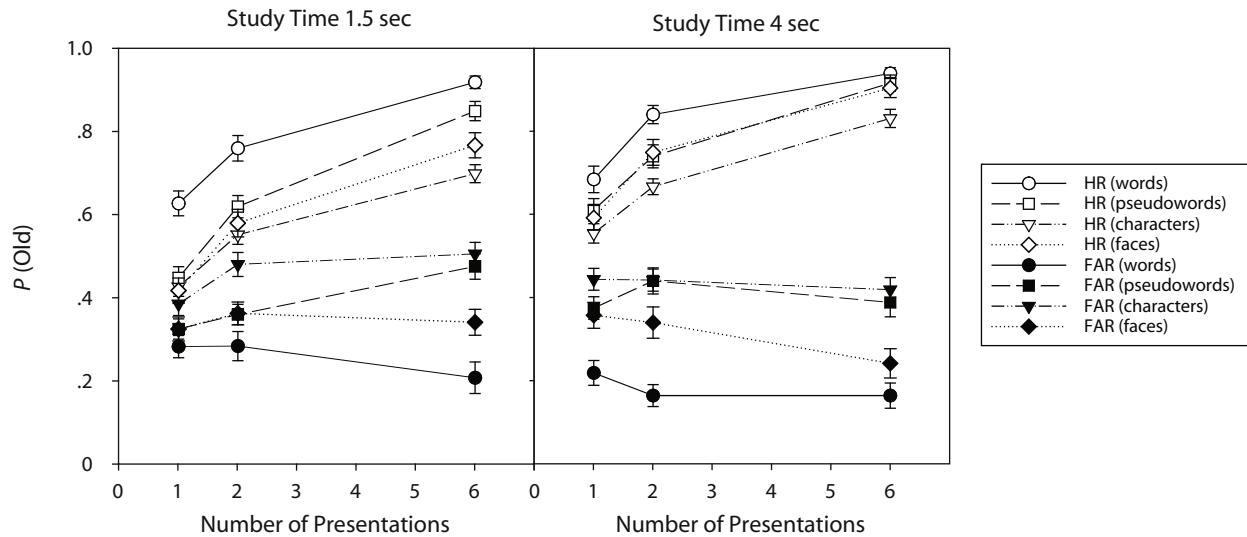


Figure 4. Mean hit rates (HRs) and false alarm rates (FARs) as a function of target presentations, showing the effects of stimulus type, presentations, and study time on associative recognition in the experiment.

did not reliably affect FARs [$F(2, 448) = 1.43$], but FARs decreased with longer study times [$F(1, 224) = 8.46$]. Thus, accuracy increased as a function of repetition and study time. A mirror effect was not observed as target presentations increased, replicating Kelley and Wixted (2001). There was a reliable mirror effect for increases in study time, but it was far smaller (approximately .05, on average) than the ones observed for pair type (approximately .21, on average). However, the interactions of pair type with study time and repetition warrant a word of caution about these main effects.

Interactions between pair type, presentations, and study time. There was no three-way interaction between stimulus type, presentations, and study time on HRs ($F < 1$): HRs always increased with increases in the number of target presentations and study time (all p s < .0005). There was, however, a reliable three-way interaction for FARs [$F(6, 448) = 3.47, p < .002$]. First, consider the interaction between pair type and study time [$F(3, 224) = 3.58, p < .02$]: The effect of study time on FARs is reliable for words [$F(1, 47) = 28.03$], but not for pseudowords, faces, or Chinese characters (all F s < 2.04). For words only, the FARs decreased significantly with increases in study time. Again, however, several simpler interactions between pair type and study time warrant a word of caution. The interaction for words versus faces and study time was not reliable [$F(1, 96) = 3.26, p = .07$], but the interactions for words versus pseudowords [$F(1, 111) = 11.58, p < .001$] and words versus Chinese characters [$F(1, 111) = 4.50, p < .04$] and study time were reliable. In addition, the interaction between pseudowords versus Chinese characters and study time was not reliable [$F(1, 113) = 2.27$]. Thus, the effect of study time on FARs was similar for pairs of words and faces (although stronger for words than for faces) and for pseudowords and Chinese characters.

Next, consider the interaction between repetitions and stimulus type. For words [$F(2, 94) = 3.54, p = .03$] and

faces [$F(2, 98) = 3.32, p < .04$], there was a reliable effect of presentations: FARs decreased slightly as presentations increased. For pseudowords, the effect of repetitions was in the opposite direction [$F(2, 128) = 5.78, p < .005$]: FARs increased as presentations increased. For Chinese characters, the increase in FARs was only marginally reliable [$F(2, 128) = 2.39, p < .10$], increasing by 5% from one to six target presentations. Thus, the interaction effect on FARs between standard strengthening operations and pair type was variable and depended on the nature of the stimuli.

Confidence ratings. To measure the use of rating scale independently of bias or accuracy, we used the probability of a rating (i.e., high or moderate confidence), given a “yes” (i.e., HR and FAR) or a “no” (i.e., correct rejection and miss) response. If R_H and R_M are the probabilities of high and moderate confidence ratings given a *yes* or a *no* response, then $R_H + R_M = 1.0$. Thus, R_H divided by the HR in a given condition is the tendency to use the high-confidence *yes* rating independently of the actual HR. We obtained similar scores for false alarms, correct rejections, and misses, and these are shown in Figure 8.

For correct responses, the tendency to use the high-confidence rating increased with presentation [hits, $F(2, 440) = 166.8$; correct rejections, $F(2, 428) = 82.6$] and study time [hits, $F(1, 220) = 97.0$; correct rejections, $F(1, 214) = 39.0$] for all pair types. In addition, the high-confidence tendency was reliably affected by pair type [hits, $F(3, 220) = 30.1$; correct rejections, $F(3, 214) = 24.5$]: It was strongest for words and least for Chinese characters. For false alarms, the tendency to use the high-confidence *old* rating increased with increases in presentations [$F(2, 276) = 32.2$] and study time [$F(1, 138) = 30.9$]. This tendency was highest for words and least for Chinese characters [$F(3, 138) = 2.50, p = .06$]. As compared with the other types of responses, the use of the ratings scale was little affected by repetitions [$F(1, 72) = 0.62$], study

time [$F(2,144) = 0.59$], or pair type [$F(3,72) = 0.23$] when the response was a miss.

Discussion: Model Analysis

AR was best for words and worst for Chinese characters, and AR accuracies for pseudowords and faces were at a similar, intermediate level. Importantly, the pair type manipulations produced robust mirror effects: HRs increased and FARs decreased as performance decreased for different pair types. In addition, HRs always increased with increases in the number of target presentations and study time. However, strength manipulations affected FARs for the four types of stimuli differently: FARs decreased with longer study time for words, but not for the other stimulus types, and increases in target repetitions decreased FARs for words and faces but increased FARs for pseudowords and Chinese characters. Thus, there were large mirror effects between pair types but small and inconsistent repetition and study time effects on FARs. Importantly, the patterns of FARs for words and faces were similar, and the patterns of FARs for pseudowords and Chinese characters were similar, which is consistent with the hypothesis that the frequency with which different types of stimuli are encountered in everyday life determines the contribution of recollection to AR.

