Psychological Review
1998, Vol. 105, No. 4, 724–760

0033-295X/98/$3.00

Familiarity Breeds Differentiation: A Subjective-Likelihood Approach to the Effects of Experience in Recognition Memory

James L. McClelland
Carnegie Mellon University
and the Center for the Neural Basis of Cognition

Mark Chappell
Carnegie Mellon University and University of Queensland

With repeated exposure, people become better at identifying presented items and better at rejecting items that have not been presented. This *differentiation* effect is captured in a model consisting of item detectors that learn estimates of conditional probabilities of item features. The model is used to account for a number of findings in the recognition memory literature, including (a) the basic differentiation effect (strength-mirror effect), (b) the fact that adding items to a list reduces recognition accuracy (list-length effect) but extra study of some items does not reduce recognition accuracy for other items (null list-strength effect), (c) nonlinear effects of strengthening items on false recognition of similar distractors, (d) a number of different kinds of mirror effects, (e) appropriate *z*-ROC curves, and (f) one type of deviation from optimality exhibited in recognition experiments.

When we become familiar with something—an object, a person, or an item in a memory experiment—we come to recognize it more reliably. This general observation applies across a range of paradigms, including perceptual identification and recognition memory. *Familiarization*—defined here in terms of the amount or duration of exposure to an item—increases the probability of correct identification and recognition. In this article we consider the nature of this familiarity effect.

One way familiarity might affect identification and recognition is through what might be thought of as a bias effect. The essential idea is captured by many detector-based models, such as the logogen model of Morton (1969). In such models, familiarization of an item increases the resting activation or bias of the detector for that item. This in turn increases the probability that an input will be identified as that item.

James L. McClelland, Department of Psychology, Carnegie Mellon University, and the Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania; Mark Chappell, Department of Psychology, Carnegie Mellon University, and Department of Psychology, University of Queensland, Gold Coast, Queensland, Australia.

The work reported in this article was supported by Grants MH00385 and MH47566 from the National Institute of Mental Health. Preliminary versions of the model were presented at the 27th and 28th Annual Mathematical Psychology Meetings and at the 35th Annual Meeting of the Psychonomic Society.

We would like to thank Simon Dennis, Randy O'Reilly, Marius Usher, and the PDP Research Group for useful discussions. We would also like to thank Stephan Lewandowsky and Roger Ratcliff for helpful comments on earlier versions of this article. Mark Chappell would like to thank Brian Junker and Jay Kadane for making him welcome in their statistics classes and for useful discussions regarding Bayesian statistics.

Correspondence concerning this article should be addressed to James L. McClelland, Center for the Neural Basis of Cognition, 115 Mellon Institute, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213 or to Mark Chappell, who is now at the School of Applied Psychology, Gold Coast Griffith University, PMB 50, Gold Coast Mail Centre, 9726, Queensland, Australia. Electronic mail may be sent to jlm@cnbc.cmu.edu or to m.chappell@bhm.gu.edu.au.

The assumption that familiarity affects a bias term makes some sense because familiarity often covaries with likelihood of occurrence: Typically, stimuli that are more likely to occur will be more familiar. If we have partial or ambiguous information, we increase our likelihood of being correct if we select more probable alternatives. Thus, a tendency to identify uncertain items as examples of familiar ones should lead to an increased overall likelihood of correct identification, relative to an unbiased strategy. Such biases may actually affect what we see as well as what we are willing to say we might have seen in a perceptual identification task. For example, in the interactive activation model of visual word identification (McClelland & Rumelhart, 1981), word frequency determines the resting activation level of each word unit, thereby exerting a biasing effect on the tendency to see the corresponding word, and, by means of top–down influences, the tendency to see the letters in that word. These perceptual activations then serve as the basis of overt identification responses.

Morton's (1969) model and the interactive activation model have both been useful in many ways, but in our view their treatment of familiarity effects seems to miss a crucial element. In both cases, the implicit assumption appears to be that there is no effect of familiarization on our knowledge of the item itself. Intuitively, one might suppose that the more familiar one is with a particular item, the more one would know about how to differentiate it from other items.

The idea that familiarity increases our ability to differentiate one item from another has a long history in the animal conditioning literature.[1] In conditioning studies, it has frequently been found that preexposure to stimuli to be used in discrimination training often facilitates later discrimination learning, especially when the to-be-discriminated stimuli are similar to each other (E. Gibson & Walk, 1956; Oswalt, 1972). E. Gibson (1940) introduced the concept of *differentiation* to account for this

[1] We are indebted to Lisa Saksida for bringing this literature to our attention.

phenomenon. The Gibsons discussed it in several later works (E. Gibson, 1969; J. Gibson & Gibson, 1955) but offered no specifics on the underlying mechanism that might give rise to differentiation.

Thus far we have considered these matters in the context of perceptual identification experiments, but similar issues arise in experiments on recognition memory for items such as words that are presented for study in lists. Because there are extensive relevant findings, we focus attention on this *single-item recognition* paradigm in this article. A key finding that has been the focus of a great deal of attention concerns the effects of familiarization of items during the study phase of such an experiment, produced either through longer exposure relative to other items or through repetition. Recently, a serious challenge to most existing models of recognition memory has arisen in the form of a (null) finding that appears to suggest that differentiation plays an important role: Familiarization of one subset of items in a study list does not appear to affect recognition accuracy for *other* items from the same list.

The situation can best be appreciated by considering it within the context of a classic model, Gillund and Shiffrin's (1984) search of associative memory (SAM) model of recognition memory. In the original version of this model, a memory image (analogous to one of Morton's logogens) is created when an item is studied, and associations are formed between the memory image and the study context and between the memory image and the actual stimulus. For example, if the word *piano* appeared as a study list item, a memory image would be formed for it, and a positive strength value would be assigned to the association between this image and the stimulus word and between the image and the study context.

Familiarization of an item in the original SAM model strengthens both the image–stimulus association and the image–context association. The strengthening of the image–context association tends to have an effect much like that of strengthening the bias term in Morton's (1969) model. Specifically, it tends to lead to an increase in the activation of the memory image of a more familiar item, relative to a less familiar item, whether the item itself or some other item is presented as a probe of memory. Thus, there is an increase in the activation of the memory image of that item, both when the item is presented as a recognition stimulus and when a different item is presented. The extra, spurious activation arising in the latter case tends to degrade performance, leading the model to predict that recognition accuracy for a particular studied item (as measured by d') will *decrease* when the familiarity of other items in the list is increased. In fact, however, no such effect is found (Ratcliff, Clark, & Shiffrin, 1990). This null effect of familiarization with some items on accuracy of recognition of other items is now widely known as the *null list-strength effect.* It has become one of the focal points of recent studies of recognition memory because it poses difficulties not only for the original version of the SAM model, but for many other prior models of recognition memory as well.

To account for the null list-strength effect, Shiffrin, Ratcliff, and Clark (1990) introduced the idea that, as the associations of a memory image with its corresponding stimulus and context are strengthened, the associations of the memory image and all other potential stimuli are correspondingly *weakened.* These alterations of association strength have the combined effect of decreasing the activation of the memory image of a particular item when some other stimulus is presented at the time of test. Adopting the terminology previously used by the Gibsons, Shiffrin et al. (1990) described the resulting effect as a *differentiation* of the response of the system to old and new items: Familiarization of an item appears to increase the activation of the corresponding memory image when the stimulus item is presented at test, but does not appear to produce an increase in activation of the memory image when other stimulus items are presented. In a subsequent discussion, Murnane and Shiffrin (1991a) conceptualized differentiation in terms of the assumption that

> the similarity between a stored item and a cue item with which it was not rehearsed decreases as the stored item becomes stronger or better learned, presumably because the differences between the two become more salient. The more dissimilar the item cue and the stored item, the less the stored item is likely to be activated by the item cue. (p. 871)

In addition to the null list-strength effect, another key aspect of the recognition memory literature can be viewed as reflecting the effects of differentiation. Indeed, in this case the relation to differentiation is more direct: When all of the items on a list are given extra exposure compared with a baseline condition, participants become *less* likely to falsely recognize new items at the same time as becoming more likely to correctly recognize the old, studied items. This effect, which we call the *strength-mirror effect,* would naturally be expected if a differentiation process were operating in recognition memory experiments.

The present article seeks to provide an account of the basis of the differentiation process. The concept of differentiation we present here appears to be consistent with the concept as construed both by the Gibsons and by Shiffrin et al. (1990). Indeed, it may be interpreted as a theory of what differentiation amounts to and how it might actually occur as a result of experience.

Our central hypothesis is simply this:

> As we become familiar with something, we come to have more accurate knowledge of its characteristics. That is, we come to have clearer knowledge of what properties it has and what properties it does not have.

We suggest that this increased clarity of knowledge of the properties of the item is a crucial aspect of familiarization and is the basis of the phenomenon of differentiation.

Our model assumes that when an item is presented to an individual, the item creates a psychological experience that can be characterized as a set of experienced features or attributes. We suppose further that these features are probabilistically related to the identity of the item. For example, the word *piano* evokes a psychological experience of auditory, visual, and other properties, and this psychological experience will differ from occasion to occasion.

We can then characterize the memory trace a person forms for an item in terms of a set of quantities that capture the estimated likelihood of occurrence of the various possible features in presentations of the item: Stated precisely, each of these quantities can be viewed as an estimate of the conditional probability of experiencing the corresponding feature, given that the experience was produced by the item. Using these ideas, we

can capture our proposal about the effect of familiarization in the form of the following, slightly more technical, version of our hypothesis:

> Familiarization with an item gives rise to a refinement of psychological quantities that serve as estimates of the probabilities that particular features will be experienced upon the presentation of the item.

Given this characterization of our hypothesis, we have framed our model explicitly in probability estimation terms. For example, the items of information a participant stores from an experience of an item or computed from a test presentation of an item are represented as estimates of conditional probabilities. We do not suppose that participants actually experience these probabilities per se; rather, we suggest that it is useful to construe the items of information stored as isomorphic to estimated probabilistic quantities and to construe the operations performed by the memory system as preserving this isomorphic relationship. There are two reasons why this approach seems useful to us: First, it allows us to draw ideas from statistics and probability theory into the formulation of our computational model, and second, it allows us to determine to what extent the behavior of our model is "rational" or "optimal" from the point of view of probabilistic inference. The enterprise bears some relationship to the "rational analysis" approach advocated by Anderson (1990). We agree with Anderson that it is useful to explore the extent to which human behavior can be construed as an optimal response of a system with physical limitations to a probabilistic world, though we have found ourselves repeatedly observing that model features that can be seen as rational or optimal from some points of view lead to real performance limitations that can be far from optimal in certain task situations. Several aspects of our model, including the ones we next consider, appear to have this characteristic, as we will discuss later.

A key feature of the model is the assumption that the initial representation of an item, based, say, on a single brief presentation, will contain relatively noisy and generic or undifferentiated estimates of the probability of occurrence of the item's features. As we become more and more familiar with the item, the estimates of these probabilities become more dependent on the actual features of the item and move away from their initially noisy and generic values. When the item is later encountered in a memory test, these refinements in the feature probability estimates lead to an increase in the estimated likelihood that the entire set of features experienced are indeed those of the item. In our model, this estimated likelihood corresponds to the subjective "sense of familiarity" that we might have upon encountering an item in a memory test, and the change in this estimate is the basis for the increase in the sense of familiarity associated with an item as the number of times we have encountered it increases. The increase in the estimated likelihood is not just a bias; in fact, it is accompanied by a simultaneous decrease in the estimated likelihood when the set of features arises from some other, unrelated item. Thus, familiarity breeds differentiation.

Illustration of the Process of Differentiation

To make these ideas concrete, we present here a simple example of the differentiation process at work. In this example, there are four possible features that an item might have, and on average half of these are active in any given item. We will suppose that a particular set of two of these features tend to be activated by the presentation of a specific word, which might as before be *piano,* whereas the remaining two features tend not to be activated when this word is shown. The presentation of the word *piano* during study results in the formation of a detector for this item. Initially, the detector forms estimates of these probabilities that are moved $\pm.10$ from the generic value of .50 toward the values actually encountered in the presentation of the word, as shown in Table 1.

With these initial estimates of the probabilities of occurrence of the individual features of the word *piano,* we can consider the probability of observing a particular set of features, given that they actually represent the stored word *piano.* We consider two possible sets of features that might be observed: those corresponding to the features of *piano* itself, and those that might be characteristic of some other word, say, *sofa.* To represent the average case of another item generated at random, we will imagine that *sofa* also has two of the four features, one of which it shares with *piano,* and the other of which it does not. The computation of the estimate of the likelihood of observing the entire ensemble of features embodies the simple assumption that the probability of an entire pattern is just the product of the probabilities of the individual elements of the pattern. This assumption, which is often invoked in probabilistic inferences as a simplifying approximation, may or may not be correct in reality, so this is one of the ways in which the quantities computed must be viewed as corresponding to estimates of probabilities rather than true probabilities. We apply this computation first to the feature values corresponding to the word *piano*: The first value is a 1, and this is expected to be a 1, with probability .60. The second value is a 0. Because a 1 is expected with probability .4, and 0 is the complementary event, the probability of the 0 is also .60. Similar reasoning applies to the rest of the features. Taking the product of the four feature-value probabilities, all of which are .60, we obtain an estimated probability of .130 of observing the values actually characteristic of the word *piano.* This relatively low probability essentially reflects our uncertain estimates of the probabilities of the features of the word.

Now we apply this same procedure to determine the probability that the features characteristic of the word *sofa* were produced by the item corresponding to our parameterized detector for *piano.* We get the same results as before where the features of *sofa* and *piano* agree, but we get smaller feature probabilities in those cases where the features disagree. Thus, the detector's estimate of the probability that the third feature is a 1 is .40, and a 1 is observed for this feature in the case of the word *sofa.* Furthermore, the estimate of the probability that the fourth feature is a 1 is .60, but a 0 is observed for this feature when the input comes from the word *sofa,* so the probability of observing a 0 in this case is also .40.

On the basis of the initial estimates of the individual feature probabilities, the estimated likelihood of observing the values experienced for the word *sofa* is .058. Clearly, there is a difference between this and the value obtained for the word *piano,* but the *piano* pattern is only a little more than twice as likely as the *sofa* pattern. Suppose, however, the word *piano* is studied

Table 1
An Example of Differentiation

Example inputs, stored patterns, and resulting likelihoods	Observed or computed quantities				
	Individual features				
	1	2	3	4	Overall
Input pattern for *piano* item at study	1	0	0	1	
Stored feature probability estimates for *piano* based on one study	1.60	.40	.40	.60	
Test input pattern for *piano*	11	0	0	1	
Likelihood that above matches studied *piano* given stored estimates	1.60	.60	.60	.60	0.130
Input pattern for new *sofa* item at test	11	0	1	0	
Likelihood that above matches studied *piano* given stored estimates	1.60	.60	.40	.40	0.058
Stored feature probability estimates for *piano* after two study presentations	.70	.30	.30	.70	
Likelihood that studied *piano* matches test input *piano* given two-study estimates	.70	.70	.70	.70	0.240
Likelihood that studied *piano* matches test input *sofa* given two-study estimates	.70	.70	.30	.30	0.044

Note. See Illustration of the Process of Differentiation for complete explanation of the feature probabilities and likelihoods that result from the storage and testing of patterns.

twice, and as a result of the second study the estimates of the probabilities of the features were moved further from their generic values toward the values actually representative of *piano* by an additional step of 0.10. Now the estimated likelihoods obtained for test presentations of *piano* and *sofa* become much more differentiated. The estimated likelihood of observing the values found in the word *piano* would increase to 0.24, whereas the estimated likelihood of observing the values found in the word *sofa* decreases to .044. Thus, the additional experience has led to a differentiation of the response of the detector: It gives a higher estimated likelihood when the item that it stands for is presented and a lower estimated likelihood when some other item is presented.

The assumption that the detectors are initialized with generic estimates of the probabilities of the individual features and adapt these estimates relatively slowly deserves some discussion in the context of the issue of optimality, as mentioned above. One might think it would make more sense to change the estimate immediately to 1 upon encountering a feature value of 1. Two perspectives on this matter might be entertained. According to one of these, gradual learning could be seen as rational or optimal in the sense that it avoids overcommitment based on preliminary evidence. This perspective can be justified by the idea that items themselves can actually give rise to varying patterns of features from time to time, so that, although some of the initially encountered values might be reliable features of the object, others might not. For example, sometimes we may think of the beautiful music a piano makes, and other times we may think how difficult it can be to play. In our simplified example described above, each presentation of an item resulted in the very same set of features, but in the complete model the feature values of items are probabilistic, so they do vary from presentation to presentation. If in reality features are probabilistic in this way, then treating the probability of a feature as 1 or 0 on the basis of a single presentation would be inappropriate and incorrect. When features really are probabilistic, starting with a generic estimate and moving gradually away from the generic estimate on the basis of experience may be a better approximation to true optimality. Another perspective might assume that the memory system is physically limited in how rapidly it can change in response to inputs. According to this perspective, the memory system may be sluggish because of physical limitations on how rapidly the underlying neural mechanisms can change. In certain cases this sluggishness will, happily enough, turn out to approximate optimality, but in others (when, for example, there is no uncertainty about features) it will no longer approximate optimality quite as well. One way of synthesizing these perspectives would be to note that biological resources (energy and molecular structures) must ultimately be allocated to make the underlying changes in the brain, and evolution may have chosen a policy that represents a reasonable compromise that approaches optimality in many cases without the overexpenditure of its resources.

A further necessary aspect of the complete model can also be viewed as a blessing or a curse, depending on one's perspective. In the actual model, both the initial estimates of feature probabilities and the calculation of estimated likelihoods are subject to noise or variability. This noise was left out of our example above for expository clarity, but it is crucial to the fit of the model to the data. There are certainly circumstances under which noise can foster optimal solutions to computational problems, though even in these cases the noise must be gradually reduced or averaged out for true optimality of performance (Hinton & Sejnowski, 1983; Movellan & McClelland, 1995). Many times, however, noise seems most plausibly viewed as an inherent limiting factor. In these cases, the presence of the noise itself can be thought of as nonoptimal. However, the decision-making processes that are applied, given the noise, may still be coping with it in the best possible way.

For the sake of accounting for data, we don't have to take a stand on the optimality or lack thereof of the various features of a model. However, our opinion is that, in most cases, the truth about the optimality of many model properties lies somewhere between the obvious extremes. That is, we think cognition is likely to depend on mechanisms that both are constrained by physical limitations and have been constrained by evolution to function sensibly, even if they are not completely optimal for every possible situation. In view of this, we don't think rationality or optimality should be the basis for deciding whether to accept a proposed model for any particular case. Rather, it pro-

vides a heuristic guide in the search for an adequate model and a basis for evaluating how close to optimal the mechanisms used by the system under study actually are.

The above example captures the essence of our claims about the nature of the differentiation process, and the further discussion of its optimality or lack thereof indicates that there are specific features of the model, in particular its noisiness and gradualness, that play a role in its account for relevant data. There remain a few additional aspects of the model that we have not yet introduced for the sake of communicating these most central points. Below, the entire model is described in complete detail, with these additional aspects included, following a presentation of the recognition memory phenomena that the model will address.

Before turning to these matters, we note that there are a number of extant models that deal with aspects of recognition memory. In addition to SAM (e.g., Murnane & Shiffrin, 1991b), such models include the rational model described by Anderson and Milson (1989), the recognition portions of the theory of distributed associated memory (TODAM; Murdock & Kahana, 1993), Minerva II (Hintzman, 1988), the neural network model described by Chappell and Humphreys (1994), the attention–likelihood model conceived of by Glanzer, Adams, Iverson, and Kim (1993), and the newer REM model as discussed by Shiffrin and Steyvers (1997).[2] Our model combines insights from a number of these models and accounts for much of the same data that some of the other models account for. A comparison of our model to these others is provided in the General Discussion.

Data

In a pure single-item recognition paradigm, items such as words are first presented in a study list. In fact, in many of the studies we cite, words were presented in sentences (Murnane & Shiffrin, 1991a) or word pairs (Ratcliff et al., 1990), generally with the intention of reducing displaced rehearsals. At test, in all studies cited here, single words were presented, some of which had been on the list, and some of which were new. In most studies cited here, the participant's task was to judge whether each test word is *old* or *new* (a small number of studies involved confidence or frequency judgments). Participants may correctly accept a word as old (a *hit*) or *correctly reject* a new word. They may incorrectly judge a new word to be old (a *false alarm*) or incorrectly judge an old word to be new (a *miss*).

Data are often characterized in terms of quantities derived from signal detection theory (Green & Swets, 1966). This theory assumes that recognition judgments are based on a continuous random variable. For recognition one may think of this variable as the sense of familiarity that a participant has upon experiencing an item at test. The value of the variable is thought to be drawn from one of two normal distributions, one characterizing the old stimuli and one characterizing the new stimuli. The theory also assumes participants adopt a criterion value of the random variable, such that they respond *old* whenever the random variable exceeds the criterion, and *new* otherwise. Standardly, the proportion of hits (*hit rate*) and the proportion of false alarms (*false alarm rate*) are the key dependent measures. Two other quantities, d' and β, are often derived from these measures. The first of these, d', is a measure of discriminability of the two distributions, reflecting their separation in units of their standard deviation. The quantity β reflects the placement of the criterion with respect to the *old* and *new* distributions. It corresponds to the likelihood that the observation comes from the *old* distribution, divided by the likelihood that it comes from the *new* distribution, at the value of the continuous random variable where the response criterion is located (these likelihoods correspond to the heights of the two distributions at the criterion). Moving the criterion up on the continuum (so that a higher value of the variable is required for an *old* response) corresponds to an increase in β, and in fact if the distributions are normal, the log of β—hereinafter $\ln[\beta]$—increases linearly with the criterion.

If the participants' response criterion is manipulated (e.g., by manipulating payoffs or the probability of presenting *old* items at test), or if participants are required to place their responses in designated confidence categories (so that the boundaries between confidence categories can be treated as distinct criteria), curves called z-ROC curves may be plotted. If the underlying distributions are normal, the slope of the z-ROC curve will be an estimate of the ratio of the standard deviations of the two distributions ($\sigma_{new}/\sigma_{old}$).

