Psychological Review

VOLUME 91 NUMBER 1 JANUARY 1984

A Retrieval Model for Both Recognition and Recall

Gary Gillund and Richard M. Shiffrin
Indiana University

The Search of Associative Memory (SAM) model for recall (Raaijmakers & Shiffrin, 1981b) is extended by assuming that a familiarity process is used for recognition. The recall model posits cue-dependent probabilistic sampling and recovery from an associative network. Our recognition model is closely related to the recall model because the total episodic activation due to the context and item cues is used in recall as a basis for sampling and in recognition to make a decision. The model, formalized in a computer simulation program, correctly predicts a number of findings in the literature as well as the results from a new experiment on the word-frequency effect.

"A critical problem of long standing in psychological study of memory is concerned with the relation between recall and recognition. In what sense are they the same, and in what sense are they different?" (Tulving & Watkins, 1973, p. 739).

This article is a preliminary attempt to formulate a theory that describes in detail the relationship between recall and recognition. The relationship is realized in a computer simulation model, and the model's predictions are checked against existing data as well as data generated in our laboratory. The model for recognition is related mathematically and logically to the Search of Associative Memory (SAM) theory of memory retrieval that has been used previously to fit recall data (Gillund & Shiffrin, 1981; Raaijmakers & Shiffrin, 1980, 1981a, 1981b).

We begin with a brief overview of theoretical treatments in the area of recognition (additional models are discussed in Section 5). Several studies carried out to test one class of models are then reported and evaluated. The introduction ends with a listing of some of the major findings in recall and recognition that the model is intended to explicate.

A rough classification of models for recall and recognition is given in Table 1. The component processes assumed to operate in recall and recognition are indicated by the symbols S, D_s, D_c, and I. S indicates an extended search, with the basic characteristic that it operates slowly and uncertainly. The speed and certainty of the search are defined empirically by results from tasks like free recall in which several seconds often elapse between successive recalls and in which many more than half the presented items may not be recalled even after many minutes of testing. Models with this property often assume serial sampling of images from memory, with or without resampling. D indicates direct access in which all the relevant information in memory is obtained in one step. D_s is a simplified form of direct access in which only the image of the tested item is contacted, so that searchlike ef-

This work was supported by Public Health Service Grant MH12717 and the Indiana University Waterman Research Award to Richard Shiffrin.

We would like to thank Lloyd Peterson, Margaret Intons-Peterson, Linda Smith, Maynard Thompson, Jim Neely, Michael Humphreys, Jeroen Raaijmakers, Roger Ratcliff, and Steven Smith for their thoughtful comments on earlier drafts.

Short annotated versions of the SAM simulation programs for recall and recognition, in PASCAL or FORTRAN, are available from the authors on request.

Requests for reprints should be sent to Richard M. Shiffrin, Psychology Department, Indiana University, Bloomington, Indiana 47405. Gary Gillund is now at the Center for Research in Human Learning, 204 Elliott Hall, University of Minnesota, Minneapolis, Minnesota 55455.

Table 1
Models for Recall and Recognition

Model	Recall	Recognition
1	$S + D_s$	D_s
2	$S + (I)$	$S + (I)$
3	$S + (I)$	$S + D_s$
4	$S + (I)$	D_c
5	$S + (I)$	$S + D_c$

Note. S = extended search; D_s = simple direct access process; D_c = complex direct access process; I = implicit recognitionlike process.

fects, such as list length and fan effects, cannot be predicted. D_c indicates a complex direct access process in which many memory images are contacted together, leaving open the possibility of predicting searchlike effects. I is a recognitionlike process that allows the subject to decide that a name recovered from a sampled image is appropriate to recall. I is placed in parentheses because it is often an implicit, rather than an explicit, component of a recall model.

Model 1 may be termed an *embedded two-process model* because the recognition mechanism is embedded in the recall mechanism. Examples of such models include Kintsch (1970, 1974) and J. R. Anderson and Bower (1972, 1974). In typical models of this type, recall is a function of two sequential stages. The first is a *generate* stage, whereby long-term store (LTS) is searched and possible word candidates are output to the *recognize* stage. In the recognize stage, a word's familiarity (Kintsch, 1970) or contextual associations (J. R. Anderson & Bower, 1972) are evaluated to determine if the generated word had been presented on the study list. During a recognition test, only the second stage is employed, and it is applied to the test word only. Direct access is presumably gained to the appropriate stored memory representation (possibly an episodic image), and the decision process then proceeds just as it does for each word generated in the recall process.

Embedded two-process models have difficulty explaining why recognition exhibits many of the same properties as recall, properties that seem to imply that recognition involves a memory search. Examples include work on organizational effects (Mandler, 1972; Mandler & Boeck, 1974; Mandler, Pearlstone, & Koopmans, 1969), test and lag effects (Murdock, 1974; Murdock & Anderson, 1975), repetition effects (Atkinson & Juola, 1973, 1974), and the list-length effect (e.g., Strong, 1912).

Perhaps the simplest alternative to Model 1 is a model proposing common mechanisms for recall and recognition. One such approach is seen in the work of Tulving (e.g., 1976). He proposed that recognition and recall both involve cue-dependent retrieval processes (Tulving, 1974), in which performance is dependent on the degree to which the cues available at the time of test (including the test itself in recognition) match the cues during presentation (Tulving, 1968, 1976; Tulving & Thomson, 1973).

A particular type of cue-dependency theory holds that recognition can be explained by the same search mechanisms typically used to model recall (e.g., the SAM model of Raaijmakers & Shiffrin, 1981b). This is indicated by Model 2 in Table 1. However, these models have a serious problem: Subjects can quickly, accurately, and with high confidence classify at least some distractors correctly (Atkinson & Juola, 1974; Fischler & Juola, 1971; Herrmann, Frisina, & Conti, 1978; Murdock, 1974; Murdock & Anderson, 1975; see also Glucksberg & McCloskey, 1981). Even a lengthy search may miss targets present in memory; therefore only a weak inference can be made that a nonlocated item is not in memory. Thus search models may possibly be able to explain accurate negative recognition responses to distractors, but only if a long and difficult search fails; conversely, fast negative recognition responses can be generated by search models but only at a considerable cost in accuracy.

It is of course possible to posit search models for recall, but assume that recognition operates differently, with direct and certain access to the image of the tested item. Model 1 had this property but had problems because the simple recognition component could not produce *searchlike* effects. One way out of this difficulty involves retaining a simple direct-access component for recognition but adding a search component as well, as indicated by Model 3 in Table 1. In two-phase models like Model 3, recognition can occur either on the basis of a simple *familiarity* process (i.e., D_s) or on the basis of a search (a slower process, on the average, than familiarity). These phases can occur sequentially or in parallel.