We considered three hypotheses to explain these findings: Different stimuli vary in the extent to which they are encoded (i.e., the *encoding hypothesis*), the extent to which recollection contributes to the discrimination between intact and rearranged pairs (i.e., the *retrieval hypothesis*), or both (i.e., the *joint hypothesis*).⁷ We conducted a series of Monte Carlo simulations designed to identify the variant of the AR model that provides the best account for the differences in accuracy due to stimulus type.

To find a *best fit* for the encoding and retrieval hypotheses, u^* and a , respectively, were allowed to vary between pair types. To find a best fit for the joint hypothesis, both u^* and a were allowed to vary between pair types in the same simulation. We considered the values of u^* and a listed in Figures 2 and 3. They provide a reasonable coverage of the parameter space, while ignoring clearly inappropriate values that lead performance to be either at floor or at ceiling, as identified in pilot simulations. The remaining parameters were fixed between stimulus types to the values listed in Figures 2 and 3. For the goodness-of-fit measure, the *mean squared difference* (MSD) between the data and the model was computed for each stimulus type. Table 1 lists, for each model, the least MSD and the parameter value(s) of interest that produced it.

Encoding hypothesis. According to this model, the pair type manipulation influences the amount of information about a pair that is stored when studied. The encoding parameter u^* was allowed to vary between .01 and .06, holding all other parameters constant. Figure 5 shows the best fit of the encoding model. The least MSD for this model across all conditions is .009. The u^* value was .044 for words, .023 for faces, .021 for pseudowords, and .014 for Chinese characters. Thus, words tended to be encoded best and Chinese characters worst, with the encoding of

Table 1
Parameter Values and Mean Squared Differences (MSDs) for Three Models

Stimulus Type	Model Assumptions						
	Encoding Hypothesis		Retrieval Hypothesis		Joint Hypothesis		
	u^*	MSD	a	MSD	u^*	a	MSD
Words	.044	.001	1.00	.002	.042	.99	.001
Pseudowords	.021	.011	.20	.004	.027	.24	.002
Faces	.023	.004	.48	.009	.023	.76	.004
Characters	.014	.020	.10	.013	.024	.09	.008
Overall fit		.009		.007			.004

pseudowords and faces falling in the middle. The trend in degree of encoding (u^*) is similar to the pattern of HRs observed in the data. HRs were greatest for words and least for Chinese characters, with HRs for faces and pseudowords falling at an intermediate level. However, the encoding model does not predict the FAR component of the pair type mirror effect. An inspection of Figure 5 shows that the model predicts very little, if any, effect of pair type on FARs and, thus, can be rejected on qualitative grounds.

Retrieval hypothesis. According to this model, only the contribution of recollection to performance is influenced by pair type. Accordingly, the retrieval parameter a was allowed to vary between .1 and 1.0, holding all other parameters constant. The fit of the retrieval model that produces the least MSD is shown in Figure 6. The least MSD for this model across all stimuli was .007. The value of a was 1.0 for words, .48 for faces, .2 for pseudowords, and .1 for Chinese characters. Thus, the variability in the contribution of recollection (a) to performance corresponds to the variability in d_a in the prior section.

In addition, FARs decreased as the number of target presentations increased for words and faces. For pseudowords, the number of target presentations had little effect, and the FARs increased for Chinese characters. Thus, the a parameter corresponded to the levels of FARs between pair types. The FARs for words were the lowest, and the FARs for Chinese characters were the highest. Note, however, that the retrieval model predicts very little change in HRs between pair types. Thus, this model cannot account for the pair type mirror effect.

Joint hypothesis. Comparing the fits of the encoding and the retrieval hypotheses shows that each model has advantages and disadvantages over the other model. These analyses show that both the encoding and the retrieval models can provide accurate qualitative accounts of our findings in terms of a composite measure of accuracy (e.g., d_a). That is, the model can predict that some pair types are better recognized than others. The main problem, however, is that neither model predicts a pair type mirror effect. The best fit of the encoding model underpredicts the FARs by as much as 20% in many cases. This is especially true for pseudowords and Chinese characters. In contrast, the retrieval model accounts for the pair type effects on FARs better than it does those on HRs. The retrieval model overpredicts the HRs for nonverbal stimuli by as much as 21%.