Signal detection theory is, of course, of tremendous relevance to single-item recognition and other psychological phenomena, but we have found it important to look directly at the primary data as well as the derived signal detection measures. The derived measures help summarize the effects of certain variables, but their use can obscure important relationships in the data. In particular, researchers often tend to focus on d' and to pay less attention to β. This may be in part because of the belief that changes in β can arise from criterion shifts that could be extraneous to the underlying effects of interest.

Although extraneous factors can influence β, other factors can as well. In fact, β will change if the criterion remains constant and there are changes in the means or standard deviations of one or both of the distributions produced by old and new stimuli. Thus, changes in β may not reflect effects of experimental variables on the placement of the criterion with respect to the psychological variable; instead, the experimental variables might be affecting the psychological variable itself. Because, as we shall see, such changes occur in our model and produce β effects, we consider both d' and β effects in our presentation of the data and in our later discussion of the accounts offered by our model for the data.

Differentiation, the Strength-Mirror Effect, and the Null List-Strength Effect

We now describe the particular effects within the single-item recognition paradigm in which we are interested. We consider

[2] After we presented a preliminary version of our model at two conferences (Chappell & McClelland, 1994; McClelland & Chappell, 1994), we learned that Shiffrin was simultaneously and independently developing a model with some similar characteristics (Shiffrin, oral conference presentation, March 27, 1995). A technical report describing this model (Shiffrin, 1995) and a subsequent publication (Shiffrin & Steyvers, 1997) have since appeared.

first the results obtained in comparing performance with a pure weak list, where items are presented once during study, to performance with a pure strong list, where items are either presented a number of times or just once but for a longer time. These list types are called *pure* to contrast them with a *mixed* condition described below. Table 2 shows data from the pure conditions of several such list-strength experiments. Ratcliff et al.'s (1990) experiments manipulated strength by means of presentation time. Experiment 1 presented single words. In all their other experiments, words were presented in pairs. In Murnane and Shiffrin's (1991a) experiments, words were presented in sentences, and sentences were presented three times for the pure strong lists.

In all cases the increase in the hit rate as one goes from the pure weak to the pure strong list is accompanied by a decrease in the false alarm rate. This is the strength-mirror effect, which we take to be due to the differentiation process. Note, however, that the changes in false-alarm rate are generally somewhat smaller than those of the hit rate in Table 2. The data from Ratcliff et al. (1990) suggest a trend toward a reduction in $\ln[\beta]$, but the effect was not completely consistent across the experiments in that article and was not apparent in Murnane and Shiffrin (1991a).

The two conditions in Table 2 are from experiments that also included a mixed condition, in which half of the items were seen once, and half were seen three times or for a longer time. The former are called *mixed weak* items, and the latter are called *mixed strong*. Interest in these experiments focuses on whether there is a difference in performance between pure weak and mixed weak items and between mixed strong and pure strong items, as measured by d'. When the first list-strength experiments were performed by Ratcliff et al. (1990), they pointed out that all extant models predicted that d' would be worse in the mixed weak condition than in the pure weak condition and worse in the pure strong condition than in the mixed strong condition. In fact, they found no difference, and many subsequent experiments have confirmed this conclusion. Table 3 shows the mixed conditions of the Murnane and Shiffrin (1991a) studies, which, when compared with the corresponding entries in Table 2, exemplify the null list-strength effect: d' for mixed weak items tends to be about the same as d' for pure weak items, and d' for mixed strong items tends to be about the same as d' for pure strong items.

In a meta-analysis of just the pure weak and mixed weak conditions of a large number of list-strength experiments, Chappell and Humphreys (1994) pointed out that there is a trend for both the hit rate and the false alarm rate to decrease as one goes from the pure weak to the mixed weak condition. This trend is exemplified by the Murnane and Shiffrin (1991a) data shown in Table 3. In agreement with this, Hirshman (1995) has done an extensive meta-analysis of list-strength experiments, as well as a number of his own experiments, focusing on the change in β as one goes from pure weak to mixed weak and from mixed strong to pure strong lists. He concluded that β increases for each of these comparisons.

In the preceding section, we saw that when familiarity of items is strengthened, sensitivity increases, and there is a strength-mirror effect. In this section, we see that the strengthening of other list items does not affect d' for items of a given strength but does appear to lead to a decrease in both hits and false alarms and a corresponding increase in β. Many treatments have focused on the d' effects but have ignored the influences on β. In our model, both the d' and the β effects of strength and list-strength are explained by the differentiating effects of familiarization.

Table 2
Pure-Strength Data

	Pure weak				Pure strong			
Experiment	HR	FAR	d'	$\ln[\beta]$	HR	FAR	d'	$\ln[\beta]$
				Ratcliff et al. (1990)				
1	0.646	0.228	1.12	0.207	0.740	0.202	1.48	0.140
2	0.665	0.165	1.41	0.385	0.700	0.159	1.52	0.358
3	0.677	0.143	1.52	0.464	0.745	0.111	1.87	0.531
4a	0.659	0.345	0.80	0	0.827	0.272	1.55	−0.261
4b	0.625	0.338	0.74	0.039	0.748	0.255	1.32	−0.010
				Murnane and Shiffrin (1991a)				
1	0.77	0.15	1.78	0.26	0.86	0.10	2.36	0.24
2 B1	0.81	0.18	1.79	0.03	0.82	0.13	2.04	0.22
2 B2	0.78	0.17	1.73	0.16	0.88	0.12	2.35	0.00
3 B1	0.70	0.24	1.23	0.11	0.81	0.14	1.96	0.20
3 B2	0.67	0.25	1.11	0.13	0.84	0.14	2.07	0.09
4 eq	0.76	0.19	1.58	0.14	0.85	0.13	2.16	0.10
Mean	0.75	0.20	1.54	0.14	0.84	0.13	2.16	0.14

Note. The values of the signal detection quantities d' and $\ln[\beta]$ are calculated from the experiment-mean HR and FAR. Means in the last row are just for the Murnane and Shiffrin (1991a) experiments. HR = hit rate; FAR = false alarm rate; B1 = Block 1; B2 = Block 2; eq = equal arithmetic condition.

Table 3
Mixed Data

Experiment	Mixed weak HR	FAR	d'	$\ln[\beta]$	Mixed strong HR	FAR	d'	$\ln[\beta]$
			Murnane and Shiffrin (1991a)					
1	0.66	0.11	1.64	0.67	0.87	0.15	2.16	−0.1
2 B1	0.66	0.16	1.41	0.41	0.92	0.16	2.40	−0.49
2 B2	0.78	0.17	1.73	0.16	0.86	0.11	2.31	0.17
3 B1	0.67	0.18	1.36	0.32	0.81	0.16	1.87	0.11
3 B2	0.73	0.23	1.35	0.09	0.81	0.21	1.68	−0.06
4 eq	0.69	0.17	1.45	0.33	0.86	0.15	2.12	−0.05
Mean	0.70	0.17	1.49	0.33	0.855	0.155	2.09	−0.07

Note. The values of the signal detection quantities d' and $\ln[\beta]$ are calculated from the experiment-mean HR and FAR. HR = hit rate; FAR = false alarm rate; B1 = Block 1; B2 = Block 2; eq = equal arithmetic condition.

Similarity and Strengthening

Another experiment that illustrates the differentiation process at work was performed by Hintzman and Curran (1995). Following Hintzman, Curran, and Oppy (1992), they presented a list of nouns in which individual nouns appeared 1 time, 3 times, 8 times, or 20 times. At test, experimenters requested judgments of frequency (JOF), and distractors included very similar words to those on the list: a singular noun if the plural had occurred on the list, or a plural if the singular had occurred on the list. Hintzman and Curran were particularly interested in false recognition of these similar distractors. Our consideration of this data concerns only whether the participant gives a JOF response of 0 (corresponding to a new response in our model) or a JOF response greater than 0 (corresponding to an old response). In fact, Hintzman et al. (1992) were strongly of the opinion that participants make a recognition decision first, then go on to make a frequency judgment only if they judge the item to be old. Not only was there a bimodal distribution over frequency judgments in evidence, with a mode at 0, but in two experiments they asked for recognition judgments and obtained a very similar pattern of results to that obtained with JOF = 0 versus JOF > 0. Thus, it seems appropriate to model these data in a model of item recognition.

Figure 1 shows JOF = 0 proportions. The decrease and then increase in proportion of JOF = 0 indicates that false recognition of similar items first increases up to three presentations of the studied item, then decreases again with further presentations. The decrease is clearly quite small, and in fact Hintzman and Curran (1995) emphasized how poorly participants learn to reject similar distractors, even after as many as 20 presentations. But all of Hintzman and Curran's and Hintzman et al.'s (1992) experiments showed a U-shaped effect. In our model, the U-shaped trend in JOF = 0 responses is indicative of the differentiation process at work. The fact that JOF = 0 responses to similar distractors level off below 100% suggests that differentiation is generally incomplete. We see later that the probabilistic nature of item representations in our model provides an explanation for this aspect of the data.

List-Length Effect

Rather than strengthening some list items, another manipulation is to increase the length of the list. The decline in performance when this is done is called the *list-length effect.*

Table 4 shows the hit and false alarm rates, as well as d' and $\ln[\beta]$. Ratcliff and Murdock (1976) had words studied individually and reported proportions of high confidence hits and correct rejections for various output positions and three list lengths. The entries in Table 4 are averages across output position (described in the Table 4 *Note*) and show two possible list-length comparisons. Murnane and Shiffrin's (1991a) experi-

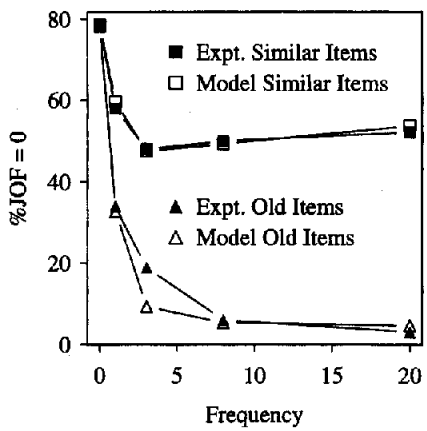

Figure 1. Data (filled symbols) showing initial generalization and later differentiation of recognition responses to items similar to studied items, as a function of the frequency of study presentations of each item. The dependent measure, % judgments of frequency (JOF) = 0, is treated as 1 minus the recognition rate. Expt. = experiment. From Experiment 1 (neutral condition) of "When Encoding Fails: Instructions, Feedback, and Registration Without Learning" by D. L. Hintzman and T. Curran, 1995, *Memory & Cognition, 23,* p. 217. Copyright 1995 by Douglas L. Hintzman and Tim Curran. Reprinted with permission of the authors. Also shown are simulation results (open symbols) from the model described later.

Table 4
List-Length Data

Experiment	Long list HR	Long list FAR	Long list d'	Long list ln[β]	Short list HR	Short list FAR	Short list d'	Short list ln[β]
Ratcliff and Murdock (1976)								
32–16	0.74	0.20	1.48	0.15	0.83	0.11	2.18	0.30
64–32	0.66	0.35	0.80	−0.01	0.74	0.20	1.48	0.15
Murnane and Shiffrin (1991a)								
1 L3	0.68	0.25	1.14	0.12	0.77	0.15	1.78	0.26
2 B1–B2–L3	0.70	0.24	1.23	0.11	0.80	0.18	1.76	0.10
3 B1	0.71	0.31	1.05	−0.03	0.70	0.24	1.23	0.11
3 B2	0.70	0.29	1.08	0.02	0.67	0.25	1.11	0.13
Mean	0.70	0.27	1.12	0.06	0.73	0.20	1.47	0.15
Gronlund and Ohrt (1994)								
1	0.661	0.131	1.54	0.888	0.742	0.069	2.12	0.542
2	0.738	0.207	1.45	0.131	0.787	0.160	1.80	0.174

Note. For Ratcliff and Murdock's (1976) studies, we have averaged different numbers of output blocks in different conditions, so as to average over the same number of tested items (e.g., four blocks of 8 items in length 16 condition compared with two blocks of 16 in the length 32 condition). HR = hit rate; FAR = false alarm rate; 32–16 = long list has 32 items, short has 16; 64–32 = long list has 64 items, short has 32; L3 = last third of long list compared with short list; B1–B2–L3 = average Block 1 and Block 2 and compare with third section of long list; B1 = Block 1; B2 = Block 2.

ment studied words in sentences. The short condition had 50 words in 10 sentences; the long condition had 150 words in 30 sentences. For their Experiment 2, we averaged over Block 1 and Block 2 and compared with the L3 items, which were at the end of the long list, thus controlling for lag.

Gronlund and Ohrt (1994)[3] presented and tested single words, controlling for lag in the two conditions. Gronlund and Ohrt's Experiment 2 presented all words in one list, but they came from two categories of different sizes.

In general, these experiments show a slightly smaller effect of list length on hit rates than false alarm rates. The ln[β] measures shown in Table 4, computed from the mean hit and false alarm rates shown, indicate a trend for a smaller ln[β] in the long condition, though this trend is clearly reversed in one of the reported studies (Experiment 1 of Gronlund & Ohrt, 1994).

Frequency-Mirror Effect

Mirror effects in item recognition have most commonly been studied in the context of such variables as preexperimental word frequency and concreteness (for early studies of the word-frequency effect, see, e.g., Balota & Neely, 1980; Mandler, Goodman, & Wilkes-Gibbs, 1982; Rau & Proctor, 1984). Words of lower frequency and words of higher concreteness tend to produce larger values of d', and with very few exceptions (e.g., Hoshino, 1991); when there is a difference in d' between two conditions in a recognition experiment, the hit rate will be higher for the superior condition, and the false alarm rate will be lower (for reviews, see Glanzer & Adams, 1985; Glanzer et al., 1993). It is worth noting that the frequency effect can be produced by a within-list design. For example, if a single list containing high- and low-frequency words is studied and the test list similarly contains such a mixture, the frequency-mirror effect is observed (e.g., Morris, 1978), so that low-frequency test words produce both a higher hit rate and a lower false alarm rate than words of higher frequency.

Hintzman, Caulton, and Curran (1994) made the point that describing a mirror effect in the terms used above is not sufficiently precise. Thus, if a large decrease in false alarm rate as one goes from the weak condition to the strong condition is accompanied by a miniscule increase in the hit rate, this should not simply be classified as a mirror effect. In fact, in a number of Hintzman et al.'s experiments they found the equivalent of a more conservative (higher) β in the strong condition for both frequency and abstract-concreteness manipulations. Hoshino (1991) also reported such tests, finding significant differences in the same direction and in fact no significant increase in hit rate in three out of four experiments. Calculating ln[β] from the hit and false alarm rates reported by Glanzer and Adams (1990), we found the same trend reported by Hintzman et al. (1994). However, there are experiments, such as Experiment 4 of Ratcliff, McKoon, and Tindall (1994), in which the effect, if present, is relatively small (the increase in hit rate from high to low word frequency, averaged over several strength and list-strength conditions, is .082, whereas the average decrease in false alarm rate is .099). Thus, it appears that there is a tendency for a higher ln[β], as well as a higher d', with lower frequency

[3] The data shown in Table 4 were provided by Daryl Ohrt (personal communication, December 1994).

and higher concreteness, but the extent of this tendency appears to vary considerably.

z-ROC Curves

Many investigators have considered the effects of various manipulations on the slopes of *z*-ROC curves. As we detail below, slopes are usually less than 1. From a signal-detection theory point of view, slopes less than 1 would be consistent with larger variance in the distribution of familiarity values produced by old items, relative to new.

Effects of experimental manipulations on *z*-ROC slopes are not entirely consistent across studies. In some studies, there is no reliable effect; in others, *z*-ROC slope tends to decrease as d' increases. Ratcliff, Sheu, and Gronlund (1992) plotted *z*-ROC curves for all conditions in a list-strength experiment, both by varying the proportions of old and new words at test and through confidence ratings. They found straight lines, with slopes around 0.8. The slope varied little between the various list-strength conditions. Similarly, Yonelinas (1994, Experiment 3) found slopes of 0.79 and 0.80 for weak and strong conditions, stronger items having been presented for longer. It is clear, however, that the slope must change at some point, as it must approach 1 for very weak items. Ratcliff et al. (1994) verified this, finding slopes of the order of 0.95 when items were presented for 50 ms, 0.9 for 100 ms, and 0.8 for 400 ms. In recent studies, Glanzer, Kim, Hilford, and Adams (in press) have shown *z*-ROC slopes decreasing as d' increases as a function of several variables affecting item strength.

Gronlund and Elam (1994) studied the effect of list-length on the *z*-ROC curve slopes. In their first experiment, they found that slopes for both conditions were not significantly different from 1. In that experiment, participants participated in 4 to 7 sessions, with 8 short and 8 long lists per session. In a second experiment participants saw and were tested on only one list. The average slopes were 0.60 for the short condition and 0.69 for the long condition. This finding gives some weight to the suggestion by Glanzer et al. (1993) that the constant slopes found by Ratcliff et al. (1992) came about because of the large number of lists participants saw in those experiments (the Glanzer et al., 1993, model predicts that slope should decrease with item strength). Interestingly, Ratcliff et al. (1994) found no dependence of the *z*-ROC slope on list length; in their experiment, participants saw 18 lists per session for 7 or 10 sessions. Conversely, Yonelinas (1994) presented participants with similarly large numbers of lists and did find an effect of list length on the *z*-ROC curve slope. Overall then, the evidence suggests that *z*-ROC slopes do tend to decrease with shorter lists and that the effect is more robust when the possibility of cross-list influence is minimized.

Turning to studies manipulating item variables, Glanzer and Adams (1990) found that there was a strong trend for the *z*-ROC slope to be lower in the stronger (higher d') condition, when strength was determined by word properties such as frequency or concreteness. They were in agreement that the slope should be less than 1 and in the range of 0.6–0.8. Ratcliff et al. (1994) also found a significant effect of word frequency on *z*-ROC curve slope, with low-frequency words having a lower slope.

The findings of the series of studies and meta-analyses by Glanzer et al. (in press) are consistent with the following summary: *z*-ROC slopes are generally less than 1. Most experimental variables that affect d' also affect *z*-ROC slope, but under many conditions the effects are quite small and therefore are not always reliable. In some studies, strength-related variables that affect d' do not affect *z*-ROC slopes once d' exceeds a value of about 0.5, but other studies find strength-related effects on *z*-ROC slopes over a wider range of d'.

Cumulative Confidence Rating Data

A recent study by Balakrishnan and Ratcliff (1996) addressed a set of issues related to the extent and nature of the information that participants use for placing response criteria in recognition memory experiments. Consider the situation in which an observer must decide whether an input represented as a value on a continuum (such as a familiarity continuum) corresponds to a "signal" item (an old item on a memory test) or to a "noise" item (a new item in a memory test). With complete information about the distribution of values generated by signals and the distribution of values generated by noise, an ideal or optimal observer could compute the likelihood that the input value received on a particular trial comes from the signal distribution and the likelihood that it comes from the noise distribution. By dividing the first value by the second, the observer would then have a likelihood ratio indicating the relative likelihood that the value represents signal versus noise. Response criteria that form boundaries between response and confidence categories could then be assigned to particular values of the likelihood-ratio variable. For example, a *most-sure–new* response might be assigned to likelihood-ratio values less than a criterion at .05, *moderately sure–new* to values between .05 and a second criterion at .25, *unsure–new* to values between .25 and a third criterion at 1.0, and *unsure–old* to values between 1.0 and a fourth criterion at 2.0, etc. What is crucial here is not the example values but the fact that they are quantified in terms of the likelihood-ratio variable, not the underlying familiarity variable.

Balakrishnan and Ratcliff (1996) considered the possibility that participants in recognition memory experiments can and do compute the relevant likelihood ratios and compare them to criteria that are fixed at particular likelihood-ratio values. They showed that under these conditions, cumulative confidence rating curves obtained in the different conditions of a list-strength experiment should cross. A cumulative confidence rating curve is a curve in which confidence category rank (ordered from *most-sure–new* to *most-sure–old*) is plotted on the horizontal axis against the cumulative total proportion of responses falling in that category or any lower category. The predicted crossings are shown in Figure 2. As illustrated in the figure, the three curves for new stimuli from the pure-weak, pure-strong, and mixed conditions should all cross; the two curves for weak–old stimuli from the pure-weak and mixed-weak conditions should cross; and the two curves from strong–old stimuli from the pure-strong and mixed-strong conditions should cross. The reason for this effect is that strengthening old items lowers the likelihood ratio for items with low familiarity and raises the likelihood ratio for items with high familiarity.

Balakrishnan and Ratcliff's (1996) analysis led them to test

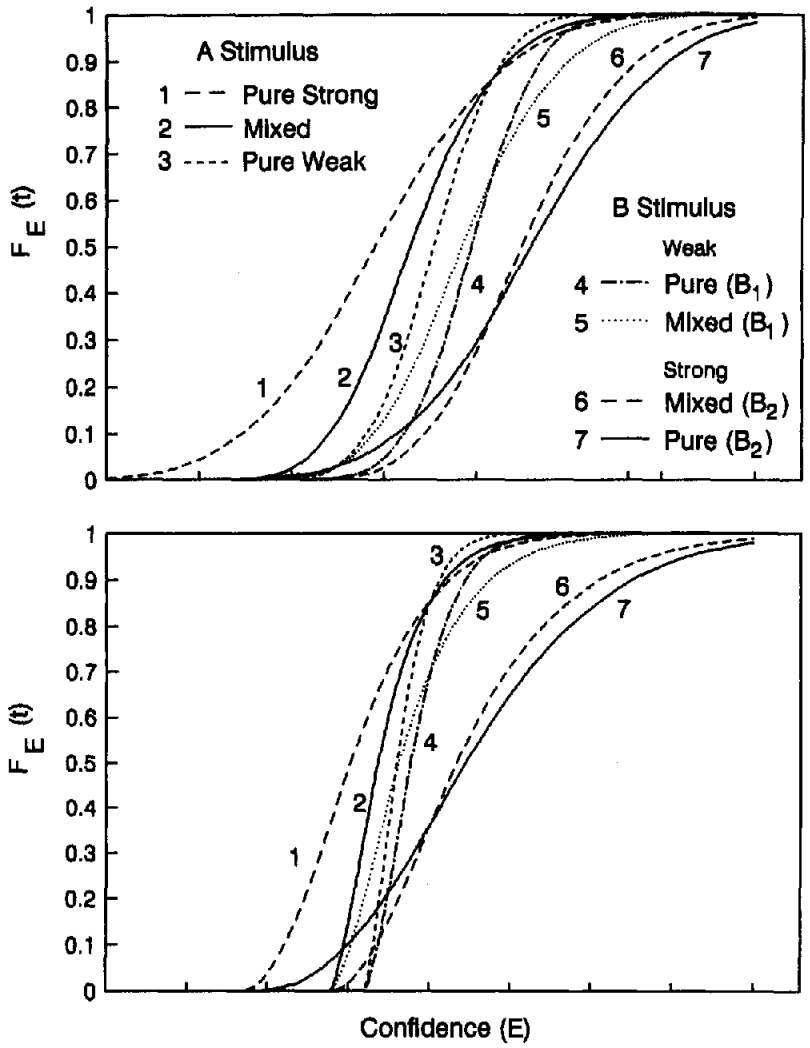

Figure 2. Distribution crossover predictions of the likelihood ratio model considered by Balakrishnan and Ratcliff (1996), for the case in which the "familiarity" distributions for new, weak–old, and strong–old stimuli are all normal and have equal variance. Crossing also occurs in other cases. See text for additional explanation. Reprinted from Figure 3 of "Testing Models of Decision Making Using Confidence Ratings in Classification" by J. D. Balakrishnan and R. Ratcliff, 1996, *Journal of Experimental Psychology: Human Perception and Performance, 22,* p. 620. Copyright 1996 by American Psychological Association. Reprinted with permission of the authors.

for the predicted crossings of cumulative confidence ratings curves, but they found that in fact the curves from the appropriate conditions of a single-item recognition memory experiment run for this purpose did not cross. This finding poses a strong challenge to the model of Glanzer et al. (1993) and any other model in which participants are thought to adjust their likelihood-ratio computations on the basis of the characteristics of items in particular list conditions.