Two phase models have been proposed by Juola, Fischler, Wood, and Atkinson (1971; see also Atkinson & Juola, 1973, 1974; Atkinson, Herrmann, & Wescourt, 1974; Atkinson & Wescourt, 1975) and Mandler (1972, 1980; see also Mandler & Boeck, 1974; Mandler et al., 1969). The Atkinson and Juola (1973, 1974) models are illustrated in Figure 1. A direct-access familiarity process gives rise to a familiarity value that is used for judgment: If the value is quite high (above a *yes* criterion) the subject responds with a positive, rapid, recognition response. If the value is quite low (below a *no* criterion) the subject responds with a negative, rapid, recognition response. If the value is intermediate (between the criteria), then a search is carried out. Mandler (1972, 1980) has proposed a model roughly similar in general outline. Both the Mandler (1980) and the Atkinson and Juola (1973, 1974) approaches posit rapid familiarity decisions and slower search processes, on the average. It is this general approach to recognition that we set out to evaluate in a preliminary series of studies.

If it is assumed that the familiarity process, at least on the average, is faster than the search process (and almost entirely responsible for rapid, accurate, high confidence, negative responses), then it should be possible to separate the familiarity and search components experimentally. We reasoned that if subjects are forced to respond rapidly (producing, say, average response times of 500 ms) the responses would be based primarily on familiarity. How-

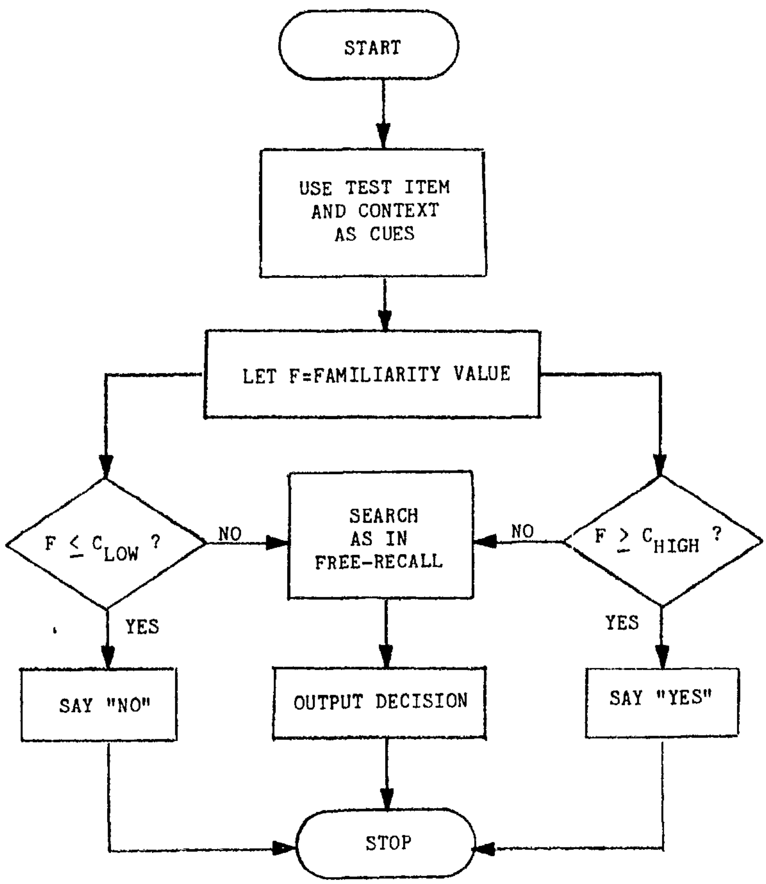

Figure 1. Retrieval from long-term store for recognition: a flow chart of a two-process recognition model. (The recognition response is based on either a familiarity judgment or a search process.)

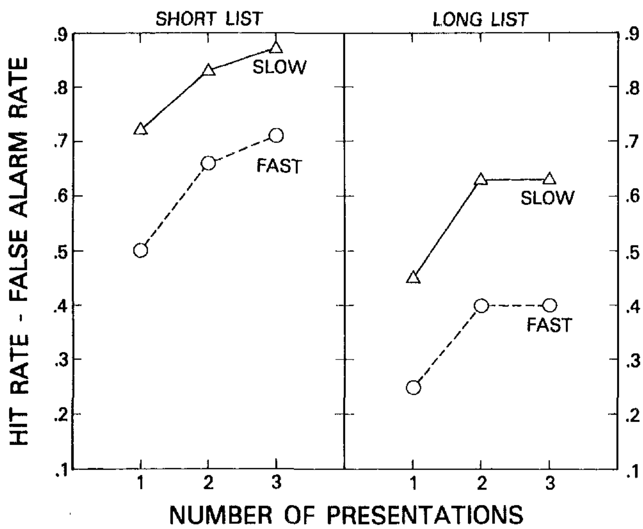

Figure 2. Recognition performance as a function of list length (36 or 96), number of presentations, and slow or fast recognition tests.

ever, if subjects are encouraged to take their time and are not even allowed to respond before 1 (or 1.5) s, then a larger proportion of responses should be based on search processes.

The experiments involved visual, sequential, presentations of words in lists. Following each list a yes–no recognition test was given: The list words were mixed with an equal number of distractor words and were tested one at a time. In fast tests, the items were presented every second, and subjects were required to respond to each item in under 900 ms. In slow tests, responses were not allowed before 1 (or 1.5) s had elapsed, and 3 (or 4) s were given between tests. In the fast test, subjects' latencies averaged about 500 ms, and in the slow test, the latencies averaged about 2.5 to 3 s (which compares with average latencies in unconstrained settings of about 800 to 1500 ms). Different lists were tested with slow and fast test techniques, with subjects practiced on the respective techniques and warned in advance of each test (but not each study) concerning the upcoming test speed. The results of three experiments that employed this procedure can be seen in Figures 2, 3, and 4.

The first experiment varied list length and number of presentations. The second experiment varied depth of encoding (pleasantness vs. pronounceability ratings) and number of presentations. In the third experiment, the type of distractor was manipulated.[1] In all situations, the variables had effects that were expected based on prior findings in the literature: Subjects were better on repeated items than on items presented only once (e.g., Atkinson & Juola, 1973; Ratcliff & Murdock, 1976). Performance was better on words rated for pleasantness than on words rated for pronounceability, although this factor did not quite reach statistical significance, $F(1, 71) = 3.21$, $p = .078$ (e.g., Jacoby & Dallas, 1981). Accuracy was higher for words from the short list than for words from the long list (e.g., Atkinson & Juola, 1973; Ratcliff & Murdock, 1976).

In the third experiment, the distractors on each list were of four types. Control distractors were not related in any systematic fashion to the list items. Other distractors were synonyms of particular list items, were similar graphemically to particular list items, or were similar both graphemically and phonemically to particular list items. (Also, the fast vs. slow test variable was a between-groups measure in this experiment.) Subjects made more errors (false alarms) on distractors that were related gra-

[1] The experiment on different types of distractors was completed with Tim Feustel. We gratefully acknowledge his contribution in all stages of the experiment.