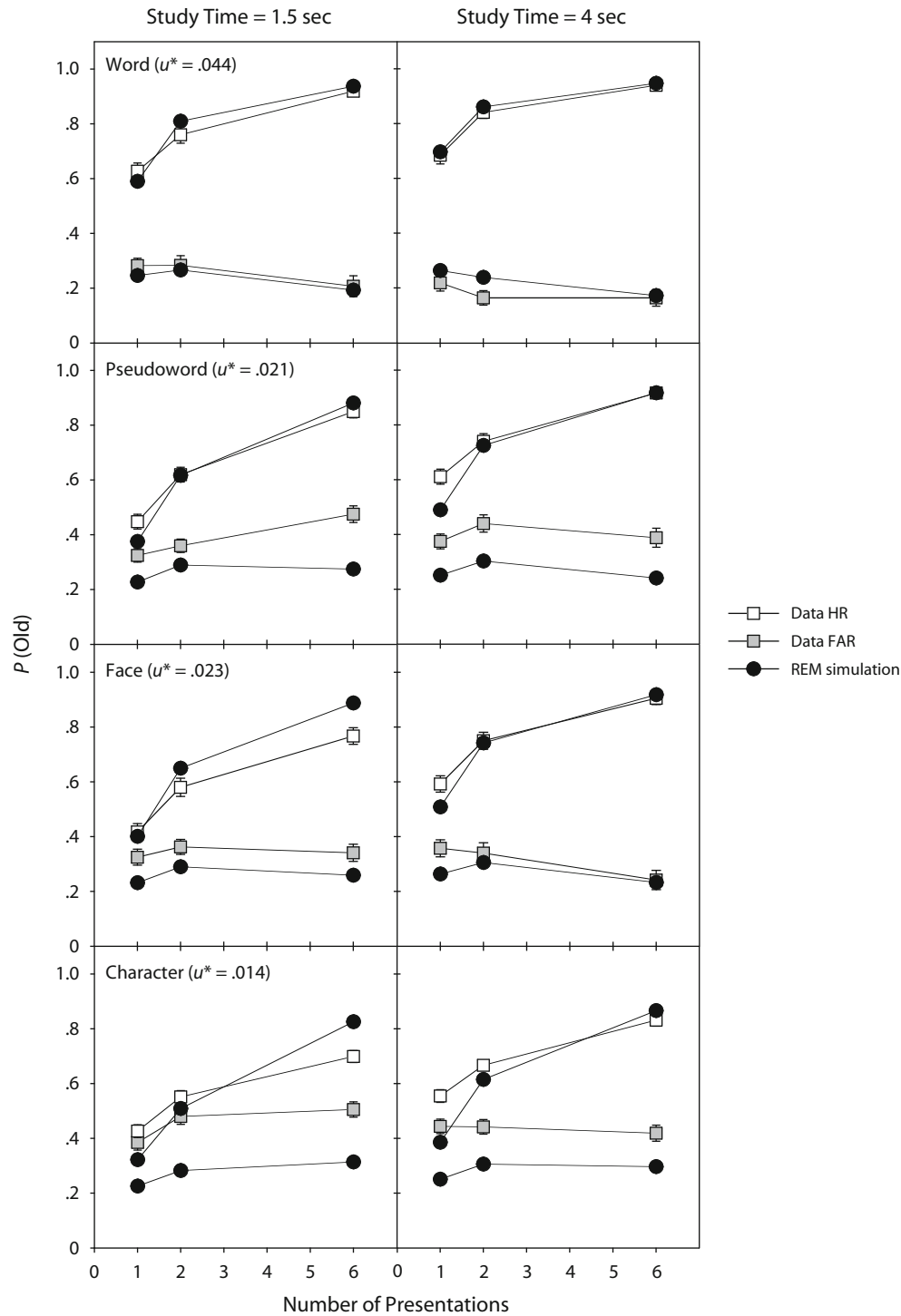


Figure 5. Best fits for the encoding model and the u^* parameters that generated them. Other parameters were fixed as $a = 1.0$, $w = 20$, $t_1 = 7$, $c = .7$, $g = .4$, $b = 1.0$, *old-new* criterion = 1.0, and $\gamma = 0.9$. HR, hit rate; FAR, false alarm rate; REM, retrieving effectively from memory.

Figure 7 shows the joint-model fit that minimizes its difference with the data. The encoding parameter, u^* , is much greater for words (.042) than for the nonverbal stimuli (faces, .027; pseudowords, .023; Chinese characters,

.024). Thus, the joint model suggests that more features are stored in a unit of time for words than for nonverbal stimuli, perhaps because there are more features available for storage. The model also suggests that there is a much

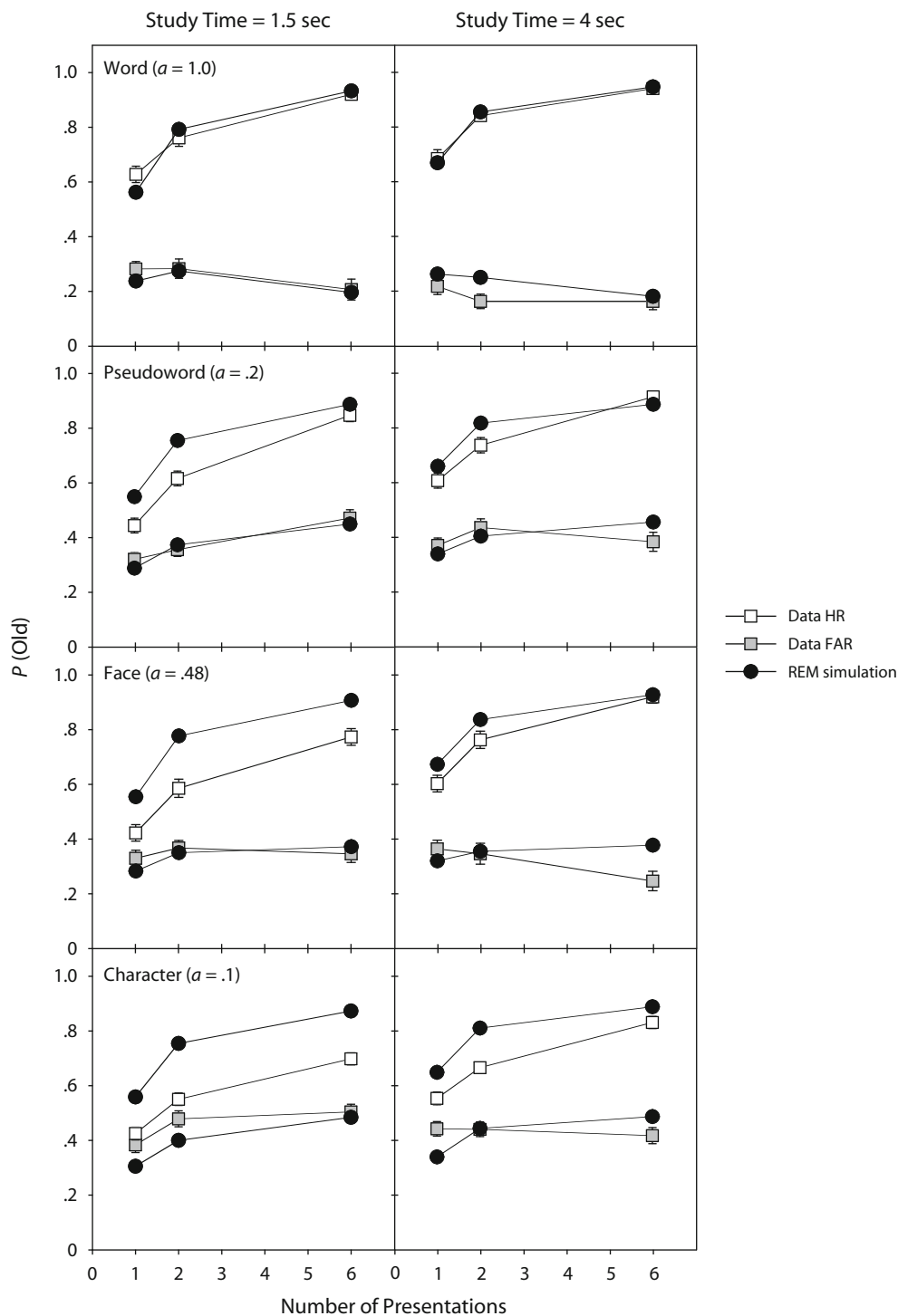


Figure 6. Best fits for the retrieval model and the a parameters that generated them. Other parameters were fixed as $u^* = .04$, $w = 20$, $t_1 = 7$, $c = .7$, $g = .4$, $b = 1.0$, *old-new* criterion = 1.0, and $\gamma = 0.9$. HR, hit rate; FAR, false alarm rate; REM, retrieving effectively from memory.

lower contribution of recollection to performance for the uncommon stimuli than for stimuli that are processed routinely in everyday life. The retrieval parameter, a , is the greatest for words (.95) and somewhat less for novel faces

(.76); a is much less for pseudowords (.24) and Chinese characters (.09).