It might be thought that our model would make the same prediction as the class of models tested by Balakrishnan and Ratcliff (1996) for the crossing of cumulative confidence curves because, like them, our model makes responses on the basis of likelihood ratios. However, our model differs from the models mentioned above in that the hypotheses our model contrasts in the likelihood ratios its detectors use do not rely on information about the distribution of values associated with other items used in the experiment. Each detector relies solely on information about the single item it detects and general statistical properties of all items, rather than other items that may be on the list. As a result, the cumulative confidence curves calculated in our model are not constrained a priori to cross. There are other factors that might cause the curves to cross, but as is seen in our simulations, the patterning of the curves in the simulation is quite consistent with the pattern seen in the Balakrishnan and Ratcliff data.

Summary

The literature on recognition memory includes a broad range of findings, including the strength-mirror effect; the null list-

strength effect; the list-length effect; z-ROC curve slopes less than one and often, but not always, affected by item strength, list length, or frequency; a U-shaped effect of item repetitions on false recognition of similar items; the mirror effect due to stimulus properties such as preexperimental word frequency; effects of list strength and word frequency on the signal-detection parameter $\ln[\beta]$; and noncrossing of cumulative confidence curves in a list-strength experiment. Table 5 summarizes all of these effects. A number of these findings appear to implicate processes related to differentiation; that is, the findings appear to suggest that many variables that increase correct recognition of old items also facilitate correct rejection of new items.

It should also be noted that many of these findings present problems for current memory models, *especially when they are considered in combination.* Consider first the list-length effect together with the null list-strength effect. The increasing false alarm rate in the list-length effect indicates that one presentation of an item increases the false alarm rate. We do not have data for two presentations, but all the data reviewed for list-strength (Chappell & Humphreys, 1994) and pure-strength manipulations (see earlier in this section) indicate that as we go from one to three presentations the false alarm rate certainly does not increase, and instead appears to decrease (differentiation). Thus the relationship between the number of presentations of an item and false alarm rate is *nonlinear,* and in fact *nonmonotonic.* This is why this combination of findings is such a problem for the linear global-matching models (Shiffrin et al., 1990). Hintzman et al.'s (1992) and Hintzman and Curran's (1995) experiments illustrate this same general nonmonotonicity more graphically, with distractors that are similar to the studied items.

As we have already noted, several of the newer models do address some aspects of the data. However, as is seen when we consider these in the General Discussion, none of these models have addressed all of the phenomena listed in Table 5, and some of the models make incorrect predictions for some of these effects. It is thus of considerable interest to see whether the model we present in the next section, which synthesizes some of the key aspects of other models, can accommodate all of these experimental findings.

Model

The model considers a situation in which a human participant is given a series of study items, one at a time, from a single multi-item list, and forms memory representations of these items in the course of the study phase of the experiment. During a later test phase the participant is given a series of items, some of which occurred during study and some of which did not. We assume that each test item is compared to the stored representation of each of the list items and that for each the participant computes an estimate of the likelihood that the test item is a representation of the item that generated the stored representation. If any one of these estimates exceeds a criterion, the participant will respond *old;* otherwise, the participant will respond *new.*

Table 5
Summary of Recognition Memory Phenomena to Be Modeled

Variable	Effect
Item strength	d' increases
List length	d' decreases
List strength	d' unaffected
	$\ln[\beta]$ increases with list strength
	Cumulative confidence curves do not cross
z-ROC slope	Generally < 1
	Decreases with most variables as d' increases
	Sometimes constant with increasing strength for $d' > 0.5$
Item repetition	Similar distractor false alarm rate increases, then decreases slightly
Word frequency	d' increases for low-frequency words
	$\ln[\beta]$ often increases with d'

Several of the essential features of the model are qualitatively described in the introduction. Here, we present these features in full detail, accompanied by other features that were previously left out in an effort to make the essential elements of the model clear. We begin with a description of the featural representations of items themselves, and then turn to the ways in which these are represented and processed in the memory system. The presentation uses mathematical formulas for precision, but we have endeavored as much as possible to spell out key points in words, so that readers interested in the general points may follow the presentation even if they do not follow all of the details of the notation.

Items and Their Features

During a study or test presentation of an item, we assume that what the memory system has access to is a pattern of values, S, across a set of N feature dimensions. For concreteness, one may think of the dimensions as represented by units and the values as represented by the activations of the units. Only two values, 1 or 0 (*active* or *inactive*) are allowed. The set of feature values elicited by an item is not completely fixed; rather, each stimulus used in the experiment may be thought of as a probabilistic generator of patterns of activation. This corresponds to the idea that each time we view a particular stimulus or think about a particular word we represent it slightly differently. The generator item determines the probability that each feature will be active in our representation of it, but there is occasional variability in the specific features that are actually present in the representation. Therefore, we are concerned with the probabilities of particular features being active when a particular item is being presented. We denote these probabilities: $p_i^{\mu} = p(S_i = 1 \mid I^{\mu})$, where I^{μ} represents the condition that the features were produced by the generator for item μ. Note that we follow the convention of using Greek superscripts to individuate items and Latin subscripts to individuate features that may be present in items.

In the item generators, the probabilities for each feature unit being active are conditionally independent of the activities of other feature units. That is, the probability that a particular unit will have the value 1 or 0 is not affected by the actual values assigned to other units. This assumption is adopted as a simple default in lieu of other more complex assumptions, and it is unlikely that the principles under consideration are affected by this assumption. For example, if certain features tended to occur together or not at all, the computations of actual probabilities would be far more complex, but the effects of experience that we will describe would still operate in the same general way. (Correlated features add less information per feature than uncor-

related features, and in the limit, perfectly correlated features can be treated as single features.) As a result, the conditional probability of a given pattern of activation, S, across the units, given that it was generated by item μ, is

$$p(S|I^\mu) = \prod_i p(S_i|I^\mu) = \prod_i (p_i^\mu)^{S_i}(1 - p_i^\mu)^{(1-S_i)}. \quad (1)$$

That is, if the unit is active we multiply by the conditional probability, otherwise, the complementary event has occurred, and we multiply by 1 minus the conditional probability. We may also say this is the probability of this pattern, S, being generated by the item generator, I^μ.

In the general case, the p_i^μ could take different values for all values of i. However, we restrict our analysis to the simple case where an expected fraction of units, f, has a high conditional probability, p_1, whereas the remainder have a low conditional probability, p_0.

Item Representations

The memory system in the model works with vectors, one to represent the current input, S, and one to represent the detector formed by the model for each list item that has been studied. The current input, S, is a pattern of 0s and 1s, these being the only two values allowed for input units. The ith element of this pattern is represented by S_i.

The vector associated with each detector, D^μ, is a pattern of real numbers in the $\{0, 1\}$ interval. Every time a new item is encountered during study, a new detector unit is assigned to it. Thus, the μth detector unit represents the μth study item. The value of the detector's ith element is the detector's estimate of the probability that the ith input will have a value of 1, when the input was generated by item μ. These estimates are represented using the notation ρ_i^μ. We use ρ instead of p in these and the remaining quantities to indicate that they are estimates, rather than the true probabilities.

Operation of Item Detectors

In this section, we develop the equations characterizing the behavior of each item detector on presentation of a test item, where we take its task to be the computation of a quantity representing an estimate of the log-odds that the input, S, was generated by the item, I^μ, for which the detector stands. This quantity depends on the likelihood computation discussed in the introduction. The actual computation involves additional factors that were not highlighted earlier because they are not affected by differentiation. We call this quantity the *log-odds*. In this and other cases, we avoid the continual repetition of the word *estimated* for simplicity, but we retain the use of Greek letters for the actual quantities as a reminder that all of these quantities are estimates.

The computation of the log-odds is taken directly from a Bayesian hypothesis-testing framework. The log-odds is the logarithm of the ratio of estimates of what are called the *posterior probabilities* of two hypotheses, one being the main hypothesis of interest, and the other being some null hypothesis or reasonable contrasting alternative (Kass & Raftery, 1995). The posterior probability of a hypothesis is simply the probability of the hypothesis, given the "evidence"; in this case, the evidence is the set of values represented over the input units, S.

In the log-odds computation, each detector contrasts two hypotheses: (a) the hypothesis that S was generated by the item corresponding to the detector, which we represent as the event I^μ, and (b) the hypothesis that it was not generated by that item, which we represent as the event $\overline{I^\mu}$. The ratio of the posterior probabilities is called the *odds ratio;* the logarithm of that ratio is the log-odds.

The odds ratio of the posterior probabilities of two hypotheses can be expressed in terms of the ratio of two products, each proportional to one of the two posterior probabilities. Each product consists of two factors: an estimate of the prior probability of the hypotheses and an estimate of the likelihood of the evidence, given the hypothesis. For our case, this relationship takes the form

$$\frac{\rho(I^\mu|S)}{\rho(\overline{I^\mu}|S)} = \frac{\rho(S|I^\mu)\rho(I^\mu)}{\rho(S|\overline{I^\mu})\rho(\overline{I^\mu})}. \quad (2)$$

Here, $\rho(I^\mu|S)$ is the estimated posterior probability of the hypothesis that S was generated by I^μ, and $\rho(\overline{I^\mu}|S)$ represents the estimated posterior probability of the contrary hypothesis.

The odds ratio can be viewed as the product of two factors, $\rho(S|I^\mu)/\rho(S|\overline{I^\mu})$ and $\rho(I^\mu)/\rho(\overline{I^\mu})$, called the *likelihood ratio* and the *prior odds,* respectively. Taking logs of both sides, we obtain the log-odds that item μ generated the input S, represented by $\omega^\mu(S)$. The log-odds is the key decision variable used in our model. It is a sum of two terms, the log of the likelihood ratio, called the *log-likelihood,* and the log of the prior odds:

$$\omega^\mu(S) = \ln \frac{\rho(S|I^\mu)}{\rho(S|\overline{I^\mu})} + \ln \frac{\rho(I^\mu)}{\rho(\overline{I^\mu})}. \quad (3)$$

We now consider separately how the log likelihood and the log of the prior odds are estimated.

The numerator in the likelihood ratio, $\rho(S|I^\mu)$, is the key factor we considered in the introduction, namely the estimated probability of the input given that it was generated by item μ, and $\rho(I^\mu)$ represents the estimated prior probability that the input would be generated by item μ. Equation 1 indicates how the true conditional probability $p(S|I^\mu)$ can be computed if the actual probabilities of occurrence of the features, p_i^μ, are available. As already indicated, the detector forms estimates of these conditional probabilities, ρ_i^μ, during study, using a "learning rule" described below. Using the ρ_i^μ, our detector can estimate the conditional probability of an input pattern using a version of Equation 1, in which we replace the true probabilities with their estimated values

$$\rho(S|I^\mu) = \prod_i \rho(S_i|I^\mu) = \prod_i (\rho_i^\mu)^{S_i}(1 - \rho_i^\mu)^{(1-S_i)}.$$

This computation incorporates the assumption that the feature values in the input patterns are conditionally independent, given that the input pattern was generated by item μ. This may not

always be true in real experience, and we view it as an approximation that might be built into the memory system to allow the estimation of relevant probabilities to proceed.

Now we consider $\rho(S|\overline{I^\mu})$, the estimated probability of the input, given that it was not generated by item μ. The event $\overline{I^\mu}$ encompasses all of the events corresponding to presentations derived from other list items as well as those corresponding to all nonlist items. We assume that the detectors have no actual knowledge of quantities being computed by the other detectors. Instead, each detector uses an estimate of the probability of the input, given that it was generated by some completely unknown item generator. This estimate is just the probability of observing the particular ensemble of features, under the assumption that each feature occurs independently of the others, and with a probability equal to the overall probability of occurrence of each feature in patterns. This estimate would be correct if in fact the feature values used in the item generators were assigned independently. Again, this independence assumption is true of the inputs actually used in our simulations but may not be true of real items encountered in experiments, so again the model treats the memory system as employing assumptions that allow the estimation of relevant probabilities. In the inputs we use in our simulations, the actual probability that an individual feature will take the value 1 is simply equal to $fp_1 + (1 - f)p_0$ and is the same for all features. Although in principle this quantity would have to be estimated (and might in principle be estimated separately for each feature), we assume that the estimate used for each feature in each detector, represented by ρ_a, actually corresponds to the true probability, to avoid proliferation of free parameters in the model. Given this assumption, the estimated probability of the input given that it was generated by an unknown generator becomes

$$\rho(S|\overline{I^\mu}) = \prod_i \rho_a^{S_i}(1 - \rho_a)^{(1-S_i)}.$$

Combining the numerator and denominator and taking logs, we obtain the following expression for the log likelihood:

$$\ln \frac{\rho(S|I^\mu)}{\rho(S|\overline{I^\mu})} = \sum_i \left[S_i \ln\left(\frac{\rho_i^\mu}{\rho_a}\right) + (1 - S_i) \ln\left(\frac{1 - \rho_i^\mu}{1 - \rho_a}\right)\right]. \quad (4)$$

This quantity plays a central role in our model. For ease of reference, we represent it as $\lambda^\mu(S)$.

Turning now to the prior odds, the numerator, $\rho(I^\mu)$, represents the participant's estimate of the probability that a particular input will in fact be item μ. The true probability that a particular test item is study list item μ is just the probability that a particular test item will be drawn from the study list, $p(list)$, times the probability of choosing item μ given that an item is chosen from the study list, $p(I^\mu|list)$. The latter probability is $1/L$, where L is the length of the list. We assume that the participant uses estimates of the relevant quantities, denoted $\rho(list)$ and Λ, to estimate the probability of occurrence of the item using $\rho(I^\mu) = 1/\Lambda\ \rho(list)$. The denominator of the prior odds is an estimate of the probability that the input will not be produced by study item μ. We assume that the participant uses the complement of the estimated probability that the item would be item μ, which is $\rho(\overline{I^\mu}) = \Lambda - 1/\Lambda\rho(list) + 1 - \rho(list)$. The prior odds then simplifies to

$$\frac{\rho(I^\mu)}{\rho(\overline{I^\mu})} = \frac{\rho(list)}{\Lambda - \rho(list)}. \quad (5)$$

In summary, substituting the various expressions we have derived above back into Equation 3, we obtain the following expression for the log-odds for item μ:

$$\omega^\mu(S) = \sum_i \left[S_i \ln\left(\frac{\rho_i^\mu}{\rho_a}\right) + (1 - S_i) \ln\left(\frac{1 - \rho_i^\mu}{1 - \rho_a}\right)\right] + \ln \frac{\rho(list)}{\Lambda - \rho(\overline{list})}. \quad (6)$$

The sum over i is the log-likelihood $\lambda^\mu(S)$, and the added term to the right is the log of the prior odds. We now consider how the values of the crucial ρ_i^μ variables that enter into the calculation of $\lambda^\mu(S)$ are learned.

Learning

A key idealization we have adopted in the model is the following: We assume that every presentation of a particular item within the study phase of a memory experiment always results in the reselection of the same detector unit. How the right detector would be selected for reactivation is not currently modeled, though we imagine that a realistic system for reactivation might rely on computations quite similar to those that occur during test in the model as currently implemented. In such a system, it is likely that the idealization would not be completely accurate, in that a re-presentation of an item would sometimes be treated as a new item, and some new items may occasionally produce false reactivation of a previously studied item. We return to this matter in the General Discussion.

In any case, the learning process as it takes place in the model is as follows. When a detector is first created, it is initialized with noisy, generic values for its estimates of the probabilities of features occurring in the item for which it stands. The actual observed features of the item are then used to adjust these estimates, according to the learning rule described below. On all subsequent re-presentations of the item during study, the same detector is again selected, and the observed-feature values from these subsequent presentations are used each time to make further adjustments to the feature probability estimates.

The learning rule for adjusting the estimates has its basis and history both in the neural network and the Bayesian inference literature and is based on the idea that if a sequence of 0s and 1s is generated by an independent, random process with a fixed probability of producing a 1, then a running average of the sequence of 0s and 1s is a useful estimate of the probability.

Several neural network and connectionist researchers (Grossberg, 1976; Rumelhart & Zipser, 1986; von der Malsburg, 1973) have explored a learning rule based on this property. The rule produces connection weights that are estimates of the probability that the unit on the input side of a connection is active, given that the unit on the receiving or output side of the

connection is active (see also Levy, Colbert, & Desmond, 1990, for development and physiological support of similar equations). We use a version of this rule to calculate estimates of the conditional probability that an input feature is active in instances of a particular item:

$$\rho_i^\mu(t+1) = \rho_i^\mu(t) + \epsilon[S_i(t) - \rho_i^\mu(t)]. \quad (7)$$

We will often use the term *weight* for the feature probability estimates, ρ_i^μ, both because of the connectionist inspiration for the learning rule and for ease of reference.

According to this rule, when the weights of detector μ are updated at a given time t, the weight for feature unit i will increase (toward a maximum value of 1) if the ith input value is 1 and decrease toward 0 otherwise. It is easy to show that if this rule is applied repeatedly, then the value of the $\rho_i^\mu(t)$ will come to hover around the conditional probability that the ith feature equals 1 for the μth item. In fact, if the learning rate, ϵ, decreases over time, the weight will actually converge to the correct value of the conditional probability (White, 1989). For this reason, we gradually reduce ϵ each time a detector learns. The specific approach we use for reducing ϵ is based on the following Bayesian considerations.

In a Bayesian approach, each probability estimate can be thought of as having an associated distribution, called the *beta distribution*, with parameters a and b (e.g., Novick & Jackson, 1974). Here, a can be thought of as a representation of the number of times a 1 has been observed for S_i in a presentation of item μ, and b can be thought of as a representation of the number of times a 0 has been observed. The best estimate of the conditional probability ρ_i^μ that $S_i = 1$ in item μ is then simply $a/(a+b)$, which is the mean of the beta distribution. Initial values of a and b establish a prior value for this mean. The procedure for updating the beta distribution parameters is to add 1 to either a or b after each observation of the S_i, depending on whether S_i is a 1 or a 0. The updated estimate of the conditional probability then becomes

$$\frac{a}{a+b} \rightarrow \frac{a+S_i}{a+b+1}.$$

We have found that initial values of a and b can be chosen such that $a/(a+b) = \rho_a$ and

$$\frac{a+S_i}{a+b+1} \approx \frac{a}{a+b} + \epsilon\left[S_i - \frac{a}{(a+b)}\right]$$

$$= \rho_a + \epsilon[S_i - \rho_a].$$

When ϵ is small, its value approximates $1/(a+b)$.

The above equations suggest the scheme that we have used in our simulations for reducing ϵ as experience with an item accumulates. The learning rate for the sth presentation of an item is given by

$$\epsilon_s = \frac{1}{n_0 + s}, \quad (8)$$

where n_0 (not necessarily an integer) is chosen to produce the desired initial value for ϵ, namely

$$\epsilon_1 = \frac{1}{n_0 + 1}. \quad (9)$$

It may be easily shown that if $\boldsymbol{l} = l_1, \ldots, l_s$ is the sequence of values (1s and 0s) that an estimate is exposed to during learning over s presentations, then

$$\rho(\boldsymbol{l}) = \frac{n_0 \upsilon + \sum_{j=1}^{s} l_j}{n_0 + s}, \quad (10)$$

where υ is the estimate's initial value. The estimate is computing a running average, and n_0 reflects the relative influence assigned to the initial value of the estimate in the running average. In particular, if $n_0 = 0$, then $\rho(\boldsymbol{l})$ is just the running average of the observed sequence of activations.

Random Initial Values of the Weights

Now that we have a learning rule, we must decide what starting value the weights ρ_i^μ should have. As previously noted, the weights are initialized with noisy, generic values. Specifically, the weights are given random initial values whose mean is equal to the average probability of a 1 occurring across all of the elements of all items:

$$\rho_a = fp_1 + (1-f)p_0.$$

The actual values are obtained by generating a value, x, according to a normal distribution with mean μ_n and standard deviation σ_n, and then taking the logistic of these numbers,

$$\text{logistic}(x) = \frac{1}{1+e^{-x}},$$

to obtain initial weights between 0 and 1. We treat σ_n as a free parameter and find the value of μ_n that yields the desired value of ρ_a (see Appendix). The use of the logistic-normal distribution for the initial values of the weights, ρ_i^μ, means that $\text{logit}(\rho_i^\mu)$ will be normally distributed, where $\text{logit}(x) = \ln x/(1-x)$. Thus, we can see the initialization as specifying a normally distributed random initial estimate of the log-likelihood of the feature, given that the experience arose from I^μ.