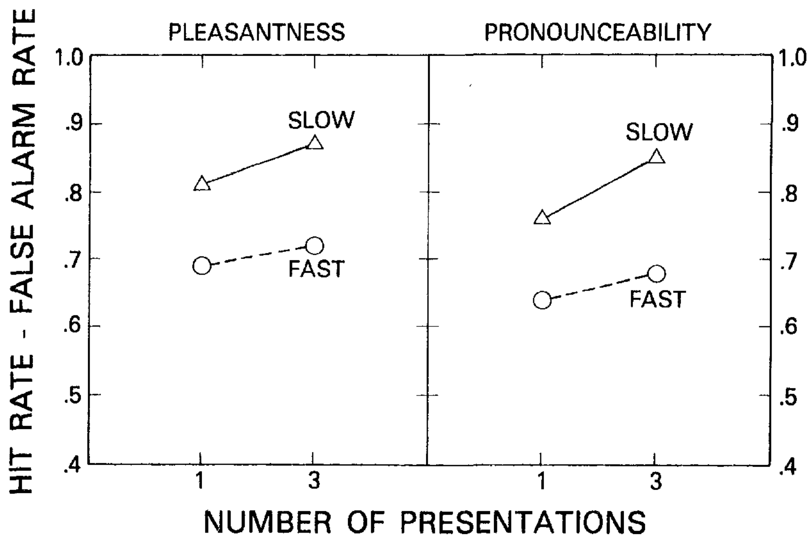

Figure 3. Recognition as a function of orienting task (pleasantness or pronounceability ratings), number of presentations, and slow or fast recognition tests.

phemically and phonemically to target items than to control items or synonyms. Such results have been found before (e.g., Cramer & Eagle, 1972; see also Eagle & Ortof, 1967; Nelson & Davis, 1972).[2]

In all three studies, performance was higher for slow tests than for fast tests—a result that is in accord with much work on speed–accuracy trade-offs—and research using the signal-to-respond technique (e.g., Dosher, 1981; Pachella, 1974; Reed, 1973; Wickelgren & Corbett, 1977).

Most important for present purposes, there were no significant interactions between the type of test (fast vs. slow) and any of the other variables in any of the three experiments.[3] Thus these variables responded similarly in the fast and in the slow test conditions. If search had occurred more often in the slow test conditions, and if search and familiarity respond differently to these variables, then an interaction would have been expected.[4] At the least, doubt is cast on Model 3 of Table 1. The findings therefore suggest either that search is not being utilized in these studies (Model 4 of Table 1) or that both search and familiarity processes are being utilized but that they respond to these types of variables in similar ways (Model 5 of Table 1).[5] Note that Model 4 is completely embedded in Model 5 and therefore cannot predict the data better than does Model 5. It is our plan in this article to explore a version of Model 4 because we do not wish to propose a more complex model unless the simpler version can be shown to fail. Furthermore, even if Model 4 can be shown to fail, its study could prove highly instructive.

Note that the direct-access component (D_c or D_s in Table 1) is often denoted *familiarity.*

---

[2] It may appear surprising that no more false alarms were made to synonyms than to control words in view of the large number of studies that have found higher false alarms for synonyms than for control words (for examples, see Anisfeld & Knapp, 1968; Grossman & Eagle, 1970). However, we have replicated this finding several times in our laboratory (see also Cramer & Eagle, 1972). One possible explanation is that the types of errors reflect the bias in encoding, and our subjects were emphasizing low-level features (acoustic or physical) rather than semantic features of the words. Some support for this notion can be found in the literature (e.g., Cermak, Schnorr, Buschke, & Atkinson, 1970; Davies & Cubbage, 1976; Elias & Perfetti, 1973). Other explanations are discussed later in this article.

[3] Although the larger advantage for physical similarity under fast testing in Figure 4 did not reach significance, the result, if reliable, may suggest that fast tests are more sensitive to low-level similarity features.

[4] The use of different methods of measuring performance in our studies, such as utilization of d' measures, did not change our findings or our conclusions.

[5] There is a possibility that no interactions occurred because familiarity is not faster than search; this is an unattractive alternative because the puzzle of rapid, accurate, negative recognition responses is difficult to explain.

This terminology seems appropriate for most laboratory recognition studies because the tasks are episodic and the decision is usually one of discriminating a recently presented item from preexperimentally known items. We use the terms *familiarity* and *direct access* interchangeably in this article.

We have assumed and defined the familiarity component to be faster on the average than the search process, because the search process must predict interresponse times in free recall that are many seconds in length, whereas the familiarity process must allow accurate negative recognition responses to be made in under a second (on many trials). Of course it is possible that the D_c or D_s process is a sequential one, similar to the search process but far more rapid (and efficient). The difficulty of distinguishing serial and parallel processes in general is well known (see Townsend, 1971), and we shall not try to do so here. For present purposes, it suffices to note that D_c (or D_s) cannot be a search process with temporal characteristics close to those needed for search in free recall.

We conclude this introduction by indicating the scope of the present article and the types of data we try to explain. Our present version of the SAM model is meant to deal with all of the recall phenomena mentioned in our earlier papers; these phenomena lie largely in the domain of episodic memory (see Raaijmakers & Shiffrin, 1980). The extension of the model to recognition is meant to apply primarily to situations in which long-term memory is being accessed (i.e., in which no contamination of recognition from short-term store [STS] occurs) and to situations in which subjects are asked to make episodic judgments that some item has occurred recently (when the distractors have not been presented recently and hence are unfamiliar).

The phenomena we wish our model to explain include the following: (a) the large number of factors that influence recall and recognition performance similarly, such as the amount of time given for study (Ratcliff & Murdock, 1976; Roberts, 1972), list length (e.g., Murdock, 1962; Roberts, 1972; Strong, 1912; see also Figure 2), test delay (e.g., Shepard, 1967; Strong, 1913; Underwood, 1957), and test position effects (e.g., Ratcliff & Murdock, 1976; Raaijmakers & Shiffrin, 1981a); (b) the factors that influence recall and recognition differently, such as word frequency (e.g., Gregg, 1976), types of encoding (e.g., Tversky, 1973), types of rehearsal (e.g., Woodward, Bjork, & Jongeward, 1973), and the effects of context shifts (e.g., Smith, Glenberg, & Bjork, 1978); and (c) the types of data that

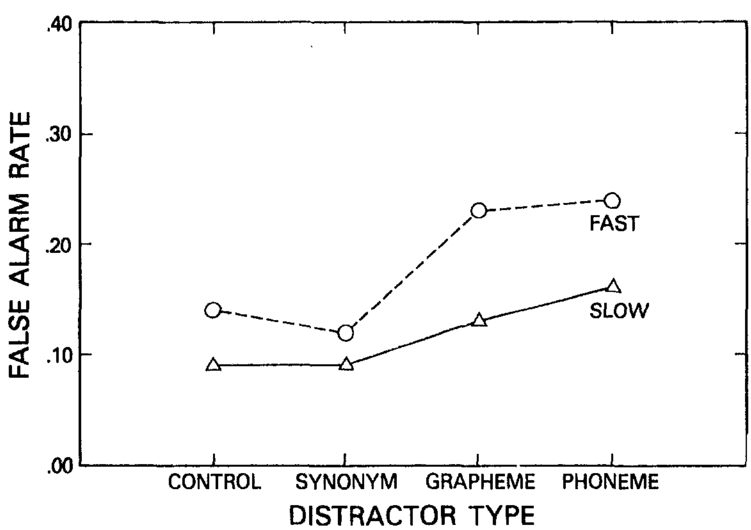

Figure 4. False alarm rate as a function of distractor similarity: control = not related; synonym = semantically related; grapheme = one letter in word altered, no rhyme; phoneme = one letter in word altered, and rhyme.

have been thought to indicate searchlike processes in recognition (e.g., Murdock, 1974; Tulving & Thomson, 1973).