Importantly, the pair type mirror effect is predicted by the joint model: Both the HRs and the FARs vary as a

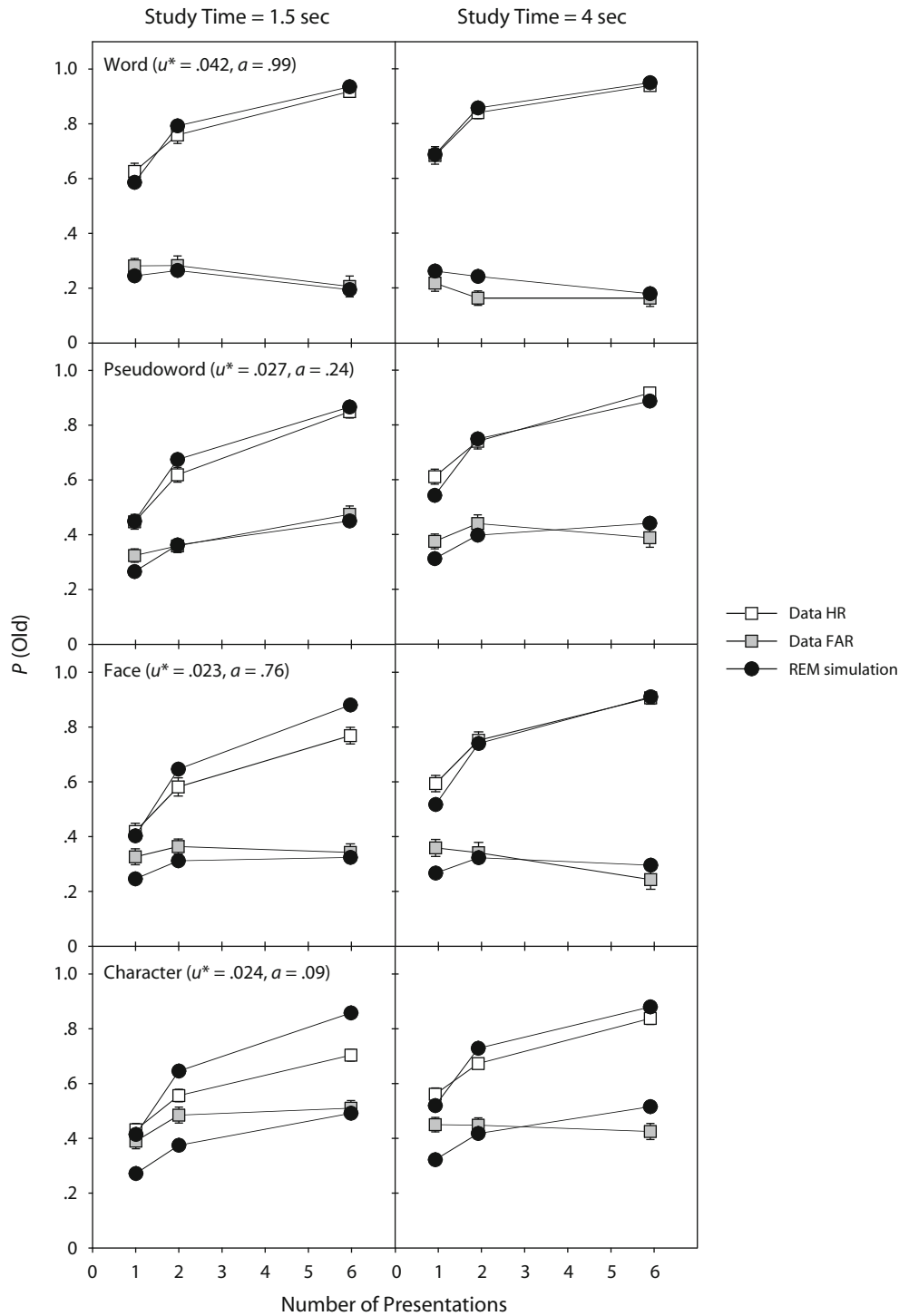


Figure 7. Best fits for the joint model and the a and u^* parameters that generated them. Other parameters were fixed as $w = 20, t_1 = 7, c = .7, g = .4, b = 1.0, \text{old-new criterion} = 1.0$, and $\gamma = 0.9$. HR, hit rate; FAR, false alarm rate; REM, retrieving effectively from memory.

function of pair type. The joint model also provides a better quantitative fit (least MSD = .004) to the data than does either the encoding or the retrieval model. On average, there was a 56% and 44% reduction in the average

deviation of the best fit from the data when the joint model was used, as compared with when the encoding and the retrieval models, respectively, were used. Thus, the joint model captures the qualitative effects of pair type and pro-

vides a more accurate quantitative fit than does either the encoding or the retrieval model.

It is worth pointing out that the joint model captures almost all of the data with relatively few free parameters. There are 48 data points and only 8 free parameters in the fit of the joint model. There are four levels of a and four levels of u^* . All other parameters were fixed to the values used in pilot simulations (Figures 2 and 3). There are, of course, more free parameters in the joint model than in the encoding or the retrieval model. However, both of these models can be rejected on a qualitative basis, because they do not predict the pair type mirror effect.

GENERAL DISCUSSION

In this section, we first will discuss in broad terms the present theoretical framework and how it accounts for the present findings. We then will discuss the effect of pair type on bias, the implications of the present findings for other models of recognition memory, and possible extensions of the accuracy model to the use of confidence ratings. Last, we will discuss the effect of pair type on AR.

Modeling Recognition Memory: Mirror Effects Versus Increases in Hit Rates

The discovery of the mirror effect for single-item recognition instigated profound changes in theories of memory. Prior to Glanzer and Adams's (1985) discovery of the mirror effect, theories of memory did an adequate job at a gross level of measurement in predicting changes in single-item recognition accuracy. However, Glanzer and Adams (1985) showed that variables that produce simultaneous increases in HRs and decreases in FARs required a new class of models, and the *likelihood* class of global familiarity models became dominant (Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).

An equally influential discovery is the finding that HRs sometimes increase but FARs remain steady when targets and foils cannot be discriminated on the basis of random factors. Such findings include those from the process dissociation procedure (see Jacoby, 1998, for a review), the registration-without-learning procedure (Hintzman et al., 1992; Malmberg, Holden, & Shiffrin, 2004), and AR (Cleary, Curren, & Greene, 2001; Kelley & Wixted, 2001). The universal conclusion of these researchers was that at least two different sources of information are required in order to explain the patterns of HRs and FARs observed in the different paradigms.

The REM theory explains associative recognition and single-item recognition within a global memory framework (Malmberg, Holden, & Shiffrin, 2004; Malmberg, Zeelenberg, & Shiffrin, 2004; Shiffrin & Steyvers, 1997). There are two sources of information: One is a continuous random variable (familiarity), and one is a discrete random variable (the presence vs. the absence of episodic details). It considers recognition memory to be a flexible system that adapts to task, situational, and subjective demands (i.e., costs and rewards): Subjects adopt a recognition strategy that maximizes the trade-off between speed

and accuracy by utilizing different sources of information. In this sense, the recognition strategy can be seen to be efficient, effective, or optimal for a given situation (cf. Malmberg & Xu, 2007).