Decision Noise in Item Detectors

As noted in the introduction, a source of variability that remains invariant regardless of the amount of learning proved to be necessary to account for all of the data. Without this, the model's performance tended to improve too quickly with repeated exposures. We incorporated a source of such variability by assuming that some of the weights on each item detector remain permanently fixed at their initial values. The fraction of weights that can learn, called κ, thus becomes a parameter of the model. Because the weights at each detector are assigned random initial values, as explained above, the presence of weights that do not learn effectively adds noise independently to the decision process at each detector. The noise varies from trial to trial depending on which input features have values of 1, but noise is uncorrelated across detectors because the initial

values of the weights are independent. Other ways of adding independent noise to each detector would be expected to have comparable effects.

Analytic Solution and Theoretical Distributions

An approximate analytic solution, characterizing the distribution of log-likelihoods that would be computed by a detector under various conditions, is given in the Appendix.

Figure 3 uses this solution to allow us to show how the distributions of log-likelihoods for stimuli produced by the generator corresponding to the detector (same stimuli; see Figure 3, right curves) and for stimuli produced by other generators (*different* stimuli; see Figure 3, left curves) will change with the number of presentations. The distribution at the back of the figure represents the response of the detector to either kind of stimulus before any learning has actually taken place; the distribution is the same for both cases. In the Appendix, formulae are derived for all moments of these distributions. To produce Figure 3 we assume, approximately correctly, that these distributions are normal.

The distribution at the back of Figure 3 is the single distribution of log-likelihoods produced for all items by a new detector before any learning has occurred. The remaining distributions come in pairs, showing that the means of these distributions for same and different stimuli move apart as the number of presentations increases. This movement of the means as a function of experience in the full model corresponds to the movement of the estimated conditional probabilities in the introductory presentation illustrated in Table 1. Though the standard deviations of the distributions also increase somewhat, the difference between the means is increasing faster than the standard deviations. Further, the mean for distributions on the left is moving away from the origin faster than the standard deviation is increasing, leading to differentiation (with a fixed criterion).

Estimated Log Odds and the Criterion for Recognition Responses

An important constraint in the design of our model concerns the question of the placement of the criterion for determining whether an item is old or new. In some approaches, the placement of the criterion has been left out of consideration as a nuisance variable, and attention has focused only on the sensitivity variable d'. Our approach has been, instead, to assume that participants in a given experiment adopt a fixed criterion for responding that they apply to all conditions. The reason for this is that many of the effects of greatest interest in the recognition memory literature, including the numerous "mirror" effects, involve statements about the relative movement of hit and false alarm rates, which are discarded in a pure sensitivity analysis. To make predictions for such variables, we must take the placement of the decision criterion explicitly into account.

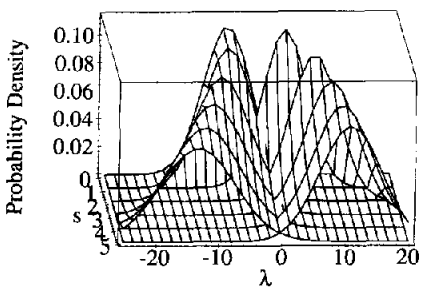

Figure 3. Distributions of log-likelihoods for a single detector, as a function of the number of presentations. The distribution at the back left is the distribution for all stimuli just after the detector is initialized, before any learning has occurred. The distributions to the right of this baseline distribution are for stimuli produced by the generator corresponding to the detector, after successive stimulus presentations. The distributions moving forward and to the left of the baseline distribution are for stimuli produced by generators for other items.

One approach is to allow differences between experimental conditions to affect the placement of the response criterion. We have avoided this approach, instead preferring to consider the possibility that changes in hit and false alarm rate actually reflect the effects of experimental variables on the actual sense of familiarity rather than the placement of the criterion. In particular, we assume that the performance of the model depends on whether the estimated log odds of any of the detectors exceeds a criterion ω_C. Although ω_C may vary as a function of instructional variables, such as the degree to which the experimenter encourages participants to be liberal or conservative in their recognition responses, we have assumed that it is held constant within experiments or sets of experiments from the same laboratory in which variables such as list length, item and list strength, and preexperimental word frequency are manipulated.

Our use of the log-odds as the crucial decision variable, rather than the log-likelihood, allows the model to incorporate effects of changes in list length and the probability that a test item is old without abandoning the use of a fixed criterion across conditions that manipulate these variables. The estimated log-odds differs from the estimated log-likelihood by a term that depends on the estimated list length and the estimated probability that a test item comes from the list.

The effect of using the log estimated odds is equivalent to using the estimated log-likelihood ratio with a stipulated criterion adjustment determined by the variables Λ and $\rho(list)$, which would be the standard practice under signal detection theory (Green & Swets, 1966). Our choice to treat these variables as affecting the computations performed by the detectors reflects the possibility that the memory system itself is sensitive to the number of list items. This could occur if the detectors actually compete in some way, as they would in a neural network implementation like the one we consider in the general discussion, or in models such as ACT* (Anderson, 1983), where spreading activation is divided among the recipients, which in this case would be the detectors.

The probability that the log-odds of any of the detectors exceeds criterion is equivalent to the probability that the largest estimated log-odds value exceeds the threshold. Thus, the statistics of performance depend ultimately on the distribution of the maximum of the estimated log-odds over all the detectors. We consider the distributions of these maxima as we present relevant aspects of the simulations and also analytically in the Appendix.

Simulation of a Participant's Performance in an Experiment

Each simulation run corresponds to the running of one simulated subject in a recognition memory experiment. The proce-

dure used in the simulation serves as a summary of the model's account of the processes that take place in such a situation. The first step in each run, before simulating the memory experiment itself, is the creation of a set of generators for the study and test items that will be used in the experiment. Each such generator, represented by $\boldsymbol{p}^{\mu}$, is simply a list of values specifying, for each of the N elements of an item representation, what the probability of that element having a value of 1 will be. These elements, represented by p_i^{μ}, are set independently for each feature of each item and take the value p_1 with probability f or p_0 with probability $1 - f$.

Next comes the simulation of the study phase of the experiment. For the study phase, some of the items are designated study items, and examples of these are generated one at a time using the $\boldsymbol{p}^{\mu}$ associated with one of the item generators. That is, a vector of 0s and 1s is generated, where the probability of a 1 in each component is the corresponding component of $\boldsymbol{p}^{\mu}$. For each unique study item, a detector is created at random, with its weights initialized as described above. At this stage a proportion, κ, of the weights are randomly chosen to be the ones capable of learning. These weights then begin learning the conditional probabilities of the item using Equation 7. When an item is presented more than once, the same detector is again selected, thereby refining its estimates of the probabilities of the features present in the corresponding item. On the sth presentation, the learning rate is given by Equation 8.

In the test phase, items based on both studied (old) and not-studied (new) item generators are presented. For each test item, an input pattern $\boldsymbol{S}$ is generated according to the corresponding generator $\boldsymbol{p}^{\mu}$. Each of the learned detectors then computes its estimate of the log-odds for $\boldsymbol{S}$, using Equation 6. If any detector has an estimated log-odds above the criterion, ω_C, the model makes an *old* decision; otherwise, the response is *new*.

Parameters and Fitting Method

In assessing the adequacy of the model, we wished to show that the model could account for the findings summarized in Table 5, particularly those that have been robustly replicated across experiments. In addition, we sought quantitative fits to certain carefully chosen illustrative experiments. As described in the Appendix, in order to fit our stochastic model we used our approximate analytic solution in combination with the gradient descent algorithm PRAXIS (Gegenfurtner, 1992).

The fitting process was greatly simplified by setting some simulation variables that in principle could be treated as estimates to their actual values. Specifically, the estimated list-length parameter Λ was set to the actual length of the list used in the experiment being simulated. Likewise, the estimated probability that a test item would be old $\rho(old)$ was set to the actual probability that an old item would be tested (which was almost always .5). Also, as previously mentioned, the estimated probability that a feature will be 1, ρ_a, was set to the actual expected value as determined by the item generator parameters f, p_0, and p_1. This left only eight parameters to be fitted in the simulations (see Table 7, below).

Assessing fit with confidence intervals. What constitutes an adequate fit? Ideally, we would adopt the following approach: For each experimental point (e.g., Hit Rate in some condition) of interest, we would calculate an empirical confidence interval (Loftus & Masson, 1994) and a confidence interval for the model's prediction and use these to decide if the two are substantially different by computing a confidence interval for the difference between the means. We could be, say, at least 99% confident that the difference between the means was not greater than δ, the width of the 99% confidence interval for the difference between the experimental and simulation values, if that confidence interval spanned 0. The confidence interval spans 0 if and only if the absolute value of the difference between the values is less than $\delta/2$. The smaller the width of the confidence interval, the more meaningful this would be. Thus, there would be two "criteria" for our fits; the confidence level and the width of the confidence interval, δ. We would then consider a fit to be adequate if for all the data points of interest the simulation means are within $\delta/2$ of the empirical means. If we assume (almost certainly conservatively) that the various condition means are independent, and there are c of them included in our confidence intervals, then we can be at least 0.99^c confident that we are within δ of all of them. Three issues arose with respect to the application of the above approach.

Estimating experimental confidence intervals. For one of the experiments we fit, the statistics of interest are z-ROC slopes and intercepts, for which statistics allowing computation of experimental confidence intervals have been provided. However, in other cases, the published papers do not provide data that would allow calculation of empirical confidence intervals. In such cases, we have nominated values for these confidence intervals that seem reasonable, trying to err slightly on the side of making the confidence intervals narrower than they might really be, so as to call attention to cases of potential misfit between the model and the experimental data.

Minimizing simulation confidence interval widths. For hit and false alarm rates, we estimated that the experimental confidence intervals we are working with are in the range of 3% to 4% (i.e., ± 1.5% to ± 2%). These intervals are relatively wide, and if a similar width was allowed for the simulation, it would be too easy to get the model's predictions inside the resulting combined confidence interval. To avoid this problem, we ran very large numbers of replications of the simulation at the end of the parameter search, so that the model's confidence intervals would be on the order of one tenth of the size of the estimated experimental confidence intervals. The contribution of the variability of the simulation to the overall confidence interval then becomes negligible. Given that we have the computing power to make these confidence intervals this small, there is no good reason not to do so. Thus, the total confidence interval is taken to be the experimental confidence interval. In the case of other statistics such as z-ROC slopes and intercepts, we were not able to calculate the simulation confidence intervals. Here, we ran very large numbers of simulation replications so that, once again, the noise in the simulation should be negligible, with the goal of producing simulated values within the experimental confidence intervals.

One fit or many separate fits? Finally, we confronted the problem of consistency of simulation parameters across experiments. One approach to this issue could have been to try to simulate the results of several experiments all at once with a single simulation, adopting a common set of parameters for all.

The extreme version of this approach proved to be unworkable because there were large procedural differences between experiments that result in relatively large differences in overall performance levels. On the other hand, piecemeal simulation of individual experiments was highly unsatisfactory as well. The approach we adopted, then, was to carry out just three separate simulations, each fit separately to a set of findings from an ensemble of related experiments.

In general, the model was allowed to fit the data using different parameters in each of the three simulations. However, it turned out to be possible to use most of the same parameter values for both the first and the second simulations. In what follows, then, we report the complete set of parameters for the first and third simulations. For the second simulation, we present only those parameters that differed between the first and the second simulations.

Simulations

Overview

We have addressed the empirical phenomena described previously through three main simulation experiments. The first addresses the strength-mirror, list-length, and null list-strength effects. This simulation targets the data from the relatively homogeneous set of experiments by Murnane and Shiffrin (1991a). Although these investigators did not consider z-ROC slopes in their experiments, our reading of the literature suggested that a slope of 0.8 would be expected, and so we included this value among the "data points" that the parameters are optimized to fit. The results of this simulation were also used to determine whether the model produced cumulative confidence data conforming to the pattern found for human participants by Balakrishnan and Ratcliff (1996).

Our second simulation addresses the effects of repeated exposures to an item on false-recognition to similar distractors, as studied by Hintzman et al. (1992) and Hintzman and Curran (1995). They presented a number of experiments, but not with the homogeneity of conditions that characterized the Murnane and Shiffrin (1991a) experiments. We have therefore chosen one of their experiments that seemed representative, involved a large number of participants, and was closest to a recognition experiment, the "neutral" condition in Hintzman and Curran's (1995) Experiment 1. Because all of the experiments in this series showed an initial increase in false recognition up to three presentations of studied items then a decrease in the false recognition rate from three to eight presentations, we required this also from the model. Besides the false recognition behavior, this data constrains the model in a regime where the Murnane and Shiffrin (1991a) data did not, testing the model's handling of asymptotic hit and false alarm rates for very large numbers of study trials.

Our third simulation addresses the word-frequency mirror effect. Rather than consider this effect by itself, we chose to simulate the results of an experiment that simultaneously considered the strength-mirror and list-strength effects, their possible interaction with word frequency, and their influence on z-ROC slope and intercept (the latter is closely related to d'). In the experiment in question (Experiment 4 of Ratcliff et al., 1994), a frequency manipulation is crossed with a strength/list-strength manipulation, and three levels of confidence are used with both *old* and *new* responses, thus allowing the analysis of effects on z-ROC parameters. The simulation of this experiment also provides a context for a consideration of the basic frequency-mirror effect and the extent of asymmetry in this effect.

Differentiation, the Strength-Mirror Effect, and the Null List-Strength Effect

The strength-mirror, list-strength, and list-length effects are all clearly exemplified in the experiments of Murnane and Shiffrin (1991a). We have targeted the average results across several experiments for our simulation because these experiments produce relatively homogeneous results, the qualitative patterns are generally consistent, and the average over all of the experiments produces hit and false-alarm rates for the various conditions that are quite close to the data of one of the individual experiments (Experiment 4_{eq} in Table 2). The data from each of the experiments and the averages derived from them for use in the fitting process are shown in Tables 2, 3, and 4. For the list-strength simulations, there were 50 unique items in each list. Weak items were presented once, and strong items were presented three times. For the list-length simulation, items were presented once each, and there were 50 unique items in the short list and 150 in the long list. (In the simulation, the short list in the list-length simulation is the same as the pure weak list in the list-strength simulation, so the same results appear in each of these cases. Note that in the simulated experiments, some but not all of the data is common to these two conditions).

We attempted to fit the mean hit and false-alarm rate data from the various conditions of the experiments with $\delta = 0.03$; as explained above, this is achieved when the simulated mean is within $\delta/2$, or 0.015, of the data for each simulated data point. The best fit we obtained is shown in Tables 6 and 8 and is based on the parameter values shown in Table 7. The desired fit was achieved for all data points but the mixed-weak hit rate, which is just over .03 smaller in the experimental data than in the simulation.

Table 6 shows the list-strength simulation fits. Differentiation is in evidence; false alarm rates decline as one goes from the

Table 6
List-Strength Simulation

	Pure weak				Mixed weak			
Source	HR	FAR	d'	$\ln[\beta]$	HR	FAR	d'	$\ln[\beta]$
Sim.	0.74	0.20	1.56	0.17	0.731	0.16	1.68	0.34
Data	0.75	0.20	1.52	0.14	0.700	0.17	1.48	0.32
	Mixed strong				Pure strong			
Sim.	0.84	0.16	2.07	0	0.83	0.12	2.20	0.22
Data	0.85	0.15	2.09	0	0.84	0.13	2.12	0.14

Note. The data values for d' and $\ln[\beta]$ were calculated from the overall HRs and FARs for the corresponding conditions and differ slightly from the mean d' and $\ln[\beta]$ values calculated over experiments as reported in Table 2. Sim. = simulation; HR = hit rate; FAR = false alarm rate.

pure weak to the pure strong condition. In fact, differentiation is also evidenced by the fact that false alarm rates decline as one goes from pure weak to mixed weak, and from mixed strong to pure strong. The reason is that each of these comparisons involves an increase in the fraction of the studied items that are strong, and strong item detectors are less likely to be strongly activated by new items. Also, the $\ln[\beta]$ values increase as one goes to the stronger condition, in agreement with the data. Note, however, that in the simulations it is the distributions that are moving rather than the criterion, which has a fixed value.

In the simulation, there are very tiny effects of list-strength on d', consistent with the null list-strength effect. Strengthening the μth item will just slightly reduce $\lambda^{\mu}(S)$ when S is generated by another item, for example, a weak or new item. This is due to differentiation. Because both mixed weak and new items are affected, the effect on d' is small and in fact its direction depends on the parameters. In fact, enough replications of the simulation can generally be run to establish a reliable list-strength effect, but in our experience with a wide range of parameter values the effect is generally very small, and most experiments would not have the power to detect an effect of the predicted size.

The discrepancy between experimental and simulated mixed weak hit rates deserves some consideration. The discrepancy reflects the fact that in the experimental data the hit rate drops more from pure weak to mixed weak than from mixed strong to pure strong, but in the simulation the drop is about the same. The difference between pure weak and mixed weak hit rates varied widely across the experiments reported in Tables 2 and 3 (hit rate differences, pure weak − mixed weak, ranged from 20% to −6%), suggesting that the effect in the data might be unreliable.

List-Length Effect

Table 8 shows that the model captures the list-length effect as it appears in the Murnane and Shiffrin (1991a) experiments. In understanding the processes at work here, it should first be noted that the distributions for new and old items are composed differently. Recall that the decision is based on the maximum estimated log-odds across the detectors (see *Simulation of a Participant's Performance in an Experiment,* above). For new items, this is the maximum across L identical distributions. For old items, one detector will correspond to the test item, so that its distribution will have a larger mean than the $L - 1$ distributions for the other detectors. On most occasions, the maximum estimated log odds will be that of this corresponding detector, although not always.

Table 7
Parameters for List-Length and List-Strength Simulations

Parameter	Value
N	73
κ	0.29
f	0.25
p_0	.0005
p_1	.86
σ_n	0.75
ϵ	0.82
ω_C	−2.51

Note. N = number of elements in each stimulus vector; κ = probability that weight is able to learn; f = fraction of elements having high probability of being active in an item; p_0 = probability of being active, for low probability elements of an item; p_1 = probability of being active, for high probability elements of an item; σ_n = standard deviation of normal distribution generating initial weights; ϵ = initial value of the learning rate parameter; ω_C = criterion value of estimated log-odds at each detector.

Table 8
List-Length Simulation

	Long list				Short list			
Source	HR	FAR	d'	$\ln[\beta]$	HR	FAR	d'	$\ln[\beta]$
Sim.	0.69	0.27	1.17	0.08	0.74	0.20	1.56	0.17
Data	0.70	0.27	1.14	0.05	0.73	0.20	1.45	0.17

Note. The data values for d' and $\ln[\beta]$ were calculated from the overall HR and FAR for the corresponding conditions and differ slightly from the mean d' and $\ln[\beta]$ values calculated over experiments as reported in Tables 6 and 2. Sim. = simulation; HR = hit rate; FAR = false alarm rate.

Operating on these distributions are two opposing factors. One is the $\rho(list)/[\Lambda - \rho(list)]$ term in Equation 6, which will reduce the estimated log-odds by about one third in going from the short to the long list. It may seem surprising that this change does not more dramatically reduce hit and false alarm rates. But in fact there is very large variance in these ratios. The ratios vary from about −8 to 10 for old pure weak items (see distributions in Figure 4 below). A factor of one third corresponds to subtracting 1.1 from the ratios, which has a relatively small effect given the large variance.

The other factor opposing this change is that there are three times as many detectors in the long condition. This second factor has a different sized effect in the case of hits and false alarms. For hits, the decision is only occasionally affected by detectors other than the one corresponding to the test item, so that the total number of detectors does not have a large effect. In this case, the first factor tends to dominate, and the hit rate goes down. On the other hand, the number of detectors has much more of an effect on the false alarm rate, which is based on the maximum of L identical distributions. We see in the simulation that the tripling of the number of detectors dominates the effect of subtracting 1.1 from all the estimated log-odds ratios, resulting in an increase in the false alarm rate.

z-ROC Curves

Because of a large amount of data indicating that z-ROC curves should have slopes close to 0.8, and because the d's for the Murnane and Shiffrin (1991a) experiments are in a similar range to the experiments showing that, we considered the model's ability to fit z-ROC slopes of 0.8 in the first simulation.

Figure 4 shows z-ROC curves derived from the model on the basis of 300 replications. They were obtained by varying ω_C, the criterion at each detector, from −4.2 to 0. All other parame-

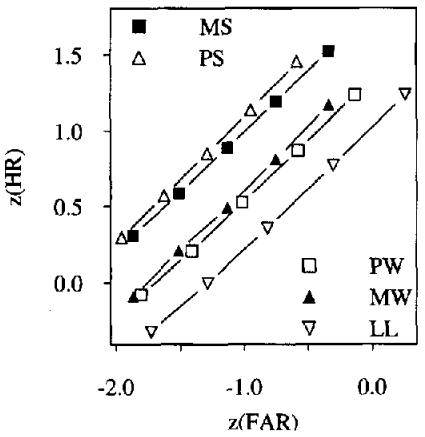

Figure 4. z-ROC curves for the list-strength and list-length simulations. PW = pure weak; MW = mixed weak; MS = mixed strong; PS = pure strong; LL = long list; HR = hit rate; FAR = false alarm rate.

ters were kept the same as in the Murnane and Shiffrin (1991a) simulations. Note that the same curves could have been obtained by varying $\rho(list)$, which would correspond to the experimental manipulation of varying the proportions of old and new items. The slopes were pure weak, 0.79, mixed weak, 0.82, mixed strong, 0.79, pure strong, 0.84, and long list, 0.79. Thus, strengthening and lengthening are not substantially affecting z-ROC slope with this parameter set.

As we noted in the model description, the recognition decision is effectively based upon whether or not the maximum estimated log-odds across the detectors is above criterion. Thus, this quantity is analogous to the familiarity of the item, which can be used as the decision variable in accord with the principles of signal detection theory. Their shapes are therefore of interest. Figure 5 shows the distributions of the maximum estimated log-odds ratios for the pure weak condition. The ratio of the standard deviations is 0.65, which is somewhat different from the observed slope of the z-ROC curve. The discrepancy is due to the fact that the distributions deviate slightly from normal.

Similarity and Strengthening

Figure 1 shows the results of simulating Hintzman and Curran's (1995) Experiment 1 (here we simulate JOF = 0 as equivalent to the model's standard *no* responses, and JOF > 0 as equivalent to the model's *yes* responses). The value of δ adopted for the simulation of this experiment was .04, so that simulated JOF > 0 rates for old and similar test items should have fallen within ± 2% of the experimental values. The use of a slightly larger value of δ seemed appropriate here because the data came from a single experiment.