1. A Model for Recall and Recognition

In Section 1, we present a model to predict the phenomena of both recall and recognition: We include (a) general features of the model that apply to both recognition and recall, (b) the recognition part of the model, (c) a summary of the model as it applies to recall (the recall model and applications have been explicated in detail in previous publications: Gillund & Shiffrin, 1981; Raaijmakers & Shiffrin, 1980, 1981a, 1981b), (d) a simple numerical example illustrating the quantitative features of the approach, and (e) certain important assumptions regarding storage and variability.

General Features of the SAM Model

Long-term store (LTS) is assumed to be composed of closely interconnected, relatively unitized, permanent sets of features called *images*. Episodic images contain many types of information, including the following: (a) contextual elements that can be used to identify the contextual-temporal setting in which an item occurs, (b) item information that can be used to name and identify the item encoded in the image, and (c) interitem information that links one image to others.

The process of retrieval is assumed to be cue dependent (Tulving, 1974). Access to LTS occurs when probe cues called a *probe set* are assembled in STS and are used to activate an associated set of information in LTS. The activated set consists of a large set of images activated to differing degrees. In recognition, some sort of integration across the entire activated set is used to generate a decision. In recall, one image is sampled from the activated set during each cycle of the memory search.

The quantitative details of the retrieval process are based on a retrieval structure (see the left-hand matrix in Figure 8). The retrieval structure is a matrix of retrieval strengths relating all possible cues to all possible images. All of these strengths are greater than zero, although many are of negligible magnitude. In theory, all cues and all images are contained in the structure, but in practice only the nonnegligible cues and images are considered. The retrieval strengths represent the relative tendency for a given cue to activate a given image and are related to the strength of the associative connections between that cue and that image.

In an episodic task, the entries in the retrieval structure are determined by two main factors. The first factor is the coding and rehearsal process used at the time of study (and any additional learning that occurs during the course of retrieval). The second factor is the match of the cue used at test to the nominally identical cue used at study: These might differ due to encoding variability and other factors. Particularly important cues in an episodic task are (a) context cues, which through similarity to the storage context give the subject access to recently presented items; (b) item cues, which have high strengths to the image containing the cue item itself and to the images containing items rehearsed with the cue item; (c) category cues, which have high (preexperimental) strengths to the images containing items from the cue category.

For computational simplicity, the retrieval structure is assumed to contain images of all presented items (rather than all images in memory). The context cue has strengths to each image in proportion to the time spent rehearsing that item. Each list item, as well as new items that are not from the list, can be a cue. If the cue item has not been rehearsed with the item in a given image, then the strength from item cue to image is set to a small residual value based on preexperimental factors; otherwise the strength is proportional to the time that both items had been rehearsed together. A category cue is given a high strength to images of items from that category and a smaller residual strength to images of items from other categories.

The SAM Model for Recognition

As indicated in the introduction, we are interested in considering Models 4 and 5 of Table 1. We present a version of the simpler model, Model 4, in spite of the fact that the basic logic of the SAM model requires that subjects should be able to carry out a memory search during recognition testing, if they choose to do so. Thus, for the purposes of the development, we adopt the working assumption that the subjects choose not to carry out a memory search during recognition.

Model 4 of Table 1 assumes that recognition is determined by a direct-access familiarity process, a process complex enough to produce characteristics that normally are the result of a search process. The model we propose does this in a very simple way. It is assumed that in a yes–no recognition test the subject probes memory with two cues: the context cue and the tested item. The total activation of LTS in response to this probe set is taken to be the value of familiarity on which a response is based. Thus when a target is tested, the activation is based not only on the activation of the image of that target, but also on the activations of all other images in memory. For this reason, we refer to the approach as a *global* model of familiarity. Searchlike effects are predicted because images of items other than the tested item are included in the activation on which a decision is based. The simplicity of the model is illustrated in Figure 5. In response to a test item, a value of familiarity is generated, and the subject responds *yes* or *no* depending on whether the familiarity value is above or below a criterion value chosen by the subject.

The quantitative details are captured in Equation 1. This equation gives the general expression for the total LTS activation when cues Q_1, Q_2, . . . , Q_M are used together to probe memory and when there are N images in LTS. The W_j are weighting factors associated with each cue, possibly reflecting the amount of attention allotted to each cue. (In previous work with the SAM model these weights had always been set to 1.0.) The term represented by the product gives the activation contributed by any given image, I_k. Thus the activation of a given image is just the product of the separate

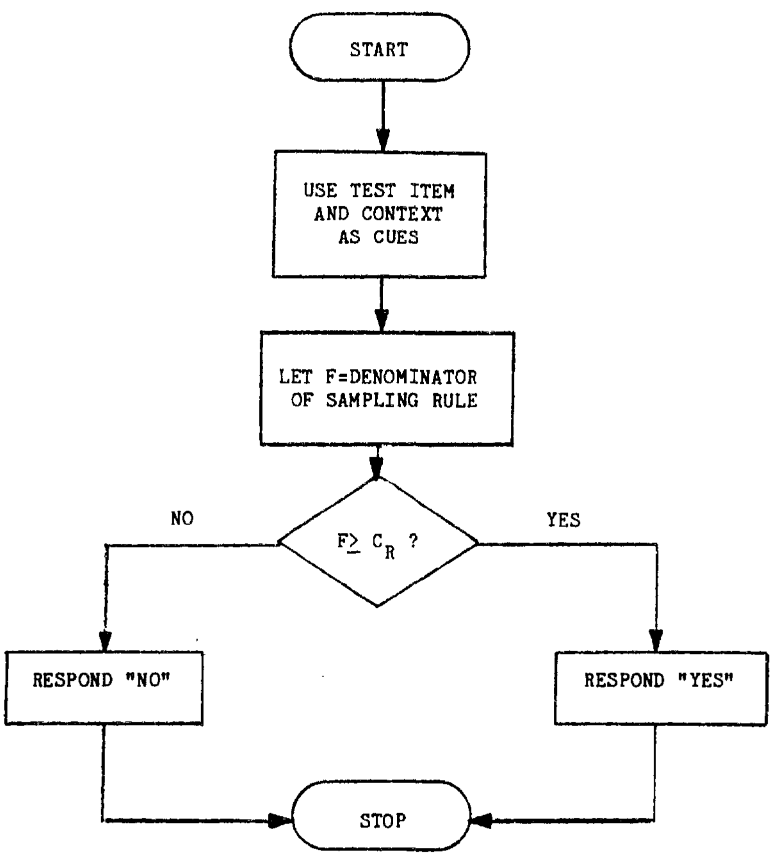

Figure 5. A flow chart for the basic SAM recognition process.