When targets and foils are randomly similar, mirror effects are predicted on the assumption that subjects do not utilize a recall-to-reject strategy (Malmberg & Murnane, 2002; Malmberg, Zeelenberg, & Shiffrin, 2004; Shiffrin & Steyvers, 1997). Targets tend to be much more familiar than foils, and thus, familiarity is a relatively accurate basis for making the single-item recognition decision. Familiarity is also assumed to become available earlier during retrieval than does recollective information (Doshier, 1984; Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; Light et al., 2004), and hence, the familiarity-only strategy is preferred for single-item recognition, because it minimizes the amount of time taken to accurately perform the task.

For AR and plurality discrimination, targets and foils are not randomly similar. In these cases, both targets and foils tend to be relatively familiar (although targets are more familiar than foils, on average), and hence, the familiarity strategy is relatively inaccurate. However, recollective details may indicate that targets were studied and foils were not, and the subjects rely on a slower but more accurate recall-to-reject strategy in order to more effectively reject otherwise familiar foils. The distinction between single-item plurality discrimination and AR is that targets and foils are less similar in AR. If so, the recall-to-reject strategy should be more effective for AR than is plurality discrimination, and it might be utilized more often if it is also less effortful.

The degree to which recollection contributes to AR performance primarily determines the patterns of FARs that will be observed. This is captured by the a parameter in the present model. We have assumed that a might be influenced by strategies and/or structural factors. However, the present model cannot discriminate between these factors. To do so, the sampling and recovery process of the SAM/REM framework need to be fully implemented, which we will leave for future research. In the present study, a was influenced by pair type, which suggests that the sampling and recovery process might be slower for uncommon stimuli; hence, the recall-to-reject strategy is used relatively infrequently as a basis for the AR decision.

Pair Type Effects and Response Bias

The joint model provides a qualitative and quantitative account of the interaction between repetitions, study time, and stimulus type. The fits are, however, imperfect. Although some departures of the model from the data are almost certainly due to influences of random factors on the data, there is one systematic deviation: Figures 2 and 3 show that changes in a or u^* do not affect FARs in the one presentation condition as much as is observed in the data (see Figure 7). As we have noted, differences in FARs are due primarily to differences in the contribution of recollection to AR. When a pair has been presented only once, relatively fewer features are stored in the joint model, and according to Equation 3, the number of correctly stored

features limits (c and u^*) the recollective contribution. Hence, variations in the recall parameter (a) have little effect on FARs in the one presentation condition but have much greater effects as items are repeated more often.

We considered several ways that FARs might be affected in the one presentation condition. One augmentation to the joint model is suggested by a recent set of nine experiments that indicates that subjects are more biased to respond *old* to nonverbal stimuli than to verbal stimuli (Greene, 2004). If this was the case in the present experiment, it could have produced the large pair type effects on FARs in the one presentation condition. Indeed, the present subjects adopted a more lenient criterion (C_a), on average, for pseudowords and characters than for faces or words. The mean C_a for faces, words, Chinese characters, and pseudowords was .02 ($SE = .037$), $-.05$ ($SE = .058$), $-.08$ ($SE = .032$), and $-.14$ ($SE = .040$), respectively. A post hoc t test showed that the mean C_a for faces was not significantly different from that for words [$t(95) = 1.05$, $p = .30$] but was significantly higher than those for Chinese characters and pseudowords [$t(112) = 2.19$, $p < .05$, and $t(113) = 2.86$, $p < .005$, respectively].

Because the bias effect occurs for both pseudowords and Chinese characters, relative to faces, it is probably not caused by a verbal–nonverbal distinction. Nor does the bias appear to be caused by differences in stimulus novelty, since there was a bias to respond *old* to novel letter strings, relative to novel faces. Perhaps, however, there is a bias to respond *old* to stimuli that are not processed frequently in everyday life (e.g., pseudowords and Chinese characters vs. words and faces).

Armed with this knowledge and hypothesis, we investigated three ways to vary bias in the present model: shift the initial criterion, vary the guessing parameter, or both, between all pair types. The results showed that varying bias in the joint model improved the overall fit of the joint model, particularly in the one-presentation condition. Keeping this in mind, we nevertheless chose not to further explore the biased joint model, because we could not distinguish between the different types of bias and because it already captured the major effects on AR accuracy.

Implications for Models of Associative Recognition

The present findings did not replicate Kelley and Wixted's (2001) null effect of word repetitions on FARs. When pairs consisted of words or faces, FARs decreased in a reliable fashion as repetitions increased, but only slightly. In contrast, FARs increased as repetitions increased when pairs consisted of pseudowords or Chinese characters. Thus, the function relating repetitions to FARs was variable and depended on the nature of the stimuli.

These findings disconfirm models that predict a null effect of repetitions on FARs. According to strong versions of independent cue models, item information and associative information are separately and independently represented during study. At test, only associative information is used to probe memory. Since cues formed from rearranged pairs are only randomly similar to the contents of memory, there should be no effect of repetitions on FARs.

Weaker versions of the independent cue model assume that associative recognition is based on a combination of item familiarity and associative familiarity. For instance, Kelley and Wixted's (2001) some-or-none model assumes that item and associative familiarity sum for intact pairs to provide the evidence on which to make a decision. For rearranged pairs, associative familiarity is subtracted from item familiarity. Importantly, in some cases, only item familiarity is available. Hence, the some-or-none model could generate a number of different functions relating FARs to repetitions, depending on the degree to which associative information influences recognition.

Nevertheless, the some-or-none model is limited in several ways. First, it is not amenable to accounting for the performance of other recognition tasks in which foils are similar to targets—particularly, plurality discrimination (Hintzman et al., 1992) and exemplar discrimination (Morrell et al., 2002). Second, it is a measurement model; it provides no explanation for how items are encoded, represented, or retrieved, and hence, it would have to be extended to explain how different stimuli affect the balance of item and associative familiarity that gives rise to the form relating repetitions to FARs. To say that the contribution of associative information varies between pair types or as function of repetitions is simply a description of the data. Last, the model is somewhat circular, in that in order to determine whether to add associative familiarity to item familiarity or to subtract associative familiarity from item familiarity, the model needs to know whether an intact or a rearranged pair is being tested, which is, after all, the task to begin with.

The variable pattern of FARs disconfirms models that predict that FARs will decrease as the number of target presentations increases, such as single-process recall-only models (Yonelinas, 1997). For all types of stimuli, HRs increased with increases in repetitions, presumably as the result of increasing the probability of encoding the pairs. If so, increasing repetitions should have increased the probability of recalling that two items were not studied together when rearranged pairs were tested. However, this occurred only for words and faces, and not for pseudowords or Chinese characters; thus, recall-only models can be rejected as general models of associative recognition.

Confidence Ratings

The present findings equally raise problems for dual-process models that assume that the highest confidence ratings are reserved for instances in which recollection is the basis of the *old–new* judgment. For instance, the dual-process model described by Yonelinas (1997) combines a single high-threshold recollection model with an equal-variance signal detection model of familiarity. Accordingly, only familiarity can be the basis for an *old* decision when memory is probed with a foil. Hence, the highest confidence *old* response should never be used when a false alarm is made.