Similarity was modeled by generating pairs of vectors so that if one has a p_1 unit, the other one also does with probability η (0.765). Otherwise, the parameters were the same as for the simulation above, except for the number of units (N = 138), the initial learning rate (ϵ = 0.42), and the criterion used (ω_C = −2.2).

In agreement with the data, false recognition first increases and then decreases with repetition of the studied item. The simulation actually exhibits a very gradual decrease, which tends to slow to a stop when more than 10 or so repetitions have occurred, similar to the effect found also in other studies of this type reported by Hintzman and Curran (1995). Thus, the simulation captures these important qualitative trends in the data. The fit is also quantitatively quite close to the data, coming within $\delta/2$ of the experimental results, with the exception of the hit rate for items presented 3 times, where the model's performance exceeds that of the participants by about 9%.

To gain some insight into the pattern of false recognition of similar distractors, and to see why the model does not achieve a fuller ultimate degree of differentiation, we begin by considering the effect of learning a specific item on a single trial on the estimated log-likelihood (which is linearly related to the estimated log odds) for test items as a function of their similarity to the item learned. Figure 6 shows how this varies as a joint function of η, similarity of the test item to the learned item, and s, the number of learning trials. At all levels of similarity shown, the estimated log-likelihood first increases and then decreases with learning. For higher similarities more learning is required before this downturn occurs.

We might expect that differentiation as a function of frequency of presentation should follow the theoretical curves shown in Figure 6. If this were so, we would expect that with sufficient trials there would always be perfect differentiation. In principle, one feature of miss–match is enough to send the estimated likelihood ratio to 0 if the detector's estimate of the probability of the value of the feature is 1.0. In this case ω will be 1 for features whose value is 1, and 0 for features whose value is 0. However, two factors not illustrated in Figure 6 mitigate against this. One is the gradual decrementing of the learning rate. This reduction in learning rate has a minor effect compared to the second factor, which is the presence of random variation in the representation of each stimulus. This variability means that crucial differentiating features will have estimated probabilities that are not extremal, thereby allowing items that differ on a small number of features to produce respectable partial matches (moderate estimated log-likelihood ratios).

Interestingly, Hintzman and Curran (1995) have found that participants can achieve perfect differentiation, but only if they

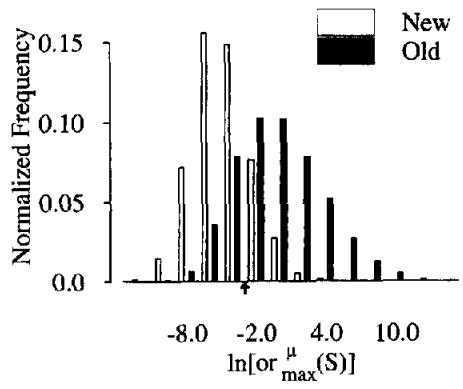

Figure 5. Distributions of maximum estimated log-odds ratios for new and old items in a pure weak condition.

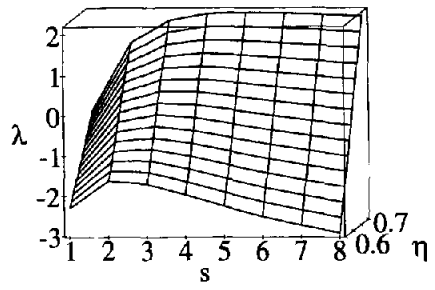

Figure 6. Estimated log-likelihood ratio λ as a function of degree of learning based on a single presentation of an item, and similarity of a test item to the studied item. The abscissa s represents the number of learning trials; the ordinate η represents the similarity of the test item to the studied item.

are forced (rather than simply urged) on a trial-by-trial basis to make explicit note of the features of a stimulus item that differentiate it from similar distractors. In their case, stimulus words were singular or plural and the similar distractor was the plural if the stimulus was singular or vice versa. When participants were required to indicate through the use of a button-press whether a fragment (e.g., *bel . . .*) of a stimulus should be singular or plural (e.g., *bell* or *bells*), they did learn to completely differentiate repeated stimuli from similar distractors. This could be modeled within our framework by assuming that Hintzman and Curran's manipulation eliminates the variability in the encoding of the relevant features. For example, p_t for the plural feature would then be either 1 or very close to it.

Effects of Word Frequency

In previous sections, we have considered how item strengthening or repetition can produce mirror effects. Here, we consider the characteristics of particular items themselves. Because pre-experimental word frequency has been extensively studied, we focus on that variable. As previously discussed, words of lower frequency produce a higher d' and there is a mirror effect, such that hit rates are higher and false alarm rates are lower for low-frequency words than for high-frequency words. In addition, the value of $\ln[\beta]$ is generally larger for low-frequency words. This is often accompanied by an asymmetry in the effect of frequency on hit and false alarm rates, the effect being larger for false alarms. A third empirical observation, also mentioned previously, is the fact that z-ROC slopes appear to be shallower for low-frequency words.

In an attempt to understand these data, it is important to bear in mind that the mirror effect occurs even when the presentation of high- and low-frequency items is mixed at both study and test. This presents a particular challenge in understanding why the false alarm rate is lower for low-frequency words than for high-frequency words. One might be tempted, for example, to suppose that, at study, participants allocate more attention to low-frequency words than to high-frequency words. This could explain why hit rates are greater for low-frequency words than for high-frequency words, but by itself it does not explain why the false alarm rates for new low-frequency words are lower than false alarm rates to new high-frequency words because, at test, both high- and low-frequency words are compared to detectors for all of the studied items. Somehow the new low-frequency items produce fewer old responses than new high-frequency items, and because they were not previously studied, this effect cannot be attributed to differences in encoding processes occurring at the time of study alone.

One approach to this issue, advocated by Glanzer et al. (1993), is to suggest that participants may attend to some features of an item but not others. Word frequency is assumed to affect the fraction of features attended, with the fraction (α) assumed to be greater for words of low frequency. In Glanzer et al.'s theory, this attention difference applies both at the time of study and at the time of test; crucially, then, at test, a larger fraction of features are sampled for low-frequency words.

Our approach involves the possibility that high- and low-frequency words differ in terms of their featural content. This is similar to the approach taken by Shiffrin and Steyvers (1997) (see General Discussion). In our approach to this matter, we examine the possibility that representations of low-frequency words are less variable than representations of high-frequency words. Support for this possibility comes from the fact that there are far more definitions of high-frequency words than low-frequency words (often the various definitions are related; Hoeffner, 1996). This assumption is captured by setting the parameter p_0 to a higher value for high-frequency words. This means that high-frequency words effectively have greater noise in their representation; low firing units fire with a slightly higher probability giving a larger number of "on-bits" that tend to be inconsistent across presentations of the same item.

To assess the model's adequacy to account for the effects of word frequency, we have chosen to fit the data from Experiment 4 of Ratcliff et al. (1994). This experiment was chosen for several reasons. First, the frequency manipulation was crossed with a strength/list-strength manipulation, and participants were asked for confidence ratings, so that several phenomena and their interplay were all assessed at once. Second, the pattern of effects obtained in the Ratcliff et al. experiment are consistent with those reported in other studies. Table 9 shows the experimental data, which are reported in terms of slope and intercept of the z-ROC curves. The z-ROC intercept is a measure of sensitivity closely related to d'. The effects of both strength and word frequency on the z-ROC intercept value were reliable: The experiment thus captured the expected effects of both frequency and strength on this variable. Considering z-ROC slope, there was a reliable overall effect of frequency, with lower slopes for low-frequency than high-frequency words. There were no reliable effects of strength or list strength on this variable. The experiment also produces a frequency-mirror effect that is slightly asymmetrical. This can be seen in the hit and false alarm data collapsed over confidence categories and is shown for the pure weak strength condition in Table 11 later in the article. Third, the authors report standard errors for both the slope and the intercept of their z-ROC curves, allowing a close assessment of the model's fit to the various aspects of the empirical data. As shown in the table, the confidence intervals are relatively wide for several of these data points. This may be responsible for some of the differences between individual conditions. In particular, the differences in intercept between corre-

Table 9
Simulation of Frequency-Mirror × Strength-Mirror Experiment

Condition	Slope			Intercept		
	M_{exp}	$CI\pm$	M_{sim}	M_{exp}	$CI\pm$	M_{sim}
HF						
Pure weak	0.801	0.091	0.82	0.566	0.119	0.69
Mixed weak	0.765	0.078	0.81	0.631	0.105	0.68
Mixed strong	0.716	0.080	0.72	0.950	0.114	0.93
Pure strong	0.825	0.103	0.74	0.879	0.122	0.95
LF						
Pure weak	0.719	0.094	0.68	1.091	0.127	1.05
Mixed weak	0.734	0.083	0.72	1.072	0.114	1.06
Mixed strong	0.703	0.097	0.63	1.550	0.127	1.42
Pure strong	0.693	0.097	0.63	1.291	0.130	1.41

Note. The 99% confidence intervals were obtained by multiplying Ratcliff et al.'s (1994) standard errors by 2771 = $t(27)$, $p < .01$, two-tailed. Data from Experiment 4 of "Empirical Generality of Data From Recognition Memory Receiver-Operating Characteristic Functions and Implications for the Global Memory Models" by R. Ratcliff, G. McKoon, and M. Tindall, 1994, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, p. 772. Copyright 1994 by American Psychological Association. Adapted with permission of the authors. HF = high frequency; LF = low frequency; M_{exp} = experimental mean; $CI\pm$ = half the width of the 99% experimental confidence interval, M_{sim} = simulation mean.

sponding pure and mixed conditions are likely to be spurious, given their lack of systematicity and the massive documentation of the null list-strength effect.

Our simulations focused initially on fitting the *z*-ROC statistics presented in Table 9. In accordance with Ratcliff et al.'s (1994) experiment, study lists had 32 items, half high frequency and half low frequency. Strength was manipulated in the experiment by varying the length of study of items; we model this by having different learning rates, ϵ_w and ϵ_s, for strong and weak items. As already discussed, word frequency was assumed to affect the p_0 parameter, so that we require two values, one for high frequency words (p_{0h}) and one for words of low frequency (p_{0l}).

In the model the overall rate of occurrence of active features is higher for high-frequency words because of the higher value of p_0. This raises an issue in our model; what value of ρ_a should we use at test in Equation 4 and for the average value of our initial weights? In the simulations reported here we use the "average" ρ_a,

$$\rho_a = \tfrac{1}{2}\,[fp_1 + (1-f)p_{0h}] + \tfrac{1}{2}\,[fp_1 + (1-f)p_{0l}],$$

although it is not clear that the two frequency classes should be given equal weight.

The fitting process was allowed to find new values of all of the parameters, and Table 10 shows the parameters found in fitting the data from this experiment. Table 9 shows the simulation results for comparison to the experimental data. With one exception, to be discussed below, the fitted *z*-ROC slopes and intercepts are within the experimental confidence intervals.

The results obtained by the fitting process show all the effects that are reliable in the experiment. Specifically, the model produces larger *z*-ROC intercepts for strong than for weak items and larger intercepts for low- than for high-frequency words. The model also captures the experimental finding of smaller *z*-ROC slopes for low-frequency words relative to high-frequency words (mean slopes were .665 vs .773 for the simulation and .712 vs .777 for the experiment). Furthermore, the model appears to capture the null list-strength effect, both for the *z*-ROC slopes and intercepts. In the simulations, the effects of list-strength (pure vs. mixed list) on *z*-ROC intercept are minuscule and, likewise, there are no consistent effects of list-strength on *z*-ROC slopes.

The simulations deviate from the experiments in two ways: First, the simulation shows a clear effect of item strength on *z*-ROC slope, contrary to the experiment. Second, the simulated intercept for the high-frequency pure weak condition falls outside the experimental confidence interval. Considering the second point first, it is quite possible that the intercept in this condition is low because of random fluctuation. The value is below the value for the corresponding mixed weak condition, and the generality of the null list-strength effect suggests that the two values should be very nearly the same. (At least 1 out of 16 experimentally determined data points should fall outside of the 99% confidence interval around its true expected value with a probability of .15, so a single discrepancy of this magnitude is not terribly surprising.)

The model's prediction that item strength will affect *z*-ROC slopes is perhaps more problematic, because with the current parameters at least the expected effect is large enough (.67 for weak items vs .76 for strong) that it might have been detectable. Clearly, this matter deserves careful consideration, because it represents a potential empirical shortcoming of the model. We will hold further discussion of this point until a later section. Figures 7 and 8 show the simulation *z*-ROC curves for all conditions.

Frequency-Mirror Effect: Is There an Asymmetry?

Although our simulation of the findings of Ratcliff et al. (1994) has concentrated on the effects of preexperimental fre-

Table 10
Parameters for Frequency × Strength Simulations

Parameter	Value
N	313
κ	0.32
f	0.04
p_{0h}	.03
p_{0l}	.001
p_1	.96
σ_n	0.405
ϵ_w	0.09
ϵ_s	0.145

Note. N = number of elements in each stimulus vector; κ = probability that weight is able to learn; f = fraction of elements having high probability of being active in an item; p_{0h} = probability of being active, for low probability elements of high-frequency words; p_{0l} = corresponding value for low-frequency items; p_1 = probability of being active, for high probability elements of an item; σ_n = standard deviation of normal distribution generating initial weights; ϵ_w = initial learning rate for weak items; ϵ_s = initial learning rate for strong items.

quency and strength on the z-ROC slope and intercept, most other studies of the frequency-mirror effect have focused on basic hit rate and false alarm rate data, noting that for low-frequency words, hit rate tends to be higher and false alarm rate to be lower than for high-frequency words. As noted in our review of data, many studies find that the frequency-mirror effect is not quite symmetric: The effect on hit rate is usually somewhat smaller than the effect on false alarm rate, and there is a trend toward a higher (more stringent) criterion in the low-frequency condition. Analysis of the results from Ratcliff et al. (1994) from this point of view, however, shows a relatively slight effect. For example, in the pure weak condition of the Ratcliff et al. experiment, collapsing over confidence categories, overall hit rates are 8.4% higher for low- than high-frequency words, false alarm rates are 9.5% lower, and the value of $\ln[\beta]$ is only slightly larger for the low-frequency words, as shown in Table 11. We assessed the frequency-mirror effect in our simulation, using the same parameters as for the main simulation of z-ROC slopes and intercepts and found very similar effects, with two different values of the criterion ω_C chosen to approximately bracket the range of the existing experimental data. There is a slight asymmetry in the expected direction. For the more stringent value of the criterion, the effect is not apparent in the hit and false alarm rates, but it is evident in $\ln[\beta]$. In this case, the criterion is well into the tails of the distributions of values produced by new stimuli, causing a compression in the effect of movement of these distributions on raw false-alarm rates.

Cumulative Confidence Curves

As previously discussed, Balakrishnan and Ratcliff (1996) considered a list-strength experiment where participants made confidence judgements about test items. They showed that under certain optimal or ideal-observer models, the cumulative confidence ratings curves for certain conditions should cross, as illustrated in Figure 2. However, in their experiment they found no evidence of the expected tendency for the curves to cross. Instead, they found the pattern presented in Figure 9. The experiment actually included three strength levels—weak (one presentation of 250 ms), moderate (two such presentations), and strong (four presentations)—and two mixed conditions—weak–moderate and weak–strong. We omitted the moderate condition because there were no special effects of this condition. Balakrishnan and Ratcliff stated the following: "Instead of following any crossover pattern, a much better description of these results is that the functions are not very different and they are ordered" (p. 625). As Figure 9 illustrates, the curves for new items are ordered, except that the curves for mixed weak and pure strong false-alarm rates tend to merge. For old items, the curves are clearly ordered by item strength. A good description of the relation between curves for items of similar strength but differing in list strength would be that they are nearly indistinguishable. (Note that the pure weak hit rate curve leads the mixed weak hit rate curve at both extremes, and corresponding curves never differ from each other by more than 4%.)

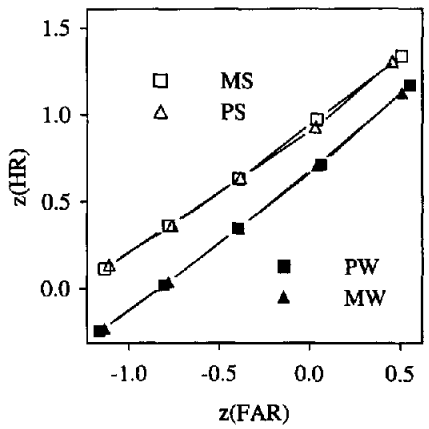

Figure 7. z-ROC curves for high-frequency words. MS = mixed strong; PS = pure strong; PW = pure weak; MW = mixed weak; HR = hit rate; FAR = false alarm rate.

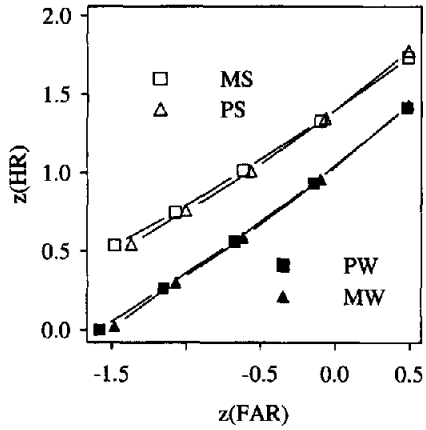

Figure 8. z-ROC curves for low-frequency words. MS = mixed strong; PS = pure strong; PW = pure weak; MW = mixed weak; HR = hit rate; FAR = false alarm rate.

Figure 10 shows such curves for our model, using parameters from our first simulation of the strength, list-strength, and list-length effects in the data of Murnane and Shiffrin (1991a). We emphasize that no additional parameter manipulations have been made specifically to fit the details of Balakrishnan and Ratcliff's (1996) results. We have simply run new simulations using the parameters from our first simulation to look at the cumulative confidence curves. In each condition, 50,000 trials were run to increase the precision of the predicted curves. The curves are formed from the simulated distributions of estimated log-odds ratios for each of the conditions and assume that participants place their criteria for determining response choice and confidence at points along the subjective likelihood continuum that remain fixed across all conditions of the experiment.

The simulated curves do not show the pattern of crossing illustrated in Figure 2, but instead conform closely to the pattern found in the empirical curves obtained in Balakrishnan and Ratcliff (1996) and shown in Figure 9. The actual shapes of the simulated and experimental curves should not be compared, because the experimental ratings categories are not likely to span equal intervals along the subjective likelihood continuum. The degree of separation of the group of curves for new items

Table 11
Frequency-Mirror Effect

	Low frequency			High frequency			Difference		
Criterion	HR	FAR	ln[β]	HR	FAR	ln[β]	HR	FAR	ln[β]
Data from pure weak condition of Ratcliff, McKoon, and Tindall (1994), Experiment 4									
	0.623	0.194	0.32	0.534	0.292	0.15	0.089	−0.098	0.17
Simulations with two different criterion values									
−1.7	0.702	0.233	.12	0.618	0.328	0.06	0.084	−0.095	0.06
−1.4	0.623	0.142	.52	0.531	0.232	0.27	0.092	−0.090	0.25

Note. Difference is obtained by subtracting the HR or FAR for high-frequency words from the corresponding quantity for low-frequency words. HR = hit rate; FAR = false alarm rate.

from the two groups of curves for old items and the degree of separation of the curves for weak–old items from the curves for strong–old items should not be compared either, because these effects depend on the details of the strength conditions, which apparently differed between the Murnane and Shiffrin (1991a) and the Balakrishnan and Ratcliff experiments. However, the ordering of the curves and their relations to each other within groups can be compared. Within the new items, the functions are ordered pure strong, mixed, and pure weak, as in the experiment. Also as in the experiment, the functions appear to converge at the top. Within the old items, the functions are ordered by item strength, with weak rising before strong. Curves

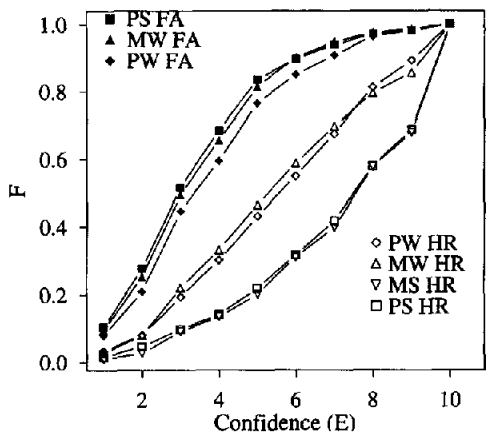

Figure 9. Cumulative frequency distributions of confidence ratings by condition in the recognition memory experiment of Balakrishnan and Ratcliff (1996). Ratings ranged from 1 (*most-sure–new*) to 10 (*most-sure–old*), with the boundary between old and new placed between Categories 5 and 6. F = frequency; PS = pure strong; MS = mixed strong; MW = mixed weak; PW = pure weak; FA = false alarm; HR = hit rate. Printed from raw data supplied by J. D. Balakrishnan, March 9, 1996. Most, but not all, of the data shown here were presented in Figure 8 of "Testing Models of Decision Making Using Confidence Ratings in Classification" by J. D. Balakrishnan and R. Ratcliff, 1996, *Journal of Experimental Psychology: Human Perception and Performance, 22*, p. 624. Copyright 1996 by American Psychological Association. Adapted with permission of the authors.

for different list contexts but the same item strength are virtually identical, again as in the experiment. There appears to be a slight discrepancy between data and experiment at the low end of the curves for new items, such that at the low end, the curves are more separated in the simulation than in the data. We return to this point in the next section where we consider discrepancies between the model's behavior and the experimental data.

Our model differs from the models considered by Balakrishnan and Ratcliff (1996) in that the likelihood ratios computed by the detectors in our model are not the same as the likelihood ratios that must be computed to perform optimally in the sense considered by Balakrishnan and Ratcliff. In particular, the likelihood ratios used in our model do not take into account the distribution of likelihoods associated with other test items in the same test list. Instead, these ratios reflect the relative likelihood that the presented test item corresponds to a particular detector versus the generic "item" in which each feature occurs with probability ρ_a. This relative likelihood is assumed to correspond to the sense of familiarity associated with the particular item. The model assumes that participants produce confidence rating data by placing multiple criteria along the continuum of values associated with this variable and do not adjust these criteria to changes in the distributions of these values in different experimental conditions. The effects of strength of studied items on cumulative confidence curves are due to effects of strengthening on the responses of the detectors.