activations of that image by each of the cues, the individual activation being the strength in the retrieval structure, $S(Q_j, I_k)$. The total activation is just the sum of the activations for each of the images.

$$F(Q_1, Q_2, \ldots, Q_M) = \sum_{k=1}^{N} \prod_{j=1}^{M} S(Q_j, I_k)^{W_j}. \quad (1)$$

In a yes–no recognition test for the presence of single items on a recently presented list, we assume that just two cues are used to probe memory: the context cue, C, and the tested item, I_j. In this case, Equation 1 simplifies to Equation 2, and when the weights are set to 1.0, simplifies even further, as indicated in Equation 3. (For a numerical illustration of Equation 3, see the first three matrices along the top of Figure 8.)

$$F(C, I_j) = \sum_{k=1}^{N} S(C, I_k)^{W_c} S(I_j, I_k)^{W_j}. \quad (2)$$

$$F(C, I_j) = \sum_{k=1}^{N} S(C, I_k) S(I_j, I_k). \quad (3)$$

The familiarity values indicated in Equations 1, 2, or 3 can produce higher than chance recognition judgments because a few components of the sum making up the total activation will generally be higher for targets than distractors. In particular, the strength of a target to its own image, $S(I_j, I_j)$, will usually be higher than the strength of a distractor to that same image, $S(I_d, I_j)$. Furthermore, the inter-item strengths are sometimes higher for targets, $S(I_j, I_k)$, $k \neq j$, than for distractors, $S(I_d, I_k)$, $d \neq k$; this will usually be the case when the test item, I_j, had been rehearsed during list presentation with item I_k.

In applications to episodic tasks, the familiarity value is calculated under the simplifying assumption that only the strengths for list items contribute nonnegligible activations (i.e., the sum in Equations 1, 2, or 3 is taken only over the N list items). This assumption leads to what may seem an unusual feature of the model: When a distractor is tested, its own image does not contribute to its familiarity; a distractor is familiar only to the degree that it activates the list items.

Equations 1, 2, and 3 specify the values of familiarity that are produced by tests of targets and distractors. As indicated in Figure 5, a decision is made only after a criterion is chosen by the subject. Positive recognition judgments are made when the familiarity value is higher than the criterion. Therefore the recognition model is fully specified by Equations 1, 2, and 3, and by a set of assumptions governing the choice of criterion (see Sections 2 and 5).

The SAM Model for Recall

The SAM recall model is quite a bit more complicated than the model for recognition, because recall is assumed to require a memory search. The search consists of a sequence of sampling and recovery operations, each cycle in the sequence containing a sample of an image from LTS and an evaluation of the information recovered from the sampled image.

We begin our discussion of the recall model by considering the basis for sampling an image from LTS. During a given cycle of the search of memory, assume that cues, $Q_1, Q_2, \ldots, Q_M$ are used together as a probe set. Then Equation 4 gives the probability of sampling image I_i:

$$P_S(I_i|Q_1, Q_2, \ldots, Q_M) = \frac{\prod_{j=1}^{M} S(Q_j, I_i)^{W_j}}{\sum_{k=1}^{N} \prod_{j=1}^{M} S(Q_j, I_k)^{W_j}}. \quad (4)$$

The denominator of this equation is of course just the total activation of LTS, in response to the cues, that was the value of familiarity used to make a recognition decision. The terms are defined similarly. Equation 4 stipulates that sampling obeys a ratio rule (Luce, 1959) in which an image's chance of being sampled is proportional to its net strength to the probe cues. That is, the numerator is just the activation of a given image, whereas the denominator is the summed activation across all images.

The justification for the product rule by which individual cue strengths are combined into a single activation for a given image is easy to see in the sampling equation. The use of a product rule ensures that the sampled image tends to be an item strongly connected to *all* cues rather than to just one. In other words, the sampled image tends to come from the most dense region of the intersection of the associative fields of the separate cues. Thus the product rule allows the subject to use multiple cues to focus the search.

Equation 4 is somewhat easier to follow in the case that has been used in simulations of simple free-recall in previous work (Raaijmakers & Shiffrin, 1980, 1981b). In these cases the W_j were all set to 1.0, and only two types of probe sets were allowed: the context cue alone or the context cue plus one item cue. If the context cue alone is used, Equation 4 becomes Equation 5 (see the top row of the second, third, and fourth matrices along the top of Figure 8):

$$P_S(I_i|C) = \frac{S(C, I_i)}{\sum_{k=1}^{N} S(C, I_k)}. \quad (5)$$

If the context cue and an item I_j are used as cues then Equation 4 becomes Equation 6 (see rows 2 through 7 in the second, third, and fourth matrices in Figure 8):

$$P_S(I_i|C, I_j) = \frac{S(C, I_i)S(I_j, I_i)}{\sum_{k=1}^{N} S(C, I_k)S(I_j, I_k)}. \quad (6)$$

Once an item has been sampled, the subject attempts to recover the information contained therein and utilize it to succeed at the given task. In the case when recall is required, then the name encoded in the image is desired; in this case the probability of recovering the encoded name has been specified in earlier work and is given in Equation 7:

$$P_R(I_i|Q_1, Q_2, \ldots, Q_M) = 1 - \exp\{-\sum_{j=1}^{M} W_j S(Q_j, I_i)\}. \quad (7)$$

In this equation, the weights do not necessarily have to be identical to those used in Equation 4. In applications to free recall, however, the weights were set to 1.0 also. When context alone is the cue, we get Equation 8 (see the top row of the lower matrices in Figure 8):

$$P_R(I_i|C) = 1 - \exp\{-S(C, I_i)\}. \quad (8)$$

When context and an item I_j are cues, we get Equation 9 (see rows 2 through 7 in the lower matrices of Figure 8):

$$P_R(I_i|C, I_j) = 1 - \exp\{-S(C, I_i) - S(I_j, I_i)\}. \quad (9)$$

Equations 7, 8, and 9 are not quite as arbitrary as they look at first glance. Let $1 - \exp\{-W_j S(Q_j, I_i)\}$ be the probability of correctly identifying the item in the sampled image, for some particular cue, Q_j, used alone. Then Equations 7, 8, and 9 give the general expressions for correct recovery under the assumption that each different cue has an independent chance of producing such recovery.