Consider Figure 8, which plots the tendency to use high-confidence responses as a function of the number of target presentations for the different pair types. For all pair types, the tendency to use the high-confidence rating

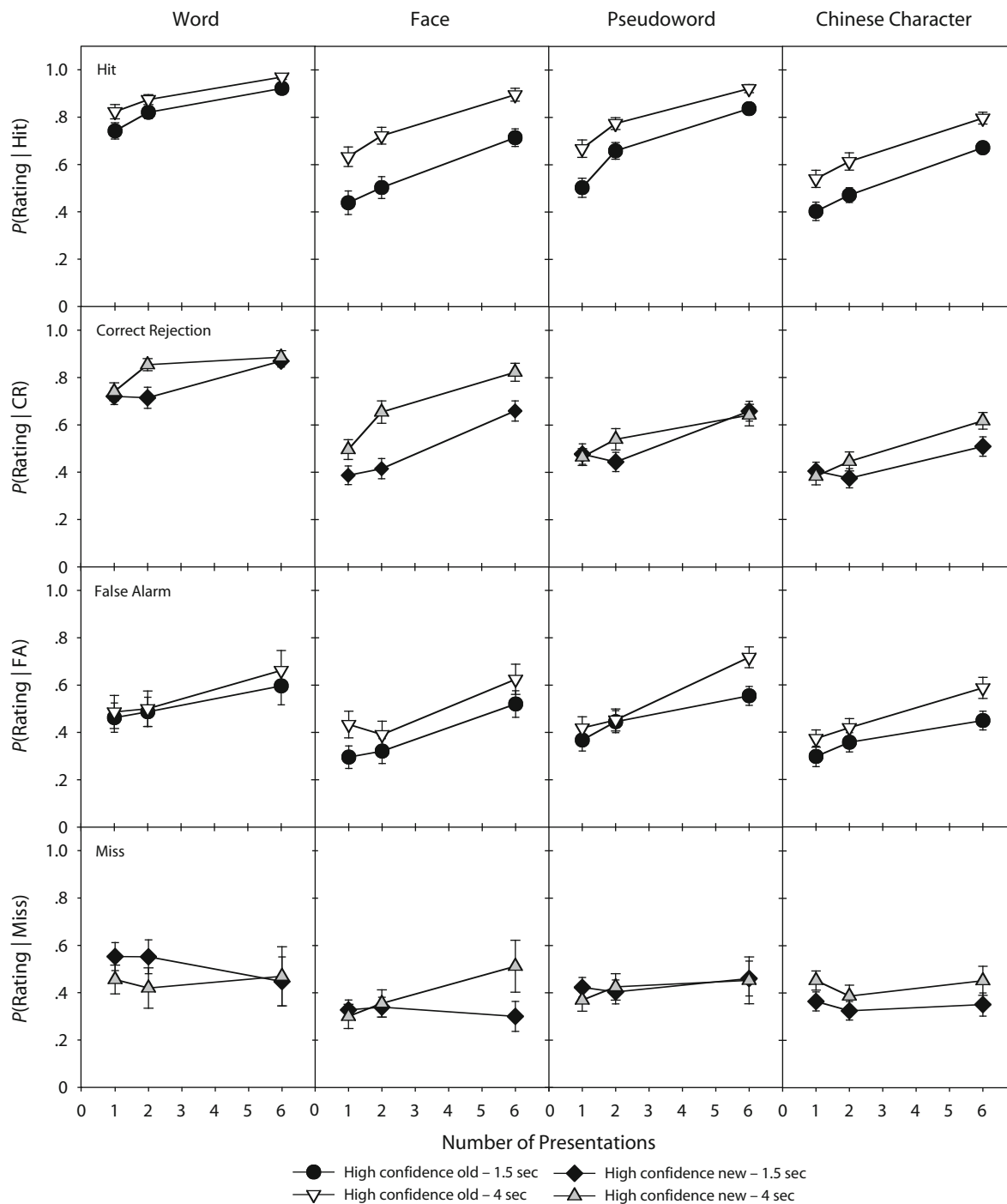


Figure 8. The tendency to use high-confidence ratings as a function of presentations, study time, and pair type. CR, correct rejection; FA, false alarm.

increases when the response is a hit or a correct rejection and might lead one to conclude that this reflects only the contribution of recollection to performance. Also note, however, that the tendency to use the high-confidence *old* response increases with repetitions when the response is a false alarm. This finding is inconsistent with the assumption that the high-confidence *old* rating is reserved for when recollection is the basis of the *old*-*new* judgment.

One might be tempted to assume that the use of a high-confidence *old* rating when a rearranged pair is tested is the result of *false recollection*. However, subjects used the high-confidence *old* response almost 50% of the time when they made false alarms (on average). It seems unreasonable to conclude that false recollection occurs this often, especially since overall accuracy is very high, particularly in the word condition. Rather, the data suggest

that highly confident responses reflect a mixture of responses based on recollection and familiarity.

A formal model of ratings is beyond the present scope, but the ratings analyses do suggest possible extensions of the REM dual-process model. According to the joint model, recollection is always veridical, and *old* responses are based on either recollection or guessing. Why, then, does the tendency to use the high-confidence *old* rating increase when a false alarm is made? One possibility is that after an *old–new* decision is made on the basis of a guess, familiarity informs the ratings decision, with a high-confidence rating given to highly familiar pairs and a moderate confidence rating given to less familiar pairs. Our modeling analyses are consistent with this hypothesis, since the use of the ratings scale mirrors the effect of differences in encoding. According to Equations 2 (familiarity) and 3 (the probability of an *old–new* decision based on recollection), differences in the extent of encoding (u^*) affect familiarity in the model, as well as the use of a recall-to-reject strategy. Figure 7 shows that u^* varies between pair types, and thus, increases in the high-confidence responses might reflect increases in pair familiarity, increases in recollective *yes–no* responses, or both.

Another possibility is that in the absence of recollection, the *yes–no* decision itself is based on the familiarity of the pair, after which ratings are made as described above. We conducted a series of simulations to ascertain whether such an augmentation could account for the effects of repetitions, study time, and pair type on the accuracy of AR. Note that the present model assumes that in the absence of recollection, the subject is highly biased to respond *old*. If one assumes that subjects always answer *old* when the initial criterion is exceeded and recollection fails, this is tantamount to assuming that familiarity is the basis for the *old–new* decision. With this in mind, it is not surprising that replacing the guessing assumption with the familiarity assumption provides a good fit to the data.

Last, consider how the ratings scale is used when the response is a miss. As compared with the other types of responses, the use of the ratings scale is little affected by repetitions, study time, or pair type when the response is a miss. According to the present model, *no* responses are made because the familiarity of the pair fails to exceed the initial criteria, because the pair is familiar and recollection indicates that the items were not studied together, or because recollection fails and the subject guesses *no*. Thus, increasing familiarity via repetitions and study time should change responses from misses to hits, and hence, the effect of repetitions on the tendency to use the high-confidence rating should be observed in hits, rather than in misses.