Although the cumulative confidence curves in our model may still potentially cross, this would be due directly to the shapes of the distributions of $\lambda^{\mu}(S)$ produced by the operation of the detectors and not to the use of optimal decision criteria in the sense considered by Balakrishnan and Ratcliff (1996). In our simulations, we have found that the curves can cross where they converge at the very high end, but the effect is too far into the tail to be detectable experimentally, yielding the pattern seen in Figure 9, in which the curves appear to converge rather than to cross.

Balakrishnan and Ratcliff (1996) suggested that the pattern of results that they found empirically—in which cumulative confidence curves are said to shift from condition to condition—can be accounted for by assuming that participants adopt response and confidence criteria that retain a fixed spacing on a

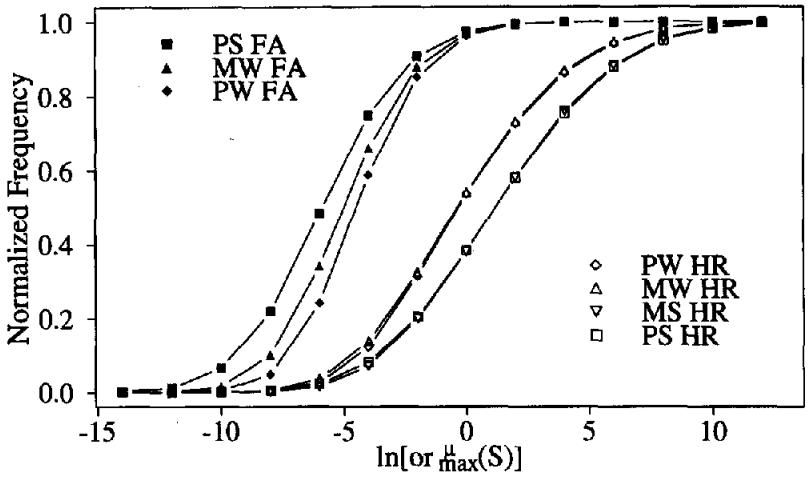

Figure 10. Cumulative curves for the distributions of log-odds for old and new stimuli from tests of recognition memory for pure and mixed lists of strong and weak items. Parameters used were those from the simulation of Murnane and Shiffrin (1991a), presented in Table 7. PS = pure strong; MW = mixed weak; PW = pure weak; FA = false alarm; MS = mixed strong; HR = hit rate.

subjective continuum across all strengthened list-strength conditions. To account for the effects that strength and list strength have on these curves, this approach assumes that participants shift the entire set of criteria along the subjective continuum by an amount that depends on the ensemble of item strengths. Thus, to account for the cumulative curves for the new item data as a function of the strength of items on the list, this approach assumes that the criteria are placed lowest for the pure weak, highest from the pure strong, and at an intermediate location for the mixed condition. This contrasts with our assumption that the criteria stay completely fixed and that it is the distributions of psychological states produced by old and new stimuli that are affected directly by the strength of the items on the list.

These two models make different predictions for the effects of list context on new versus old responses. The criterion shift account predicts exactly equivalent effects on old and new items as a function of list context. For example, the shift from strong–new to mixed–new should be the same as the shift between pure-strong–old to mixed-strong–old, and the shift from mixed–new to weak–new should be the same as the shift from mixed-weak–old to pure-weak–old. In our model, the effect on old items need not be the same as the effect on new items. Indeed, the effect tends to be slightly larger for new than for old items because the list context items make more of a difference to the distribution of estimated likelihood values for new items. For old items the response of the model is strongly dominated by the detector for the item itself, so that the detectors for other items make a relatively small difference. The data seem to hint at the sort of differential effect predicted by our model, in that among the old items, the curves for different list contexts at the same level of item strength are virtually identical, whereas there appears to be slightly more separation of the curves for new items (note that for weak items, the empirical curves actually intersect at both ends, counter to the notion that there is a shift). However, the effects are obviously too small overall to take seriously. Further research will be necessary to test this subtle difference between the two approaches.

Residual Issues Concerning the Correspondence Between the Model and Experimental Data

Overall, the simulations have captured the experimental phenomena listed in Table 5 and have come quite close to capturing the specific condition means and experimental effects where this was attempted. The model did miss one data point in each of the main simulations: The mixed weak hit rate for Murnane and Shiffrin (1991a), the pure weak *z*-ROC intercept for high-frequency words in Ratcliff et al. (1994), and the proportion of JOF = 0 for three-times presented items from Hintzman and Curran (1995). Such a number of misses is not unexpected given the number of data points fitted overall, and these cases do not appear to be particularly systematic. Indeed, in the first two of the three cases, the data points being fit seem to deviate from the general pattern that characterizes the overall body of relevant data. Of course, further study may reveal that one or more of these discrepancies is telling us something about a way in which the model may be inaccurate, or that the model is incorrect in some other completely different way.

One place to look carefully for other signs of mismatch between the model and the data is in various trends seen in the model that may not occur in the data or vice versa. Here we consider three cases where the degree of correspondence between trends in the model and data is not completely clear.

Effects of strength on z-ROC slope. With most values of the parameters, the model tends to predict that factors affecting d' or *z*-ROC intercept will also affect *z*-ROC slope: *z*-ROC slope tends to decrease as d' increases. To a large extent, the empirical data bear out this relationship: (a) Several studies (though not all) report that *z*-ROC slopes are lower for shorter lists than for longer ones. Some of the studies that do not report slope effects showed many lists of differing lengths to the same participants,

possibly contributing to a washout of effects of list length on z-ROC slope in these cases. (b) Several studies report that z-ROC slopes are lower for low-frequency than for high-frequency words, consistent with higher d' and z-ROC intercepts for words of lower frequency. (c) When items of extremely low strength are used (on the basis of very brief durations) so that d' is very small, z-ROC slopes are very close to 1, whereas with more typical strengths ($d' \geq 0.5$) slopes are definitely somewhat lower. The one difficulty is the inconsistency of experimental results concerning the effects of item strength on z-ROC slope, once d' exceeds 0.5. Although many studies find such effects, there are several studies that show no effects of item strength when d' is investigated in the range between about 0.5 to 3.0 or so (see Glanzer et al., in press, for new data and a thorough review). On the basis of current evidence, it is difficult to tell whether the absence of slope changes with d' changes in some studies is a significant problem for our model. Two points seem worth mentioning in this connection.

First, the relatively straight lines obtained in many group z-ROC curves and the relatively homogeneous z-ROC slopes across conditions may obscure facts that are only apparent in those rare experiments where sufficient data are obtained to plot z-ROC curves for each condition for each participant. One important observation that emerges from one such study (Ratcliff et al., 1994, Experiment 5) is the finding that there are substantial individual differences in z-ROC slope between participants, and indeed many participants produce z-ROC slopes that are highly nonlinear. Ratcliff et al. explored the possibility that rather uninteresting factors such as occasional random guess responses might be responsible for some of these deviations from linearity. The point here is that contaminating influences that distort linearity will also distort calculated z-ROC slopes and might contribute to obscuring a small but real effect of strength on z-ROC slope. To the eye, some of the participants that produce the most linear z-ROC curves show a tendency for shallower slopes in the stronger condition, at least in the low-frequency word condition.

Second, the model does not produce large effects of strength on z-ROC slope throughout the parameter space. In fact, the effects of strength on z-ROC slope were very small and unsystematic with the parameters used for the Murnane and Shiffrin (1991a) simulations. In fitting the Ratcliff et al. (1994) results, the parameters found were such that a moderately large effect of strength on z-ROC slope was predicted, but it is possible that an overall fit about as good could be obtained if the model was forced to find parameters producing little or no effect of strength on z-ROC slope. Thus, it may be that the model is not intrinsically inconsistent with the small (undetectable) effects of strength on z-ROC slope in some experiments. Furthermore, given the large individual differences among participants seen in Experiment 5 of Ratcliff et al., it is possible that there are real individual differences in those parameters that control covariation of z-ROC slope with d', so that some participants show such effects, but others do not. If this were the case, one might expect it to be difficult to find robust evidence of slope varying with strength in a typical multiparticipant experiment.

Changes in β. The literature on the effects of experimental variables on β is somewhat murky, and there are only two effects that appear to be relatively clear: (a) The value of β increases (becomes more conservative) as list-strength increases, and (b) β also appears to increase for low-frequency words, relative to words of higher frequency. Our model captures both of these effects. These changes in β occur even though the criterion ω_C remains the same for different conditions within a given simulation. Differences between conditions reflected in the calculated values of β are due to changes in the distributions of estimated log-odds, not to shifts in the placement of the criterion.

For effects of other experimental values, such as study time, list length, and so forth, the experimental data do not present a clear picture of the direction of shift in β. Correspondingly, we noticed in our simulations that different choices of parameters affected the amount or even the direction of change in β. Further empirical investigations of effects of these variables on β need to be undertaken so that the trends can be more clearly delineated. It is only in this way that we can have a sufficient basis to consider whether the effects are systematic enough to be worth trying to trace to experimental factors that might relate to variations in the parameters of the model.

Effects of strength and list-strength on cumulative confidence curves. Our model correctly accounts for many of the features of the cumulative confidence curves reported by Balakrishnan and Ratcliff (1996). Most importantly, it captures the empirical fact that the curves do not exhibit the strong pattern of crossing predicted by some accounts. However, there is a slight discrepancy between the model and the data of Balakrishnan and Ratcliff (1996) that deserves some discussion. This is the fact that the cumulative curves for new items are relatively more separated in the simulation than in the data, especially at the low confidence end.

One possible factor that may contribute to this discrepancy is the fact that the simulations were based on parameters used in fitting the experiment of Murnane and Shiffrin (1991a) rather than parameters specifically based on the data from the experiment of Balakrishnan and Ratcliff (1996). There are differences of method between these experiments. Words were presented in sentences in Murnane and Shiffrin but were unrelated words in Balakrishnan and Ratcliff, and very brief presentations led to much lower values of d' in the latter case. What we do not know at this time is whether parameter values that would capture the observed d' values for the different conditions would also produce an improved fit to the relative shapes of the cumulative confidence curves. It is possible that in the relevant portion of the parameter space, the effect of strengthening may be primarily to shift cumulative response curves rather than change their slope. On the experimental side, we do not know whether cumulative confidence curves from experiments like those of Murnane and Shiffrin would actually produce cumulative confidence curves for new items that are separated at the bottom but converge at the top like those seen in the simulation shown in Figure 10. Further research is needed to address these points.

General Discussion

We have introduced a new model of single-item recognition based on item detectors that estimate feature probabilities. The salient property of these detectors is their tendency to compute subjective likelihood estimates that gradually differentiate as a function of experience. As each detector learns an item, it also

becomes better at rejecting other items. Using these detectors, our model succeeds in modeling various mirror effects, the list-length effect, the null list-strength effect, *z*-ROC curves in various experiments, the main findings of Balakrishnan and Ratcliff (1996) on effects of item and list strength on cumulative confidence rating curves, and the findings of Hintzman et al. (1992) regarding the effects of strengthening and similarity.

Our model is far from the final word on recognition memory, and we have already indicated several places where its fit to experimental data may not be perfect. However, it appears to us that the correspondence to the data is good enough for it to be worth considering the key aspects of the model that appear to be crucial to its successes. Some of these aspects have implications for general issues that confront information processing models of perception and memory. Our discussion begins with these matters and then proceeds to a comparison with other models, a consideration of some remaining issues, and a brief exploration of possible extensions.

Key Aspects of the Model and Implications for General Issues

Use of patterns to represent items in memory. Our model reflects an emerging trend in thinking of a memory, not simply as a node with associative connections to other nodes, but as a pattern represented by a vector of features. This assumption is shared with the theory of distributed associated memory (TODAM; Murdock & Kahana, 1993), the Minerva II model of Hintzman (1988), the memory models proposed by Grossberg (1978; Carpenter & Grossberg, 1993), the associative memory model of Chappell and Humphreys (1994), and the attention-likelihood model of Glanzer et al. (1993). This property is also shared with a number of distributed connectionist models (Hinton, McClelland, & Rumelhart, 1986; Knapp & Anderson, 1984; McClelland & Rumelhart, 1985), although these have not typically been related to single-item recognition data as in the present article, and with the matrix model of Humphreys, Bain, and Pike (1989).

The idea that an item consists of a pattern or vector of features is a prerequisite to our notion of differentiation, because differentiation concerns knowledge about the likelihoods of these features. In this connection it is interesting to compare our account of differentiation with the account used by Shiffrin et al. (1990), who described differentiation in terms of "associations" between memory images and items. Differentiation both increased the association between an item and its corresponding memory image and decreased the association between an item and other images in memory. With the introduction of the notion that items and memory images (the item detectors in our model) are composed of a set of elements, an understanding of such associations as reflecting some sort of correspondence between the elements becomes available. This conceptualization then makes accessible a broad range of theoretical possibilities, in which experiences alter the elements of the memory representation, thereby affecting the degree of correspondence between the representation and a test item. Furthermore, it is very natural to think about a wide range of factors, including the richness of an item's representation (number of nonzero features), uniqueness (number of atypical features), similarity to other items (number of shared features), and so on. Models incorporating these factors thus provide a systematic basis for exploring the effects of such variables in ways we have only begun to hint at here.

Intrinsic variability in representation and processing. A second idea at the heart of our approach is the notion that both representation and processing in memory are intrinsically variable. The notion that features are probabilistic gives rise to the idea of using conditional probabilities to represent knowledge of the features of items and to the use of Bayesian ideas about how these probabilities might be updated and used.

In addition to variability in feature representation, we found that additional intrinsic variability in processing is necessary for the model to be able to account for all of the data. We also found that the strength-mirror effect only occurs if there is some variability in the initial estimates of the conditional probabilities embodied in the detectors when they are initialized. Without this variability, relatively weakly learned items produce relatively few false alarms, and with additional learning, the standard deviation of strengthened items changes more than the mean, so that the false alarm rate tends to *increase* rather than decrease as items are strengthened. The assumption that there is initial variability in the conditional probability estimates overcomes this tendency, thereby allowing learning to lead to a reduction rather than an increase in false alarm rates.

The fact that variability shows up in so many places in our model is consistent with the idea that a certain degree of randomness is a general principle of human information processing (Ashby & Lee, 1993; McClelland, 1993). There is a very long history of using intrinsic variability in reaction time models (e.g., Ratcliff, 1978), and both Ashby (1982) and Luce (1986) have pointed out the shortcomings of models like the cascade model (McClelland, 1979) in which variability is tacked on as an afterthought rather than built into processing itself. For another case in which incorporation of intrinsic variability appears necessary to account for psychological data, see McClelland, 1991. In any case, inclusion of the relevant sources of variability allowed the model not only to exhibit the strength-mirror effect, but several other effects considered in this article as well. The null list-strength effect, the nonlinear effect of strengthening items on the false recognition of similar items, and the mirror effect simulations particularly benefited from the various sources of intrinsic variability in the operation of the detectors.

Construing familiarity in terms of estimated likelihoods. Our model follows a thread dating back to the introduction of signal detection theory, which introduced the notion that human performance must address the inherently probabilistic relation between signals and underlying events, and a consideration of the possibility that performance might bear some relation to Bayesian inference. Starting with Anderson (1990; Anderson & Milson, 1989), several investigators have adopted a Bayesian perspective in memory research, and models that follow a Bayesian approach have been introduced for a wide range of different tasks. Perhaps the recognition memory task has seen the greatest profusion of such models, including those of Anderson and Milson (1989), Glanzer et al. (1993), Shiffrin and Steyvers (1997), and our own model based on estimated likelihoods.

Most of the older models that use feature vectors to represent

items in memory do not frame their consideration of issues in item recognition with respect to the idea of estimating likelihoods, but processes at work in these models may be closely related to those at work in our model. Indeed, the learning rule we have used is closely related to the learning rule used by Grossberg (1976), as previously noted, and it may turn out that there are strong connections between our models and those he has pioneered. Similarly, it may be possible to relate the learning processes in other models such as TODAM or Minerva II to the process of updating conditional probabilities as conceived in our current approach. In this connection it might be noted that the discovery of the null list-strength effect casts doubt on many of the earlier models, including TODAM, Minerva II, and the Matrix model of Humphreys et al. (1989). It may well be that the incorporation of a subjective-likelihood-based approach into these models could overcome these empirical difficulties. Indeed this change could bring some of these models, particularly Minerva II, into fairly close alignment with the model we have presented here.

One reaction to likelihood-based approaches to modeling memory has come from Hintzman (1994, 1997), who has asked whether human participants could possibly be as sophisticated as the models appear to require and whether they could possibility have access to the statistical quantities that the models appear to depend on. He suggested that relatively simple models might provide just as good an account as the Bayesian models. Specifically, he suggested a model in which participants do apparently simple things, such as register the occurrence of features in presentations of items and weight the effects of matches and mismatches between the features of a test item and the features of a stored item by the number of times the feature was registered during study.

We have considerable sympathy with Hintzman's position, and indeed, we have stressed that our model uses quantities that serve as *estimates* of statistical quantities. We do not assume that participants have access to the true values of these quantities, or even that they experience them as probabilistic quantities per se. Furthermore, in designing the model, we made an effort to avoid requiring participants to have detailed background knowledge of such quantities. Although the model has various parameters, these are in most cases properties of items and item detectors rather than statistical quantities we imagine that participants use in calculating likelihoods. We do assume participants have access to an estimate of the overall probability of feature occurrence, ρ_a, as well as an estimate of the list length Λ and of the probability that a test item is from the study list, $\rho(list)$. We fixed the values of the estimates of these quantities in the simulations to their actual values, but we did this only to reduce the number of free parameters and not because we believe participants have access to the true values. Inaccurate estimates of these parameters would be largely undetectable as long as they stayed constant throughout the set of conditions contrasted in a given simulation because they will mostly affect the value of the criterion parameter, Ω_C.

The model does assume that participants keep track of the values of many variables that serve as estimates of probabilistic quantities, namely the ρ_i^μ. These quantities serve as estimates of the probability of the occurrence of particular features in particular items. However, the model explicitly assumes that initial estimates of these quantities are noisy and uses a very simple procedure for adjusting the estimates. If the feature is observed in an item, the estimate of its probability is increased, and if the feature is not observed the estimate of its probability is decreased. The size of the change is proportional to the difference between the observed value of the feature in the current input and the existing estimate of its probability:

$$\rho_i^\mu(t+1) = \rho_i^\mu(t) + \epsilon[S_i(t) - \rho_i^\mu(t)]. \quad (11)$$

This is not such a complex procedure and is in fact exploited in many connectionist models. A slight complication arises from the gradual reduction of ϵ as a function of number of presentations of a particular item. As we suggested above, the reduction of ϵ might correspond to a gradual waning of a novelty reaction as a function of repeated presentation.

It is also worth considering the sophistication and complexity of the computations that must be performed at test to compute the likelihood ratio that a test item matches a particular detector. The model assumes that participants use the individual feature likelihood ratios in calculating item likelihood ratios:

$$\lambda^\mu(\mathbf{S}) = \sum_i \left[S_i \ln\left(\frac{\rho_i^\mu}{\rho_a}\right) + (1 - S_i) \ln\left(\frac{1 - \rho_i^\mu}{1 - \rho_a}\right)\right]. \quad (12)$$

The calculation may seem complex, but it can be described simply as follows: The "match" of a test item to a stored item is a sum over all of the features of the item representation. For each feature, if the feature is present in the test item, we add to the match a number reflecting the estimated likelihood of the feature being present in the item relative to the base rate, ρ_a. If the feature is absent, we add to the match a number reflecting the estimated likelihood of the feature being absent in the item relative to the base rate of feature absence, $1 - \rho_a$. The particular function relating the estimated likelihoods of feature presence or absence to the amount added is not linear, but it is monotonic, and adherence to the exact form of this function may be relatively unimportant. Thus, it may be that a relatively simple summation of feature-by-feature quantities can produce variables that have the same computational characteristics as those that might be computed by a more self-consciously Bayesian likelihood calculation.

In sum, what may seem at first to be a complex statistical computation may reduce to a few relatively simple operations. Operations of a comparable complexity are used in all of the models we are aware of. The particulars of the computations assumed here can be justified in Bayesian terms, but this does not require the mechanism that computes them to be especially complex.

Use of separate representations for items in memory. Our model assumes that items are represented by detectors, each of which is separate from all of the others in that it keeps track of statistical quantities relevant to a particular item. These quantities are updated separately as the item is learned and used separately at the time of test. This approach differs from that of many models introduced in the last two decades, including models we have participated in developing (McClelland & Rumelhart, 1985; Chappell & Humphreys, 1994). In these models, the

stored parameters in the memory system (the *connection weights*) participate in the representation of many items, and there is no separate unit for each item, but instead a population of units, each of which participates in the representation of many items. Some readers may wonder, then, whether the present model represents an insufficiency of the distributed approach or a rejection by us of that approach to memory representation.

In our view, memory really is distributed at an underlying level. However, we take the relative degree of success of the model as suggesting that for present purposes at least, a localist approach can serve as an extremely useful and valuable abstraction. The localist approach is applicable in part because the literature we have considered deals with conceptually distinct items, namely words. Some of the properties of distributed models (e.g., their ability to learn a systematic mapping such as the mapping from spelling to sound and then to be able to apply this to novel items; Plaut, McClelland, Seidenberg, & Patterson, 1996) simply do not arise in the case of single-item recognition. What the localist approach allows us is the ability to conceptualize memory in terms of representation and updating of meaningful item-specific quantities, such as estimates of probabilities of particular features occurring in particular items (the ρ_i^μ) and the likelihood that a particular input matches a particular item previously studied ($\lambda^\mu[S]$). In a distributed model, these quantities will not be kept separate, and yet it may still be the case that in some sense the distributed network is computing what abstractly amounts to these particular quantities.

There is obviously a serious research question concerning exactly what sort of distributed memory model could successfully store the necessary quantities and perform the necessary computations. A full consideration of this issue is beyond the scope of this discussion, but it is worth noting that processes related to those we see here can be exhibited by some distributed memory models. We will just note that we expect a successful distributed model to employ relatively sparse distributed representations, so that each item is represented by only activating a relatively small subset of the pool of available elements. Below, we describe one such model (Chappell & Humphreys, 1994) that is already known to capture many aspects of the data addressed by our present model. We also consider this matter further in the section on possible connectionist implementations.