Equations 4–9 describe what happens on a given cycle of the search. The whole course of retrieval consists of a series of such cycles. Figure 6 depicts the course of the retrieval process. On each cycle the subject chooses probe cues according to a retrieval plan that takes into account both prior knowledge and the outcome of the retrieval process to date. The overall plan may involve strategies such as alphabetic search or temporally ordered search or instead may involve a strategy based on the momentary outcome at each stage of search. After probe cues are selected, the relatively automatic sampling and recovery phases occur. Then the subject evaluates the outcome and decides whether to terminate or to continue search. In free-recall tasks, the termination decision is presumably based on the number or proportion of search cycles that are *failures* (these are cycles on which no new item is successfully recovered). If the search continues, then the cycle begins anew with the choice of probe cues.

To predict certain changes that take place during the course of retrieval, it is necessary to assume that some learning occurs during retrieval (such learning was termed *incrementing* in previous publications, in defiance of grammatical rules). In free recall, for example, each successful recovery is assumed to be followed by an increase in the strengths relating the cues in that probe to the image sampled (see Rundus, 1973).

Our brief review of the SAM model for recall may be completed by pointing out that separate recovery attempts that involve the same image and at least one common cue are not independent. In free-recall tasks, it is assumed that a new and independent recovery chance for an image (that has not yet had its item recovered) occurs if and only if at least one

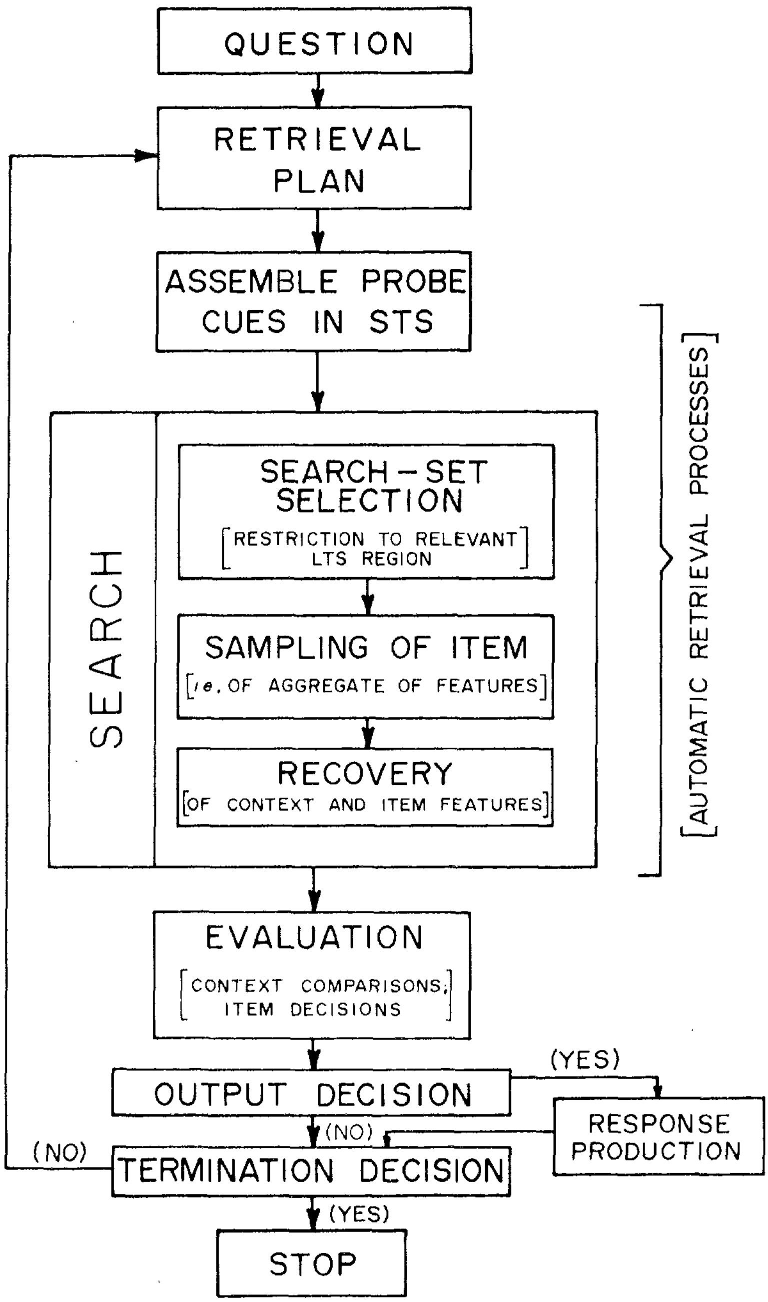

Figure 6. A generalized depiction of the course of retrieval from long-term store during recall (taken from Raaijmakers & Shiffrin, 1981b).

new cue (that has not been tried yet for the present image) is in the probe set.[6]

The manner in which the process of retrieval is implemented for a particular task is illustrated by the case of free recall. The flow chart for retrieval used by Raaijmakers and Shiffrin (1980) for this task is depicted in Figure 7. The retrieval structure is first filled according to a buffer-rehearsal process. Retrieval begins with the output of all items currently in STS (unless an arithmetic or other task is used to empty the words from STS before recall begins).

The course of LTS retrieval is indicated in the flow chart. Recall begins with the context cue alone. This cue continues to be used on successive search cycles until a word is recalled. Then an increment is given and the next probe set consists of the context cue plus the item just recalled. This probe set is used either until a new item is recalled or until L_{max} failures in a row occur; in the case of L_{max} failures, the item cue is dropped and the subject reverts to the context cue alone. Whenever a new item is recalled the probe cues to image strengths are given an increment, and the next probe set consists of context plus the just recalled item. Recall terminates whenever a total of K_{max} failures accumulate. Although strategies are surely idiosyncratic and variable, the free-recall strategy outlined here (and in Figure 7) seems to us a reasonable approximation that is fairly rational and simple to apply.

Note that cued-recall tasks are also easy to handle within the same framework. If an item is presented as a cue, and some other item is required as a response, it is natural to assume that the subject uses as a probe set the context cue along with the presented item. This probe set can be repeated until the response is found or the search is terminated.

Finally, note that the SAM recall model incorporates several processes in the recovery phase, implicitly. For example, in cued recall the recovery probability incorporates a decision process by which the subject decides a sampled item that can be named is in fact the response to the stimulus. More important for present purposes, the recovery process in free recall incorporates an implicit decision process by which the subject decides that a nameable sampled item was in fact on the recently presented list. This is the process denoted I in Table 1.