On the Effects of Novelty, Lexicality, and Familiarity

A number of findings suggest that the occurrences of verbal stimuli are easier to recollect than the occurrences of nonverbal stimuli. For instance, remember–know recognition requires subjects to introspectively decide the basis of a *yes* response (see Yonelinas, 2002, for a review).

A *remember* response is often assumed to be based on recollection of the study event, whereas a *know* response is often assumed to be based on familiarity (but see Dunn, 2004; Malmberg, Zeelenberg, & Shiffrin, 2004; Wixted & Stretch, 2004). The typical finding is that words tend to elicit more *remember* responses than do nonwords (Curran et al., 1997; Gardiner & Java, 1990). In addition, electrophysiological experiments have shown a larger P600 ERP signal when words versus nonwords are recognized, further suggesting that words are more likely to be recognized on the basis of recollection than are nonwords (e.g., Curran, 1999). The present model analysis suggests a similar conclusion for associative recognition. Within the dual-process framework, we observed a greater tendency to base *yes–no* responses on the outcome of the sampling and recovery process when word pairs were tested than when pseudoword pairs were tested.

Although our conclusion is similar to those based on remember–know results, the basis for our conclusion is different. Consider that *remember* responses follow positive recognition judgments, leading to either hits or false alarms. The proportion of *remember* responses following a hit is used to derive an index of recollection, whereas the proportion of *remember* responses made following a false alarm is considered an index of *noise* (Yonelinas, 2002).

In contrast to these assumptions, the dual-process REM framework assumes that hit rates offer little insight into the basis of the *yes–no* response, because both recollection and high levels of familiarity can contribute to a *yes* response. To the extent that the ratings task and the remember–know task are accounted for by similar models (cf. Yonelinas, 2002), moreover, *remember* responses most likely reflect a mixture of familiarity- and recollection-based responses, and we have shown that use of the high-confidence *old* response increases for rearranged pairs. Indeed, the function relating repetitions to FARs is relatively flat or even decreases, whereas the FAR function relating the high-confidence *old* response to repetitions increases, which suggests that there are different types of information used to make the *old–new* and the remember–know judgments (cf. Hintzman et al., 1992; Malmberg, Holden, & Shiffrin, 2004). Thus, the FAR function, but not the HR function, provides insight into the relative contribution of recollection and familiarity.

Last, the present findings call into question the assumption that the critical difference between associative recognition of words and pseudowords is lexicality. On the basis of the model analysis, we found that the recollective component of recognition was relatively high for words and novel faces and relatively low for pseudowords and Chinese characters. Indeed, the function relating the number of target presentations to FARs decreased for words and faces, whereas they increased for pseudowords and Chinese characters. From these observations, it seems that neither lexicality nor stimulus novelty is the factor that determines the contribution of recollection to associative recognition. It is, however, plausible to conclude that recollection is more often the basis for an associative recognition decision for items that are frequently encoun-

tered (e.g., novel faces and words) versus items that are not frequently encountered (e.g., pseudowords and Chinese characters).

Randomly Similar Foils

We assume that a single-process familiarity-only model is used when targets and foils are randomly similar (Malmberg, Holden, & Shiffrin, 2004; Malmberg, Zeelenberg, & Shiffrin, 2004; Shiffrin & Steyvers, 1997). In some experiments, however, randomly similar foils are tested along with foils that are similar to the targets. For instance, Kelley and Wixted (2001) and Malmberg and Xu (2007) found relatively low FARs for completely new pairs of words (*XY* pairs) for associative recognition, as is the case for plurality discrimination (Hintzman et al., 1992; Malmberg, Holden, & Shiffrin, 2004). Malmberg, Holden, and Shiffrin showed that the dual-process model predicts relatively low FARs for completely new stimuli and that the model fits the data well because it assumes that the familiarity of completely new pairs is relatively low and, hence, they are usually rejected (see Figure 1).

In addition, it is important to note that randomly similar foils and *XY* pairs were not studied; hence, it seems peculiar to assume that they can be rejected on the basis of the recall-to-reject strategy. Nevertheless, it is interesting to consider the performance of the model when the contribution of recollection to performance is relatively low. FARs are much lower than the HRs when the contribution to recollection is very low. In this instance, recollection is rarely the basis for a response (e.g., when $a = .2$ in Figure 2, q varies from $<1\%$ to $\sim 4.3\%$). Rather, the subjects are assumed to guess *old* 90% of time, which is pretty much what would happen if they always made their decisions on the basis of familiarity. Thus, when $a = .2$, FARs are much less than HRs, mostly because the odds associated with rearranged pairs exceed the *yes-no* criterion far less often than do those for intact pairs. This is because rearranged pairs overlap with targets only in about 50% of their features, mismatched features lower the odds, and increased repetitions or study time increase the number of mismatches, as well as the number of matches.

We have assumed that recognition memory is a flexible system. When targets and foils are randomly similar, a single-process familiarity strategy is used because targets are generally much more familiar than foils, and familiarity is available at test sooner than are the details provided by a recollection process. Hence, subjects can perform the recognition task relatively well in a relatively short period of time. On the other hand, when targets and foils are similar, the foils seem relatively familiar. Hence, a recall-to-reject strategy is used, even though it takes longer to perform.

These assumptions suggest that when *XY* pairs are added to the associative recognition test list, subjects will rely on the faster familiarity-only strategy to a greater extent than they will when only rearranged pairs are tested. If so, the predictions are that (1) the FARs for *XY* pairs will be less than the FARs for rearranged pairs, (2) the FARs for rearranged pairs will be greater when *XY* pairs are also tested than when only rearranged pairs are tested, and (3) response

latencies will be shorter when *XY* pairs are tested. All three of the predictions were recently confirmed in a series of experiments conducted by Malmberg and Xu (2007).

CONCLUSIONS

Increases in study time and repetitions have variable effects on FARs that are dependent on the nature of the stimuli, and thus, they improve associative recognition primarily by increasing HRs. For pairs of items that are encountered in everyday life, such as words and novel faces, strengthening operations decrease FARs. For pairs of items that are not encountered in everyday life, such as pseudowords and Chinese characters, strengthening operations increase FARs. In addition, pair types that are better recognized have substantially higher HRs and lower FARs than do pair types that are recognized less well. The REM dual-process model suggests that the pair type mirror effect is caused by the enhanced encoding and a greater contribution of recollection to the recognition of better recognized pair types. Encoding advantages primarily produce greater HRs, and increases in the contribution of recollection produce lower FARs. The recognition advantage of words over pseudowords is not due to the lexical status of the stimuli. Rather, pairs of items encountered every day are more likely to be recognized on the basis of recollective evidence than are pairs of items infrequently encountered.