Optimality of behavior. The introduction of Bayesian approaches to the representation and calculation of probabilistic quantities often accompanies the view that it is useful to construe human behavior and cognition as rational or optimal in some way, or at least to consider the possibility that human behavior may be rational or optimal. The optimality or lack thereof of human behavior is quite a broad issue of some social relevance, and we want to avoid creating the impression that our research sheds much light on this as a general issue. However, we can observe that in our experience, optimality can only be defined—let alone achieved—with respect to a set of assumptions about the true state of affairs that holds in the situation where the behavior or cognition occurs and about the knowledge that an individual has available for use in the situation. The gradual procedure our model uses to adjust estimates of feature probabilities may be optimal if features are probabilistic, but if the features are in fact deterministic, the gradual adjustment procedure would be decidedly nonoptimal. Also, if an individual is assumed to have access to sufficient statistics characterizing the distributions of the decision variable for each of the two classes of items (*old* or *new*) in a memory experiment, then one can consider whether this information is used optimally to set the criterion for assigning inputs to classes. If individuals do not have access to that information, they could still be behaving optimally in the sense of doing the best they possibly can on the basis of the information that is available to them.

In general, we have considerable sympathy with Simon's (1972) notion of bounded rationality, which explicitly considered the fact that true optimality generally requires omniscience and unbounded computation.

Familiarity and differentiation. Finally, we turn our attention back to the point with which we began, namely the basic idea that familiarity arises from a process of differentiation. This idea lies at the heart of the strength-mirror and null list-strength effects, and many other aspects of the model's behavior. The effect that familiarization has on cumulative confidence curves, shifting the curve for *no* responses down the confidence categories and shifting the curve for *yes* responses up the confidence categories, likewise reflects the differentiation process at work.

In our view, differentiation is also a fundamental process in perceptual learning that arises in a very wide range of psychological research contexts. There are many findings in animal learning that appear to call for some form of differentiation as a result of familiarization, including, for example, the positive effect of preexposure to similar stimuli on rate of acquisition of discriminative responses in animals (see Saksida, 1998, for recent discussion and a model of this phenomenon). There are also several findings from human perceptual and information processing that appear to suggest the idea that familiarity produces differentiation. For example, there are several cases in which participants are actually faster or more likely to reject close matches to high-frequency stimuli than to low-frequency stimuli (Cole, 1973; Van Orden, 1987). If, as in Morton's (1969) logogen model, we restrict the effect of familiarization entirely to the bias term, these findings would be difficult to account for. However, if we assume that familiarity leads to increased certainty concerning the features of an item (e.g., the letters or phonemes of a word), then we should be better at rejecting close mismatches to high-frequency items than to low-frequency items, exactly as Cole and Van Orden found.

Comparison With Previous Models

In this section, we consider three other recent models that address many of the phenomena we have addressed, and we compare and contrast these models with our own.

The attention-likelihood model. Glanzer et al. (1993) presented a model of recognition memory built around the idea that participants use likelihoods in discriminating old from new stimuli. This model has pointed the way for other models such as ours and the REM model of Shiffrin and Steyvers (1997). However, there are substantial differences between this initial likelihood-based model and the newer models. Glanzer et al. assumed that all items are represented by vectors of the same number, N, of features, and that the recognition decision is based

on the number of features out of a sample, n, from the total of N, found to be "marked" at the time of test. Prior to study, items in all conditions are assumed to have the same proportion, $p(new)$, of their features marked. During study and test, $n(i) = \alpha(i)N$ of these features are sampled. The condition is denoted by i; more features are sampled in the stronger condition (e.g., low-frequency words) at study and at test. This is said to be because more attention is paid by participants to lower-frequency words. During study, features sampled that are not already marked become marked, so that the proportion of marked features in studied items is

$$p(i, old) = p(new) + \alpha(i)[1 - p(new)].$$

At test, the decision is not based on the number, x, out of the sample, $n(i)$, found to be marked per se, but on the likelihood ratio of this number—that is, the probability of obtaining this number under the assumption that an old item is being tested, divided by the probability of obtaining this number under the assumption that a new item is being tested. Now these probabilities are given by

$$p(x|i, j) = \binom{n(i)}{x} p(i, j)^x q(i, j)^{n(i)-x},$$

where j indicates *old* or *new* and $q(i, j) = 1 - p(i, j)$. Glanzer et al. (1993) showed that the true log-likelihood ratio is given by

$$\ln L(x|i) = n(i) \ln \left[\frac{q(i, old)}{q(new)} \right] + x \ln \left[\frac{p(i, old)q(new)}{p(new)q(i, old)} \right]. \quad (13)$$

The attention-likelihood model differs in three important ways from our model. Most importantly, the likelihoods involved are quite different in nature. In our model, a test item is assessed against all of the detectors. Each detector reports its estimate of the log likelihood that the input matches it. This log-likelihood is based only on the degree of familiarization with the item the detector represents. In contrast, in the attention-likelihood model, a test item accesses only its own representation, where the number, x, of marked features out of the $n(i)$ sampled features are determined. A single likelihood ratio is then computed, which in this case depends on (estimates of) the number of marked features in new items and in old items from the same condition. The effect of this is to make the likelihood ratio values that are used as the basis of decisions dependent not just on the number of marked features of an item, but on the relationship between this number and the number of marked features of other items in the same condition. This characteristic of the attention-likelihood model appears to lead it to predict that cumulative confidence rating curves for different strength and list-strength conditions should cross in a way that is not consistent with the pattern seen in Balakrishnan and Ratcliff (1996).

A second difference between the attention-likelihood model and our model lies in the fact that our model explicitly relies on estimates of the relevant probabilistic quantities, whereas Glanzer et al.'s (1993) relies on true values of these quantities. However, as they admitted and Hintzman (1994) was at pains to point out, it is not clear that the recognition system will have access to the probabilities in Equation 13. Hintzman therefore proposed replacing p and q with expressions that explicitly indicate that the relevant variables must be estimated. Glanzer et al. showed that for some purposes their theory is quite robust to poor estimations of these probabilities, and hence to the possibility that participants may not use true likelihoods. First, d's calculated within a condition are not affected by the transformation represented by Equation 13 (this is true of any linear transformation) and thus by errors in estimation of the quantities used in the equation. However, the distances between the means for different conditions will be affected by the transformation. Glanzer et al. stated, however, that the basic mirror pattern is preserved despite quite wide ranging estimates of the $p(new)$ and the $p(i, old)$ values. They illustrated this point with a case where the recognition system only uses one estimate for $p(i, old)$, the average of those for the two conditions. It thus appears that some modification of the Glanzer et al. model using estimated rather than true likelihoods may be a viable candidate to account at least for some aspects of the data. However, for such a model to be viable, specific procedures for deriving estimates of the relevant quantities would have to be proposed. It would only then be possible to assess the model's ability to address some of the subtler aspects of the experimental data, such as changes in $\ln[\beta]$ as a function of conditions. Whether an estimated likelihood version of the model would avoid the incorrect prediction that certain cumulative confidence curves would cross cannot be determined until a specific basis for the estimation process is proposed.

A final difference between our model and the attention-likelihood model is that in our model, the actual featural content of items plays a role, whereas in the attention-likelihood model, all that matters is whether features of an item are marked. This difference allows our model access to a host of factors that are likely to play roles in recognition memory, as discussed above in the paragraph on the use of patterns to represent items. In particular, in our model we can address effects of item similarity, and we can address effects of word frequency in terms of differences in featural content or featural variability, instead of relying solely on differences in attention as the basis for such differences. Although different items may well elicit different degrees of attention, we believe that other item-specific factors also play a role in recognition memory and that these can be more adequately captured in models that make reference to item content.

The REM model. As we have already noted, there are many similarities between our model and the REM model of Shiffrin and Steyvers (1997). REM was developed independently and at about the same time as our model, and in our view, the successes of both models derive from two facts: (a) both use inherently noisy, vector-based representations of items in memory; and (b) both derive procedures for the representation and use of inherently probabilistic information within a Bayesian, likelihood framework. In REM, each study item consists of a vector of elements, and each element may take any one of an infinite number of values that are assigned numeric indices increasing from 1. The probability that an element will have the ith value

is $g(1 - g)^{i-1}$, where g is a probability. During learning, the model samples a subset of the elements of each item vector and forms a detector for each, in which the values of the sampled elements are recorded. Values are recorded correctly with probability c (with probability $1 - c$, a random value is recorded for a sampled element). During test, the number of features in the test item that match the features of each detector are counted, and the likelihood that such a number of features would match is computed, under the hypothesis that the test item matches the detector and under the hypothesis that they do not match. The likelihood ratio thus obtained is then used to compute the overall odds ratio that the sample was based on any of the stored items, and if the odds are greater than 1 that it was, the model responds that the item is *old*, on the basis of knowledge of the feature probability distribution and the parameters g and c.

In light of the fact that the models were developed independently, it is striking how many similarities exist between the two models. Both models represent items as vectors of elements, though each element takes on only one of two possible values (0 and 1) in our case instead of an infinite range of possibilities in REM. In both models, the sampled values of features are subject to random perturbation. During test, both models compute estimates of the odds that the test item came from the generator for each studied item, as represented by the corresponding detector. The decision procedure is slightly different in that our model depends on whether the likelihood computed by any single detector exceeds a threshold, whereas the REM computation involves a sum of quantities computed by the detectors, but we doubt that this sum plays a significant role in accounting for the results of most experiments. As a result of these similarities, it is not surprising that the models account for many of the same empirical phenomena.

One positive feature that is shared by our model and by REM is the fact that both provide ways of capturing the possibility that material-based mirror effects may be due at least in part to differences in the featural properties of stimuli. For example, Shiffrin and Steyvers (1997) suggested that the value of g may be lower for low-frequency words. This is tantamount to the assumption that these words tend to have rarer feature values. Rarer feature values are more diagnostic than less rare values; it is relatively unlikely for a sample and a detector to match on a rare value by chance. This same idea could be captured in our model by assuming that the values of some of the parameters of the model (e.g., f) might vary from feature to feature, with high-frequency words having more features with relatively large values of f.

Another factor that can vary between words is the consistency of the representations they evoke from presentation to presentation. In our simulations, we considered the possibility that features of frequent words may be more context sensitive and therefore less consistent than the features of infrequent words. We simulated this possibility by using a larger value of the p_0 for frequent words so that their representations would be more variable. With this manipulation we were able to show that differences would lead to corresponding mirror effects. Though Shiffrin and Steyvers (1997) did not consider the possibility that materials might vary in consistency, their model could potentially capture this factor by assuming that their parameter c varies between different types of materials. In our view, the two models have merely considered two alternative ways of accounting for one particular type of item effect in recognition memory. It is likely that either approach could have been taken within either model. Further research is necessary to see whether these approaches can be distinguished empirically.

In spite of the many similarities between the models, there are two important differences. One important difference is that our model learns estimates of the probability that features take on particular values, whereas Shiffrin and Steyvers's (1997) model assumes that these probabilities are general and known in advance (through the parameters g and c). Shiffrin and Steyvers' model is like the Glanzer et al. (1993) model in this respect. Indeed, Shiffrin and Steyvers described their model as one that calculates actual conditional probabilities rather than estimates of these probabilities. As they point out, however, this calculation of actual conditional probabilities breaks down in simulations of mixed lists of items (such as mixed lists of high- and low-frequency words). In general, as we have already stressed, it seems likely that it will be necessary to assume that participants are working not with actual probabilities but with estimates. In any case, it is necessary to specify where the probabilities or estimates that are used come from. As already discussed, our model assumes that all such quantities, with the exception of ρ_a, are initialized at random and subject to adaptation as a result of experience; even ρ_a might ultimately be the product of preexperimental learning.

The second important difference between the models is related to the first. What our model learns about items is very different from what is learned in the REM model. In REM, the value assigned to the model's representation on a particular feature dimension is determined the first time that dimension is sampled. On any given presentation, only some dimensions are sampled, and the sampling process is imperfect, thereby creating uncorrectable inaccuracies in the representation. Once it is sampled, its value becomes fixed and cannot be further refined; later presentations can only fill in features that were not sampled initially. In contrast, our model adjusts its estimate of the probability that each feature has a particular value (because features only take values of 0 or 1, the model needs only to estimate the probability that the value is 1, because 0 is the only other possibility). The same feature may be sampled repeatedly, and as sampling continues the model's estimate of the probability that the feature takes on a particular value becomes more refined. This difference has several important consequences. One is that it allows our model to learn the actual properties of the underlying generators in a more complete way. If in fact the underlying generators of stimuli are really probabilistic, the probabilities of the various element values provide a more complete characterization of the generators than the values that happen to have been recorded from a particular sample produced from these probabilities. In general, we believe that the structure of experience is probabilistic and that models that actually learn estimates of probabilities will therefore prove ultimately to be more adequate. The second consequence relates to the effects of multiple exposures to the same item during learning. In both models, the random perturbation of features during learning means that the samples on different trials may be different. In our model, repeated sampling leads to gradual convergence on the actual

underlying probabilities, whereas in Shiffrin's model, no such convergence is possible.

These two different approaches lead to two different ways of understanding the findings of the studies of Hintzman et al. (1992) and Hintzman and Curran (1995), in which participants did not generally learn to completely differentiate items from other very similar items. In REM, the failure to achieve complete differentiation is seen as a sign that initial coding decisions are not altered by later experience. In our model, the failure reflects instead the natural variability in the results of the encoding operation from trial to trial. The variability can be reduced by forcing accurate encoding of differentiating features.

In summary, our model shares many features with the REM model. REM tends to be cast in terms of calculations of actual likelihoods, but in practice it appears that some estimation is always needed in REM as well, so in the end this difference may come down to a matter of viewpoint. The other difference relates primarily to whether one thinks that it is features or estimates of the probabilities of features that are learned. It will be interesting to see if experiments can be devised that will distinguish these two alternative approaches to item representation in memory.

Chappell and Humphreys (1994). Chappell and Humphreys proposed a neural network model of recognition memory. An auto-associator was assumed to store long-term semantic memories of the items that might occur in the experiment, forming a kind of vocabulary. At study, associations were learned between context and the units making up these items. Weights within the auto-associator were also increased slightly, thereby slightly expanding the size of the attractor basins for studied words. At test, the decision was based upon whether or not convergence to the appropriate memory minimum occurred. This model gave good fits to all of the data modeled here, with the exception of mirror effects, which were not considered.

Chappell and Humphreys's (1994) model in fact did have differentiation-like processes. Thus, as any one item's weights in the auto-associator were strengthened, the critical overlaps for all other items increased slightly, which would make it slightly more difficult to converge on them. However, with the parameters used, the effect was infinitesimally small (Chappell, 1993). A slightly larger, but still small, effect was due to the global inhibition in the auto-associator being increased as items were studied.

A major difference between that model and the present one is that the Chappell and Humphreys (1994) model assumed that episodic memories are formed through associations between context and existing memory representations, and the detectors are essentially the existing memory minima in the auto-associator. The present model forms entirely new detectors during study and treats recognition memory as if the only items with which the test stimuli are compared are those that arose from the study of the list currently under test. Below, we consider how sensitivity to context may be incorporated into the present model.

Further research is planned to better understand the relationships between the two models and to see if they can be empirically differentiated (see the section below on possible connectionist implementations of our model).

Issues Facing the Model

In this section, we consider a number of issues facing the model that might be addressed in subsequent research.

How do repeated presentations access the same detector? In our simulations, we have assumed that once a detector is formed, it is accessed by all subsequent repetitions of stimuli based on the same underlying item generator. Because stimuli are in fact only probabilistically based on these generators, it is not trivial to determine which detector if any should be reactivated when a new sample from the same generator is presented. Here we ask, "How might reaccessing of the same detector occur, and how reliable might it actually be?" Regarding the first point, we suggest that the same mechanism that is used for recognition can also be used for reselection of the same detector when a repetition occurs during study. Obviously, this process would not be fool-proof, but it should be noted that probability of success in reaccessing the same detector might be considerably higher during study than it would be at test. This would be the case if encoding varied less within study than between study and test, or if there is any decay or regression of the effects of each experience on the parameters of detectors (the ρ_i^μ) as a function of time between presentations.

Factors affecting the effective learning rate. The previous paragraph suggests that the effective magnitude of changes in the ρ_i^μ may vary as a function of the length of the study–test interval. For simple cases, the effects of this can be modeled by assuming that the effective learning rate varies with the duration of the study–test interval. What other variables would affect the effective learning rate? It seems obvious and natural to suppose that the amount of time available for study of an item would also affect the initial size of the adjustments made to the weights, as we did in modelling Ratcliff et al.'s (1994) data.

In these simulations, we have explicitly assumed that the learning rate decreases with the number of presentations of the same item. Although this is justified from the point of view of leading to convergence on optimal estimates of the weights, it is somewhat less easy to see how this might be implemented. Somehow the system would have to keep track of the number of times an item had been presented or some reasonable proxy of this number. One possibility is that relatively novel stimuli lead to a novelty response that in turn amplifies the rate of learning; as stimuli become more familiar this novelty response would lessen, and with it the effective learning rate. Perhaps this novelty response is mediated by the same mechanisms that mediate recognition; that is, perhaps the amount of learning is based on the estimated likelihood that the experience was in fact produced by the generator corresponding to one of the stored detectors.

Possible connectionist implementation? Throughout this article we have alluded occasionally to connectionist ideas. We draw our learning rule from the competitive learning rule as used by Rumelhart and Zipser (1986), and we have associated the estimated conditional probabilities stored in the model with the weights in a connectionist network. Could the model in fact be implemented in a connectionist network and if so how would the implementation actually proceed? One possibility would be to use a localist connectionist model like the one actually used by Rumelhart and Zipser. Such networks have the same structure

as is illustrated in Figure 11. One set of units corresponds to our feature units (and these take activations of 1 and 0), and the other set of units corresponds to our detectors. The units are initialized with small random weights. When a stimulus comes in, the unit that most closely matches it becomes activated and adjusts its weights. It can be seen that these processes closely parallel those in our model, suggesting that it will be worth examining the relationship in more detail in subsequent research.

A second possibility is to use a more distributed connectionist model, such as the one proposed by Chappell and Humphreys (1994), as mentioned earlier in this discussion. It seems likely that the hippocampal region of the brain, which of course is crucial for recognition memory (Squire, 1992), uses a sparse, distributed model somewhat like that of Chappell and Humphreys (1994; for further discussion, see McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & McClelland, 1994). An examination of the relationship among Bayesian models like the one presented in this article, localist connectionist models such as Rumelhart and Zipser's (1986), and distributed connectionist models may ultimately lead to an understanding of the relationship between the neural mechanisms underlying human memory and their psychological consequences for performance in memory tasks.

Extensions

As suggested in the introduction and previously in this discussion, the concept of differentiation is as relevant in perception and perceptual learning as it is in the domain of memory. Although there has been some relevant modeling work related to perceptual learning in animals (Saksida, 1998; Myers, Gluck, & Grainger, 1994), this appears to us to be an area that is ripe for further exploration. Within the recognition memory framework, there are also several possible extensions. Here, we mention two extensions that seem most direct and natural.

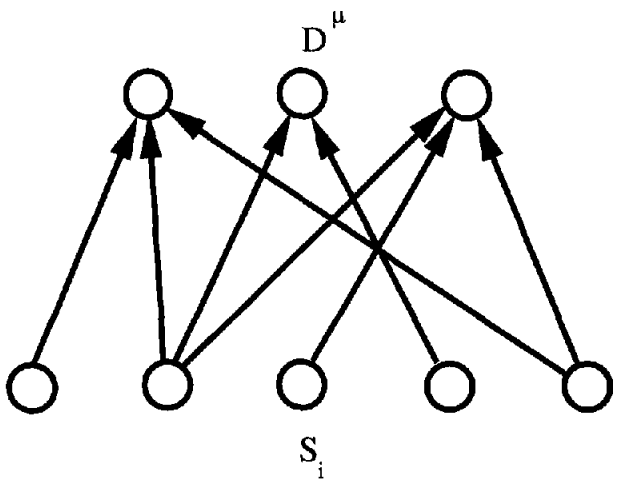

Figure 11. Schematic diagram of possible connectionist implementation of the recognition model. Input units (bottom row) allow representation of the values of the input features S_i; units in the upper row correspond to item detectors D^μ; and arrows to a detector unit from each input unit represent the connection weights that encode knowledge of feature probabilities conditional on the hypothesis that the input is an example of the item corresponding to the detector.

Context and recognition. As noted in our comparison of the present model to that of Chappell and Humphreys (1994), we have up to now neglected context in memory retrieval and the possible role of memories formed outside the experimental study context. This obviously reflects a limitation of the present model, because recognition is clearly susceptible to interference from items on different lists as well as from extraexperimental items. Our model can be extended to include influences of extralist memories while retaining the strong sensitivity to context that memory clearly exhibits, if the memory vectors include features of both the item and the context. Indeed, we might think of each experience as consisting of the concatenation of two vectors representing stimulus and context features. During study of a single list, the context stays relatively constant. If at test this context is treated as part of all of the memory probes, and if the context is sufficiently distinct from other contexts, memory vectors for experiences that occurred in other contexts will have very low estimated likelihood ratios and will have negligible probabilities of reaching criterion. Thus, our present model can be seen as a special case of a more general model in which context explicitly plays a role.

These modifications would increase the convergence between the present approach and the approach to modeling memory represented by the SAM model (Gillund & Shiffrin, 1984) and the REM model (Shiffrin & Steyvers, 1997). With respect to the latter, Shiffrin and Steyvers suggested that memory vectors encompass both context and stimulus features, but they neglected the context components in their analysis of length, strength, and mirror effects, on the basis of an argument very similar to the one that we have just given. These suggestions link the current proposals back to key features of the SAM model. In the introduction, we saw that familiarity in SAM is a product of stimulus and context terms. We also saw how our item detectors capture the differentiation assumed in the modified version of SAM, and we noted that the model's estimated likelihoods correspond to SAM's stimulus–image associations. These observations can be extended to the context–image associations: The contribution to the total estimated likelihood that would be provided by the context portion of the feature vector in our extended model (and in the full version of REM) corresponds to the contribution of the context–image associations to the total activation of a particular memory image in SAM. Because the contributions of concatenated vectors are multiplied in the computation of likelihoods, the concatenation of context and item vectors is essentially the same mathematically as the multiplication of context–image and stimulus–image associations in determining the activation of a memory image in the SAM model.