A Numerical Example

To help readers who are unfamiliar with the SAM model understand the implications of the equations and assumptions, we give a numerical example. The example is illustrated in Figure 8. The retrieval structure is indicated in the left-hand matrix. This gives the retrieval strengths linking cues to memory images. In this example it is assumed that four items are presented to the subject for study, resulting in an interconnected storage network containing four images corresponding to the four presented items. The images are denoted by I_j^* and the possible cues that might be used to probe memory are given along the left-hand margin of the matrix. The cue labeled C is the context cue (which effectively restricts the search to the just-presented list). The items of the list are possible cues and are denoted I_j. The items as cues are distinguished from the memory images of those same items because these are never identical (and sometimes can be quite different if, say, the encoding context shifts between study and test). The remaining cues denoted D_j are items that were not on the study list (and might be distractors in a recognition test).

The retrieval strengths in the matrix are assumed to result from the operation of rehearsal and coding processes during the study of the list. Note that all of the cues are connected by nonzero strengths to all of the images, even if they were not rehearsed together (as must have been the case for the D_j cues). The high strength connecting the context cue to I_3^* probably means that I_3^* had been rehearsed extensively, and the high strength relating I_3 to I_3^* (the *self-strength*) probably re-

[6] Because our recovery rule is consistent with an hypothesis that each cue has an independent chance of leading to recovery, it is perhaps more reasonable to assume that the recovery probability is based only on the strengths associated with the new cues and not on the strengths associated with the old cues that failed earlier. However, we have utilized the old assumption in the present article to be consistent with the earlier SAM papers. We doubt that important differences would result for the predictions in the present article if the other assumption is adopted.

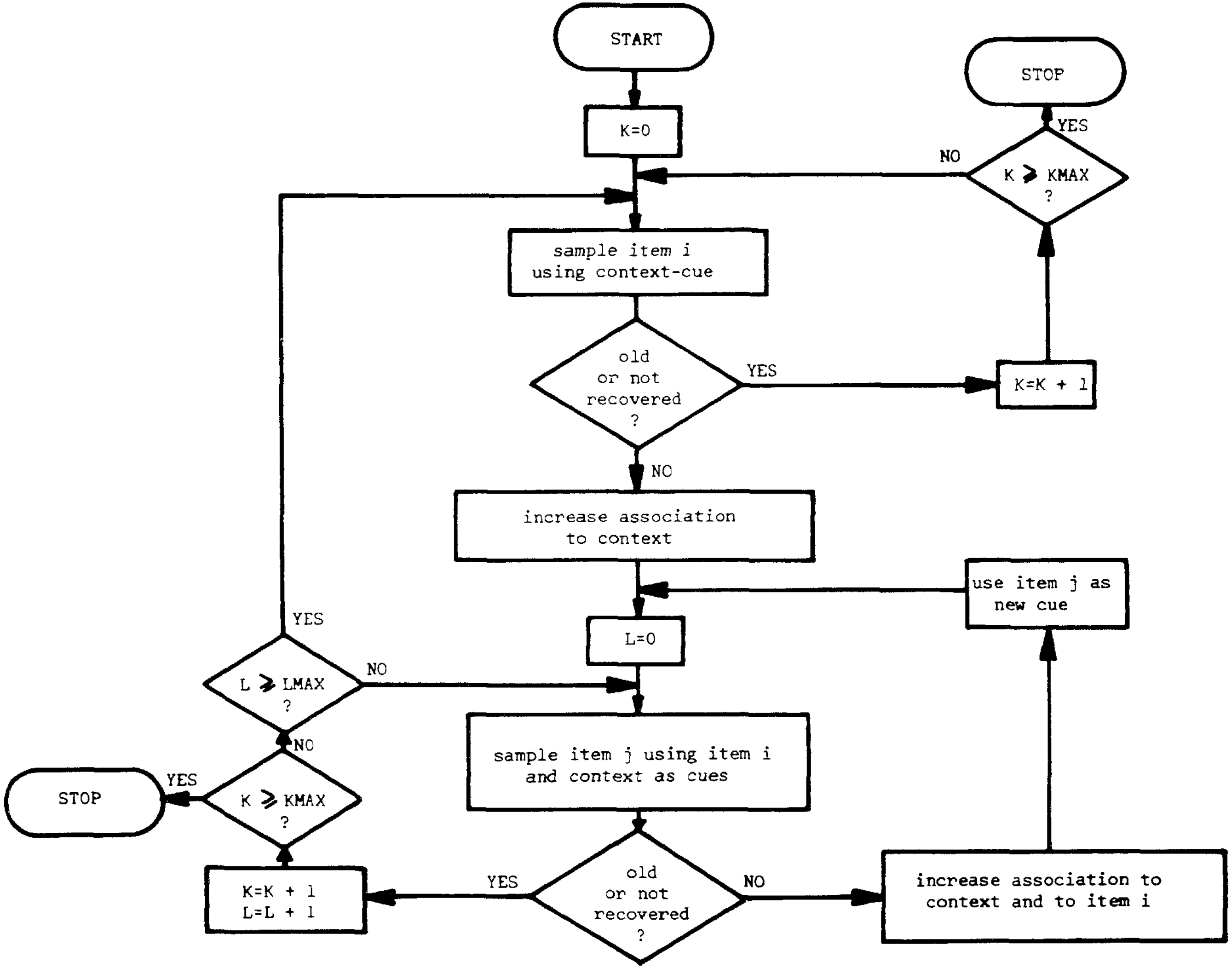

Figure 7. A flow chart of the SAM retrieval process for free-recall (taken from Raaijmakers & Shiffrin, 1980).

RETRIEVAL STRUCTURE

Memory Images

CUES	I_1^*	I_2^*	I_3^*	I_4^*
C	.5	.3	.8	.4
I_1	.3	.3	.4	.1
I_2	.3	.4	.1	.1
I_3	.4	.2	.7	.3
I_4	.1	.1	.2	.4
D_1	.1	.05	.1	.1
D_2	.2	.1	.3	.1

STRENGTH TO PROBE SET

PROBE SETS	I_1^*	I_2^*	I_3^*	I_4^*
C	.5	.3	.8	.4
$C+I_1$	.15	.09	.32	.04
$C+I_2$	.15	.12	.08	.04
$C+I_3$	.20	.06	.56	.12
$C+I_4$	.05	.03	.16	.16
$C+D_1$	.05	.015	.08	.04
$C+D_2$	.10	.03	.24	.04

SUM OF STRENGTHS

Familiarity of Probe Set
(2.0)
.60
.39
.94
.40
.185
.41

PROBABILITY OF SAMPLING AN IMAGE WITH THE PROBE SET

I_1^*	I_2^*	I_3^*	I_4^*
.25	.15	.40	.20
.25	.15	.53	.07
.39	.31	.20	.10
.21	.06	.60	.13
.13	.08	.40	.40
.27	.08	.43	.22
.24	.07	.59	.10

Sum of Strengths

	I_1^*	I_2^*	I_3^*	I_4^*
C	.5	.3	.8	.4
$C+I_1$	.8	.6	1.2	.5
$C+I_2$	.8	.7	.9	.5
$C+I_3$	.9	.5	1.5	.7
$C+I_4$	.6	.4	1.0	.8
$C+D_1$	.6	.35	.9	.5
$C+D_2$	.7	.4	1.1	.5

Recovery Probabilities

I_1^*	I_2^*	I_3^*	I_4^*
.39	.26	.55	.33
.55	.45	.70	.39
.55	.50	.59	.39
.59	.39	.78	.50
.45	.33	.63	.55
.45	.30	.59	.39
.50	.33	.67	.39

flects the same factor. The interitem strengths are not very high, but the value of 0.4 for I_3 to I_1^* may reflect some extra joint rehearsal of these two items. The residual values connecting the distractor D_1 to the list-item images are quite low, as is usually the case. However, the residual values for item D_2 are somewhat higher, possibly indicating a higher similarity of this distractor to (some of) the list items than was true for item D_1. Finally, note that even in cases where rehearsal and coding are not factors (e.g., the rows of strengths for the distractor cues), the strengths are somewhat variable. This variability is quite important; as we see, without variability, recognition performance would generally be perfect.