AUTHOR NOTE

This research partially fulfilled the master's degree requirements for J.X. at Iowa State University. We thank Veronica Dark and Vassant Hanovar for their participation on J.X.'s master's degree committee. Some of this research was presented at the 45th Annual Meeting of the Psychonomic Society in Minneapolis, the 30th Annual Interdisciplinary Conference in Teton Village, WY, and the 3rd Context and Episodic Memory Symposium in Philadelphia. Correspondence concerning this article should be addressed to K. J. Malmberg, Department of Psychology, University of South Florida, 202 East Fowler Ave., PCD 4118G, Tampa, FL 33620 (e-mail: malmberg@cas.usf.edu).

REFERENCES

- AT&T CAMBRIDGE (1994). *The Olivetti research database of faces* [Electronic database]. Available at www.cam-orl.co.uk/facedatabase.html.
- ATKINSON, R. C., & JUOLA, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory, and thinking* (pp. 243-293). San Francisco: Freeman.
- 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.
- CLARK, S. E., & SHIFFRIN, R. M. (1992). Cueing effects and associative information in recognition memory. *Memory & Cognition*, *20*, 580-598.
- CLEARY, A. M., CURRAN, T., & GREENE, R. L. (2001). Memory for detail in item versus associative recognition. *Memory & Cognition*, *29*, 413-423.
- CRISS, A. H., & SHIFFRIN, R. M. (2004). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, *32*, 1284-1297.
- CURRAN, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia*, *37*, 771-785.

- CURRAN, T., SCHACTER, D. L., NORMAN, K. A., & GALLUCCIO, L. (1997). False recognition after a right frontal lobe infarction: Memory for general and specific information. *Neuropsychologia*, **35**, 1035-1049.
- DENNIS, S., & HUMPHREYS, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, **108**, 452-478.
- DILLER, D. E., NOBEL, P. A., & SHIFFRIN, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 414-435.
- DOSHER, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed-accuracy study. *Cognitive Psychology*, **16**, 519-555.
- DUNN, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, **111**, 524-542.
- FRANCIS, W. N., & KUČERA, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
- GARDINER, J. M., & JAVA, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, **18**, 23-30.
- GLANZER, M., & ADAMS, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, **13**, 8-20.
- GLANZER, M., & ADAMS, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 5-16.
- GREENE, R. L. (1996). Mirror effect in order and associative information: Role of response strategies. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 687-695.
- GREENE, R. L. (2004). Recognition memory for pseudowords. *Journal of Memory & Language*, **50**, 259-267.
- GRONLUND, S. E., & RATCLIFF, R. (1989). Time course of item and associative information: Implications of global memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 846-858.
- HEALY, M. R., LIGHT, L. L., & CHUNG, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 768-788.
- HINTZMAN, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.
- HINTZMAN, D. L., & CURRAN, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory & Language*, **33**, 1-18.
- HINTZMAN, D. L., CURRAN, T., & OPPY, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 667-680.
- HOCKLEY, W. E. (1992). Item versus associative information: Further comparisons of forgetting rates. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 1321-1330.
- HOCKLEY, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition*, **22**, 713-722.
- JACOBY, L. L. (1998). Invariance in automatic influences of memory: Toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 3-26.
- JACOBY, L. L., & DALLAS, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, **110**, 306-340.
- KELLEY, R., & WIXTED, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 701-722.
- LIGHT, L. L., PATTERSON, M. M., CHUNG, C., & HEALY, M. R. (2004). Effects of repetition and response deadline on associative recognition in young and older adults. *Memory & Cognition*, **32**, 1182-1193.
- MALMBERG, K. J., HOLDEN, J. E., & SHIFFRIN, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 319-331.
- MALMBERG, K. J., & MURNANE, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 616-630.
- MALMBERG, K. J., & SHIFFRIN, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 322-336.
- MALMBERG, K. J., & XU, J. (2007). On the flexibility and fallibility of associative memory. *Memory & Cognition*, **35**, 545-556.
- MALMBERG, K. J., ZEELENBERG, R., & SHIFFRIN, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 540-549.
- 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.
- MORRELL, H. E., GAITAN, S., & WIXTED, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 1095-1110.
- MURDOCK, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, **104**, 839-862.
- NAVEH-BENJAMIN, M., CRAIK, F. I. M., & BEN-SHAUL, L. (2002). Age-related differences in cued recall: Effects of support at encoding and retrieval. *Aging, Neuropsychology, & Cognition*, **9**, 276-287.
- RAAIMAKERS, J. G. W., & SHIFFRIN, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207-262). New York: Academic Press.
- RATCLIFF, R., CLARK, S. E., & SHIFFRIN, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 163-178.
- ROTELLO, C. M., & HEIT, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject. *Journal of Memory & Language*, **40**, 432-453.
- ROTELLO, C. M., & HEIT, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, **28**, 907-922.
- ROTELLO, C. M., MACMILLAN, N. A., & VAN TASSEL, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory & Language*, **43**, 67-88.
- SHEFFERT, S. M., & SHIFFRIN, R. M. (2003). Auditory registration without learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 10-21.
- SHIFFRIN, R. M., HUBER, D. E., & MARINELLI, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 267-287.
- SHIFFRIN, R. M., & STEYVERS, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**, 145-166.
- SHIFFRIN, R. M., & STEYVERS, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73-95). Oxford: Oxford University Press.
- WIXTED, J. T., & STRETCH, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, **11**, 616-641.
- YONELINAS, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, **25**, 747-763.
- YONELINAS, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory & Language*, **46**, 441-517.

NOTES

1. We thank Bill Hockley for pointing out that mirror effects also seem to be observed for AR when strength is varied between, but not within, lists (cf. Clark & Shiffrin, 1992).
2. The occurrence of nonverbal stimuli might be temporarily represented by a vector of w features in a short-term memory buffer. The features making up this vector may be primarily perceptual, but they may also be retrieved from lexical/semantic traces if the stimulus reminds the subject of something or someone. Once entered into the short-term buffer, the encoding of a nonverbal stimulus occurs in the same manner as that described above.
3. In the present applications, varying c has the same effect on AR as varying u^* , because both are encoding parameters.

4. Malmberg, Holden, and Shiffrin (2004) found this assumption to be necessary to quantitatively account for performance on a plurality discrimination task.

5. Figure 2 shows that when study time is very short, increasing u^* in the model slightly increases FARs from one to two presentations, because the probability of exceeding the familiarity value increases more than the probability of successfully recalling that a rearranged pair was not studied.

6. For 2 subjects, a slope estimate of the zROC could not be computed, and they were eliminated from this analysis.

7. One might be tempted to assume that frequently encountered types of stimuli are made up of different features than are less frequently

encountered types of stimuli. For instance, the g parameter varies the distinctiveness of the memory traces, and it is used to model word frequency effects. Low g values lead to better recognition performance, and hence, one might assume, say, that words have lower g values than do pseudowords. However, this would assume that the features associated with words are less common than those associated with pseudowords (or faces or Chinese characters), which would have to be the case in order to achieve better encoding of words.

(Manuscript received September 9, 2005;
revision accepted for publication December 20, 2005.)