Latency. It may not be immediately apparent how our model might be extended to model latency data. However there is a model that "Ward (1947) developed and that Stone (1960) suggested as a choice reaction-time model" (Luce, 1986, p. 340).[4] This is the Sequential Probability Ratio Test (SPRT)

[4] We would like to thank Marius Usher for bringing this model to our attention.

random walk model. If a sequence of observations is made of a random variable S, call them S_m, then if the log-likelihood ratio is evaluated for each observation, the sum of the log-likelihood ratios is the overall log-likelihood ratio for the sequence of observations. Upper and lower boundaries may be set for this sum so that appropriate decisions are made when one of these boundaries is crossed. The sum describes a random walk.

> Wald and Wolfowitz (1948) showed that this sampling procedure is optimal in the sense that if the error probabilities are fixed, then the number of observations before a decision is reached is, on average, no greater than it would be with any other rule. (Luce, 1986, p. 340)

Swensson and Green (1977) showed that basing decisions on log-likelihoods in this way predicts that the latency distributions for, say, *yes* responses are the same for stimuli from both possible classes. This is rarely the case in data, so this model can usually be easily rejected.

However, the quantities calculated in our model are not true log-likelihoods for the S, only estimated ones. Preliminary simulations reveal that if we accumulate estimated log-likelihoods, with items generating a new S for each observation, different latency distributions are obtained for hits and false alarms. Such a model would implement an analog of Ratcliff's (1978) model, with the average estimated log-likelihood for a test item in our model being related to the average "drift" for an item in that model. And Ratcliff's decision rule is equivalent to basing the decision on whether or not the maximum of the cumulatives of the estimated log-likelihoods for the detectors has crossed the upper or lower detector boundaries. A natural extension of the present work would involve an examination of whether all the effects we have considered here would be preserved in a simulation of detector activation as a thresholded-time-dependent process.

Conclusion

Our discussion suggests there are similarities between the mechanisms in our model and those in several other models, and that the ideas we and others have explored will lead to useful extensions and elaborations. These mechanisms have allowed us to model the nonlinear length–strength dissociation and false recognition data for similar distractors, which challenged the global matching models. Our model can also produce appropriate z-ROC curves in a range of paradigms.

We believe that our model represents part of an ongoing assimilation of two key ideas into models of memory and information processing. One is the idea that memories are patterns rather than atomic items. The other is the idea that learning is a process related to the estimation of statistical quantities, such as conditional probabilities. These ideas seem likely to continue to play key roles in our efforts to understand many aspects of cognition and memory.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96,* 703–719.

Ashby, F. G. (1982). Deriving exact predictions from the cascade model. *Psychological Review, 89,* 599–607.

Ashby, F. G., & Lee, W. W. (1993). Perceptual variability as a fundamental axiom of perceptual science. In S. C. Maslin (Ed.), *Foundations of perceptual theory.* New York: Elsevier Science.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, 12,* 171–178.

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 615–633.

Balota, D., & Neely, J. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 576–587.

Carpenter, G., & Grossberg, S. (1993). Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences, 16,* 131–137.

Casella, G., & Berger, R. L. (1990). *Statistical inference.* Belmont, CA: Duxbury.

Chappell, M. (1993). *Neural network models of human recognition and cued recall.* Unpublished doctoral dissertation, University of Queensland, Queensland, Australia.

Chappell, M. (1998). Predictions of a Bayesian recognition memory model (and a class of models including it). In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). London: Oxford University Press.

Chappell, M., & Humphreys, M. S. (1994). An auto-associative neural network for sparse representations: Analysis and application to models of recognition and cued recall. *Psychological Review, 101,* 103–128.

Chappell, M., & McClelland, J. L. (1994, August). *Bayesian models of recognition memory.* Paper presented at the 27th Annual Mathematical Psychology Meeting, Seattle, WA.

Cole, R. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics, 13,* 153–156.

Gegenfurtner, K. R. (1992). PRAXIS: Brent's algorithm for function minimization. *Behavior Research Methods, Instruments, & Computers, 24,* 560–564.

Gibson, E. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47,* 196–229.

Gibson, E. (1969). *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts.

Gibson, E., & Walk, R. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology, 49,* 239–242.

Gibson, J., & Gibson, E. (1955). Perceptual learning—differentiation or enrichment? *Psychological Review, 62,* 32–41.

Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

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, 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.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (in press). Slope of the receiver operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355–1369.

Gronlund, S. D., & Ohrt, D. D. (1994, November). *The list-length effect: Controls and theoretical implications (Abstract 481).* Poster presented at the 35th Annual Meeting of the Psychonomic Society, St. Louis, MO.

Grossberg, S. (1976). On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biological Cybernetics, 21,* 145–159.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. *Progress in Theoretical Biology, 5,* 233–374.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 77–109). Cambridge, MA: MIT Press.

Hinton, G. E., & Sejnowski, T. J. (1983, June). *Optimal perceptual inference.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC.

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

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

Hintzman, D. L. (1997, November). *Bidirectional integration of familiarity as an explanation of the mirror effect (Abstract 391).* Paper presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, PA.

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

Hintzman, D. L., & Curran, T. (1995). When encoding fails: Instructions, feedback, and registration without learning. *Memory & Cognition, 23,* 213–226.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667–680.

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

Hoeffner, J. H. (1996). *Are rules a thing of the past? A single mechanism account of English past tense acquisition and processing.* Unpublished doctoral dissertation, Carnegie Mellon University.

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

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96,* 208–233.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association, 90,* 773–795.

Knapp, A., & Anderson, J. A. (1984). A signal averaging model for concept formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 617–637.

Levy, W. B., Colbert, C. M., & Desmond, N. L. (1990). Elemental adaptive processes of neurons and synapses: A statistical/computational perspective. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory* (pp. 187–235). Hillsdale, NJ: Erlbaum.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Mandler, G., Goodman, G., & Wilkes-Gibbs, D. (1982). The word-frequency paradox in recognition. *Memory & Cognition, 10,* 33–42.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

McClelland, J. L. (1991). Stochastic interactive activation and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McClelland, J. L. (1993). Toward a theory of information processing in graded, random, and interactive networks. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience* (pp. 655–688). Cambridge, MA: MIT Press.

McClelland, J. L., & Chappell, M. (1994, November). *Bayesian recognition (Abstract 171).* Poster presented at the 35th Annual Meeting of the Psychonomic Society, St. Louis, MO.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

Morris, P. E. (1978). Frequency and imagery in word recognition: Further evidence for an attribute model. *British Journal of Psychology, 69,* 69–75.

Morton, J. (1969). Interactions of information in word recognition. *Psychological Review, 76,* 165–178.

Movellan, J. R., & McClelland, J. L. (1995). *Stochastic interactive activation, Morton's Law, and optimal pattern recognition* (Technical Report PDP.CNS.95.4). Pittsburgh, PA: Department of Psychology, Carnegie Mellon University.

Murdock, B. B., & Kahana, M. J. (1993). An analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689–697.

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

Murnane, K., & Shiffrin, R. (1991b). Word repetitions in sentence recognition. *Memory & Cognition, 19,* 119–130.

Myers, C. E., Gluck, M. A., & Grainger, R. (1994). Dissociation of hippocampal and entorhinal function in associative learning: A computational approach. *Psychobiology, 23,* 116–138.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 4,* 661–682.

Oswalt, R. (1972). Relationship between level of visual pattern difficulty during rearing and subsequent discrimination in rats. *Journal of Comparative and Physiological Psychology, 81,* 122–125.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190–214.

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

Rau, K., & Proctor, R. (1984). Study-phase processing and the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 386–394.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1. Foundations* (pp. 151–193). Cambridge, MA: MIT Press.

Saksida, L. M. (1998). *Reconciling associative and non-associative accounts of preexposure effects in discrimination learning: A competitive–configural connectionist model.* Unpublished manuscript, Center for the Neural Basis of Cognition, Pittsburgh, PA.

Shiffrin, R. (1995). *A model of recognition memory: REM: Retrieving efficiently from memory* (Technical Report 136). Bloomington: Cognitive Science Program, Indiana University.

Shiffrin, R., Ratcliff, R., & Clark, S. (1990). The 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 of recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Simon, H. A. (1972). Theories of bounded rationality. In C. B. Radner & R. Radner (Eds.), *Decision and organization* (pp. 161–176). Amsterdam: North-Holland.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Swensson, R. G., & Green, D. M. (1977). On the relations between random walk models for two-choice response times. *Journal of Mathematical Psychology, 15,* 282–291.

Van Orden, G. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition, 15,* 181–198.

von der Malsburg, C. (1973). Self-organizing of orientation sensitive cells in the striate cortex. *Kybernetik, 14,* 85–100.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation, 1,* 425–464.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354.

Appendix

Derivation of Model Moments

We wish to characterize the distribution of $\lambda^{\mu}(S)$, the log-likelihood that input S was generated by item μ. The right-hand side of Equation 4 is the sum of N terms. We are interested in its expected value, in which case we may consider that we are adding N random variables from identical distributions. We thus wish to find the expected value of the log-likelihood at the ith component (the expected value for the detector will then be N times this value):

$$\lambda_i^{\mu} = S_i \ln\left(\frac{\rho_i^{\mu}}{\rho_a}\right) + (1 - S_i) \ln\left(\frac{1 - \rho_i^{\mu}}{1 - \rho_a}\right).$$

Let c be 0 for low-firing units (with probability $1 - f$) and 1 for high-firing units. Let t indicate if a 0 or a 1 occurs at test. Then we have for new items (dropping the μ and i super- and subscripts)

$$\overline{\lambda^n} = \sum_{t,l_1,\ldots,l_n,c=0}^{1} [\rho_a^t(1 - \rho_a)^{1-t}] \prod_{m=1}^{\tau} [p_c^{l_m}(1 - p_c)^{1-l_m}][f^c(1 - f)^{1-c}].$$

$$\left\{\kappa\left[\overline{t \ln\left(\frac{\rho(I)}{\rho_a}\right) + (1 - t) \ln\left(\frac{1 - \rho(I)}{1 - \rho_a}\right)}\right] + (1 - \kappa)\left[\overline{t \ln\left(\frac{\rho(0)}{\rho_a}\right) + (1 - t) \ln\left(\frac{1 - \rho(0)}{1 - \rho_a}\right)}\right]\right\}.$$

Thus, once a value of c has been chosen, a 1 occurs on each of the presentations with probability p_c, and the test item presents a 1 with probability ρ_a. The expectation of the logarithm terms are with respect to the logistic-normal distribution for v (or whatever other distribution might be chosen). The weight, $\rho(I)$, is as given by Equation 10, and $\rho(0)$ is the weight with no learning ($= v$).

What does the logistic-normal distribution for v look like? Suppose o is normally distributed:

$$f(o) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(o-\mu)^2/2\sigma^2}.$$

Taking the logistic,

$$v = \frac{1}{1 + e^{-o}},$$

and solving for o, we find:

$$o = \ln \frac{v}{1 - v},$$

so that the transformation has Jacobian (Casella & Berger, 1990):

$$|J| = \frac{1}{v(1 - v)}.$$

This leads to the following expression for the logistic-normal distribution:

$$f(v) = \frac{1}{\sqrt{2\pi}\sigma v(1 - v)} e^{-(\ln(v/1-v) - \mu)^2/2\sigma^2} \quad 0 < v < 1.$$

The formula for old items differs only in the first term:

$$\overline{\lambda^o} = \sum_{t,l_1,\ldots l_s,c=0}^{1} [p_c^t(1-p_c)^{1-t}] \prod_{m=1}^{s} [p_{c^m}^{l_m}(1-p_c)^{1-l_m}][f^c(1-f)^{1-c}].$$
$$\left\{\kappa\left[\overline{t \ln\left(\frac{\rho(\boldsymbol{l})}{\rho_a}\right) + (1-t)\ln\left(\frac{1-\rho(\boldsymbol{l})}{1-\rho_a}\right)}\right]\right.$$
$$\left. + (1-\kappa)\left[\overline{t \ln\left(\frac{\rho(0)}{\rho_a}\right) + (1-t)\ln\left(\frac{1-\rho(0)}{1-\rho_a}\right)}\right]\right\}.$$

For the Hintzman and Curran (1995) paradigm, we need expected values for similar items:

$$\overline{\lambda^{sim}} = \sum_{t,l_1,\ldots l_s,c=0}^{1} \left\{[p_1\eta + p_0(1-\eta)]^c\left[p_1\frac{(1-\eta)f}{1-f} + p_0 \times \left(1 - \frac{(1-\eta)f}{1-f}\right)\right]^{1-c}\right\}^t \cdot$$
$$\left\{1 - [p_1\eta + p_0(1-\eta)]^c\left[p_1\frac{(1-\eta)f}{1-f} + p_0 \times \left(1 - \frac{(1-\eta)f}{1-f}\right)\right]^{1-c}\right\}^{1-t} \cdot$$
$$\prod_{m=1}^{s} [p_{c^m}^{l_m}(1-p_c)^{1-l_m}][f^c(1-f)^{1-c}].$$
$$\times \left\{\kappa\left[\overline{t \ln\left(\frac{\rho(\boldsymbol{l})}{\rho_a}\right) + (1-t)\ln\left(\frac{1-\rho(\boldsymbol{l})}{1-\rho_a}\right)}\right]\right.$$
$$\left. + (1-\kappa)\left[\overline{t \ln\left(\frac{\rho(0)}{\rho_a}\right) + (1-t)\ln\left(\frac{1-\rho(0)}{1-\rho_a}\right)}\right]\right\}.$$

To calculate variances and skewness we need higher moments; the pth moment is given by

$$\mathrm{E}[(\lambda^{sim})^p] = \sum_{t,l_1,\ldots l_s,c=0}^{1} \left\{[p_1\eta + p_0(1-\eta)]^c \times \left[p_1\frac{(1-\eta)f}{1-f} + p_0\left(1 - \frac{(1-\eta)f}{1-f}\right)\right]^{1-c}\right\}^t \cdot$$
$$\left\{1 - [p_1\eta + p_0(1-\eta)]^c \times \left[p_1\frac{(1-\eta)f}{1-f} + p_0\left(1 - \frac{(1-\eta)f}{1-f}\right)\right]^{1-c}\right\}^{1-t} \cdot$$
$$\prod_{m=1}^{s} [p_{c^m}^{l_m}(1-p_c)^{1-l_m}][f^c(1-f)^{1-c}].$$
$$\left\{\kappa\left\{\overline{t\left[\ln\left(\frac{\rho(\boldsymbol{l})}{\rho_a}\right)\right]^p + (1-t)\left[\ln\left(\frac{1-\rho(\boldsymbol{l})}{1-\rho_a}\right)\right]^p}\right\}\right.$$
$$\left. + (1-\kappa)\left\{\overline{t\left[\ln\left(\frac{\rho(0)}{\rho_a}\right)\right]^p + (1-t)\left[\ln\left(\frac{1-\rho(0)}{1-\rho_a}\right)\right]^p}\right\}\right\}. \quad \text{(A1)}$$

Note that the formulae for new and old items are special cases of this, with $\eta = f$ and $\eta = 1$, respectively. Thus, this formula gives us all (noncentral) moments for new, old, and similar items. To determine variances we find the expected value with $p = 2$ and subtract out the squared mean. For skewness, we find the third moment and the third central moment is given by:

$$\mu_3 = \mathrm{E}[\lambda^3] - 3\mathrm{E}[\lambda]\mathrm{E}[\lambda^2] + 2\mathrm{E}^3[\lambda],$$

and the skewness by μ_3/σ^3 (for this work we do not subtract 3, so that a normal distribution has a skewness of 3).

Equation A1 may be somewhat simplified by noting that Equation 10 may be written

$$\rho(s, o) = \frac{(n-1)v + o}{n + s - 1},$$

where o is the number of 1s encountered by the weight during s learning trials. That is, the weight has no "memory" for the order of the 0s and 1s. With this, Equation A1 becomes

$$\mathrm{E}[(\lambda^{sim})^p] = \sum_{t,c=0}^{1} \left\{[p_1\eta + p_0(1-\eta)]^c \times \left[p_1\frac{(1-\eta)f}{1-f} + p_0\left(1 - \frac{(1-\eta)f}{1-f}\right)\right]^{1-c}\right\}^t \cdot$$
$$\left\{1 - [p_1\eta + p_0(1-\eta)]^c \times \left[p_1\frac{(1-\eta)f}{1-f} + p_0\left(1 - \frac{(1-\eta)f}{1-f}\right)\right]^{1-c}\right\}^{1-t} \cdot$$
$$\sum_{o=0}^{s} \binom{s}{o} [p_c^o(1-p_c)^{s-o}][f^c(1-f)^{1-c}].$$
$$\kappa\left\{\overline{t\left[\ln\left(\frac{\rho(s,o)}{\rho_a}\right)\right]^p + (1-t)\left[\ln\left(\frac{1-\rho(s,o)}{1-\rho_a}\right)\right]^p}\right\}$$
$$+ (1-\kappa)\left\{\overline{t\left[\ln\left(\frac{\rho(0)}{\rho_a}\right)\right]^p + (1-t)\left[\ln\left(\frac{1-\rho(0)}{1-\rho_a}\right)\right]^p}\right\}.$$

For computation, this may be further rearranged so that integrals, which must be performed to find the expected value of the log terms, are not repeatedly evaluated. When the expectations are with respect to the log-normal distribution, we do not believe they are analytially determinably, and we use numerical integration.

Now we may find the means, standard deviations, and skewness for a detector. For N units we multiply the means and standard deviations by N and $\sqrt{N}$ respectively. The skewness is found by dividing the skewness for each unit by $\sqrt{N}$ (because μ_3 increases with N, and the standard deviation with $\sqrt{N}$). These calculations may be checked by simulation to any desired accuracy.

To go further, we need to make an assumption about the distribution of λ for a detector, as this cannot be computed. To make Figure 3 we assume normal distributions. However, our simulations show that we do not get sufficiently accurate calculations assuming this distribution, as the true distributions have substantial skewness. We thus use instead Azzalini's (1985) *skew-normal* distribution. If $\phi(x - \mu/\sigma)$ is the density function for a normal distribution with mean μ and standard deviation σ, and $\Phi[(x - \mu/\sigma)]$ is the corresponding cumulative density function, then Azzalini's density function is

$$f(x|\mu, \sigma, \lambda) = 2\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right),$$

(Appendix continues)

where λ determines the amount of skewness and may be negative to give negative skewness. The cumulative density function and moment generating functions are (Azzalini, 1985)

$$F(x|\mu, \sigma, \lambda) = \Phi\left(\frac{x-\mu}{\sigma}\right) - 2T\left(\frac{x-\mu}{\sigma}, \lambda\right),$$

$$M(t) = 2e^{\mu t+\sigma^2 t/2}\Phi\left(\frac{\sigma\lambda t}{\sqrt{1+\lambda^2}}\right).$$

$T(h, a)$ is Owen's T for the computation of which Azzalini gives references for computer routines.

We find that the skewness is

$$\tau = \sqrt{2}(4-\pi)\left(\frac{\lambda}{\sqrt{\pi+(\pi-2)\lambda^2}}\right)^3.$$

Note that $-a < \tau < a$, where

$$a = \frac{\sqrt{2}(4-\pi)}{(\pi-2)^{3/2}} \approx 0.9952717463.$$

This may be inverted to give

$$\lambda = \frac{\sqrt{\pi}\sqrt[3]{\tau}}{\sqrt{\sqrt[3]{2}(4-\pi)^{2/3} - (\pi-2)\tau^{2/3}}}.$$

Thus, having computed the skewness of the distribution at the detector for a given item type, we may now determine λ.

The mean and standard deviation of the skew-normal distribution are given by (Azzalini, 1985)

$$m = \mu + \sqrt{\frac{2}{\pi}}\frac{\sigma\lambda}{\sqrt{1+\lambda^2}},$$

$$s = \sigma\sqrt{1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}}.$$

Once m and s have been computed for the detector and λ is known, these equations may be inverted to give μ and σ. $\ln[\rho(list)/(\Lambda - \rho(\overline{list}))]$ is then added to the mean (see Equation 6) to convert to log-odds. Then the cumulative density, F, may be used, with the criterion, ω_C, to compute the probability that a given type of target item gives a log-odds above criterion for a given individual detector.

As noted earlier, however, the distributions upon which decisions are effectively made are the maxima of the log-odds ratios across all the detectors. For new items we are effectively finding the maximum from L identical distributions. If they have cumulative density functions, $F(x)$, then the cumulative distribution of the maximum is (Casella & Berger, 1990)

$$F_{X_{\text{maximum}}}(x) = [F(x)]^L.$$

In general, the distribution of the maxima will have less variance than that of the individual distributions, which is the main reason our new item distribution has less variance than the old item distribution.

Finally, then, we can calculate false alarm and hit rates. For example, for the pure weak false alarm rate it will be

$$FAR_{PW} = 1 - F^L(\omega_C|\mu_N, \sigma_N, \lambda_N),$$

where μ_N, σ_N, and λ_N are parameters of the skew-normal distribution for a new item. The hit rate in this condition is

$$H_{PW} = 1 - F(\omega_C|\mu_o, \sigma_o, \lambda_o)F^{L-1}(\omega_C|\mu_N, \sigma_N, \lambda_N),$$

where μ_o, σ_o, and λ_o are parameters of the skew-normal distribution for an item presented to the detector that learned it. Writing these equations for all conditions of the list-strength and list-length experiments, one can derive predictions for the whole class of models which base their decision on the maximum of a set of detectors (Chappell, 1998).

To achieve the fits, we used the approximate analytic solution just derived. We used PRAXIS (Gegenfurtner, 1992) to find best fitting parameters for this approximation. Then a large stochastic simulation (typically 5,000 simulated participants per condition) was run with the parameters thus found. When the stochastic simulation disagreed with the analytic approximation, we estimated correction factors to apply to the estimates of performance generated by the analytic solution. These correction factors were then incorporated into the approximation, and PRAXIS was then run again with the new correction factors. This process continued until the actual stochastic simulation at the end of a PRAXIS run produced an adequate fit, or until the results stopped improving from one run to the next.

Received June 12, 1995
Revision received March 18, 1998
Accepted March 18, 1998 ■