According to the SAM model, the context cue is always used to probe memory in episodic tasks but sometimes is combined with other cues, in this case the item cues. The possible probe sets size of 1 or 2 are given on the left-hand margin of the second matrix. The numerator of Equation 6 specifies that the strength when cues are combined is the product of the individual cue strengths. These products are given in the second matrix (except for the first row, for the context cue alone, which is determined by the numerator of Equation 5). Thus the strength connecting probe set ($C + I_1$) to I_1^* is given by the product of the strength connecting C to I_1^* (i.e., 0.5) and the strength connecting I_1 to I_1^* (i.e., 0.3) for a result of 0.15.

For a given probe set the sum of the strengths to all the images is the familiarity value of Equations 1, 2, or 3 and also the denominator of Equations 4, 5, or 6. We treat this value as the total LTS activation engendered by the probe set, and the results are given in the column vector of Figure 8. If the subject were carrying out a recognition test with targets I_1–I_4 and distractors D_1 and D_2, a criterion would be chosen and responses made accordingly. Because the familiarity value of D_2 is higher than that of the two list items, I_2 and I_4, errors occur wherever the criterion is placed. Thus if the criterion were set at 0.402, a *false alarm* would be given to D_2 and a *miss* would be given to I_2 and I_4.

Equations 4, 5, or 6 determine the sampling probabilities for each image for a given probe set, for each cycle of the search process. The probabilities are simply the numbers in the rows of the second matrix in Figure 8 divided by the row sums given in the column vector. The resulting sampling probabilities are given in the right-upper matrix of Figure 8. Note that the absolute values of the strengths in the first two matrices along any row are not important for sampling; sampling is based only on the relative strengths along any row of the second matrix. If learning takes place during retrieval, then some of the strengths in the retrieval structure increase and, consequently, some of the sampling probabilities change during the course of retrieval.

Once an image is sampled, there is no certainty that a recovery of the information in that image will be complete enough that the name of the item embedded in that image will be recallable. The probability of recovery is given by Equations 7, 8, or 9 and is based on an exponential function of the sum of strengths for the cues in the probe set. The sums of strengths are given in the lower left-hand matrix of Figure 8, and the corresponding recovery probabilities from Equations 8 and 9 are given in the lower right-hand matrix. These probabilities apply whenever an image is sampled by a probe set that has not previously been used to sample that image.

Storage and Variability Assumptions

The values in the retrieval structure are the result of (a) rehearsal and coding processes that take place during the study of the list, (b) preexperimental associations, and (c) the match of the cue encodings at study and test. Storage assumptions based on a *rehearsal buffer* have been used in previous applications of SAM to free and cued recall. However, these previous storage assumptions cannot be used without modification in applications to rec-

Figure 8. A numerical illustration of a possible retrieval structure for a four-item list with context, item, and distractor cues. (The remaining matrices give quantities that are used during the application of SAM to recognition and recall.)

ognition. The previous storage assumptions allowed variability in the strengths placed in the retrieval structure only to the degree that the amount of rehearsal (or coding) varied. Thus the residual values placed in the structure (when two items were not rehearsed together) were set to a fixed value. Because the residual value was low, and less than the self-strength for any list item, every list item would have a higher familiarity value than every distractor. This is evident from inspection of Equation 3: Both targets and distractors share the values of the context strengths, $S(C, I_k)$; for the distractors, each interitem strength, $S(I_j, I_k)$, would attain a constant, small, residual value; for list items, some of the interitem strengths would be higher than the residual value due to rehearsal of an item I_j. The result of these assumptions is the prediction that recognition performance would be perfect.

Obviously what is needed for realistic recognition predictions is variability of the strengths placed in the retrieval structure. Familiarity distributions for targets and distractors would then have the possibility of overlapping. Because items obviously differ (as do many other factors besides rehearsal time), variability is a very reasonable assumption. Variability was not assumed in previous publications concerning free recall because it was not needed, but now that we introduce it for recognition, we also use the assumption for recall, and generate predictions accordingly.

Introduction of variability opens many degrees of freedom for the model. We decided to adopt a very simple approach, at least at the outset: After the retrieval structure is filled in the usual fashion, each entry is replaced by a random sample from a 3-point distribution. In particular, if the starting value is x, then the final stored value, y, is chosen according to Equation 10.

$$y = \begin{cases} (1 - v)x, & p = 1/3 \\ x, & p = 1/3. \\ (1 + v)x, & p = 1/3 \end{cases} \qquad (10)$$

This discrete distribution is sufficient for our purposes because the familiarity value results from a sum across the list items; the law of large numbers assures that the distribution of the sum approaches a continuous distribution as the list length becomes large (actually approaching a normal distribution for the distractor tests).

Although the choice of Equation 10 is fairly arbitrary, it has a property that we regard as desirable: The standard deviation associated with a value in the retrieval structure is a fixed multiple ($v\sqrt{2/3}$) of that value. Such an assumption is not compatible with a hypothesis that the variance associated with a strength based on a storage episode of time 2t is the sum of variances associated with the two successive storage episodes of time t for the same item (because storage is linear with time in our model). This *independence* assumption would lead to a linear growth of variance, rather than standard deviation, with strength. Our present variability assumption is compatible with a *dependence* hypothesis in which the random deviation associated with the strength for any subinterval of time for an item is in the same direction as the deviation associated with any other subinterval. This variability assumption has what we regard as another desirable property: Multiplication of all the values in the retrieval matrix by any constant (before noise is added) leaves unchanged the shapes and overlap of the familiarity distributions for targets and distractors.[7] Other assumptions concerning variability are obviously possible but are not explored in this article.

Summary

Our model for recognition and recall is of the type labeled *Model 4* in Table 1. Recall is carried out with a search process (containing an implicit episodic identification process embedded in the recovery phase of the search cycle). Recognition is carried out with a direct-access familiarity process that is of the global type because familiarity is a sum of familiarity values for the images of each list item. Recall and recognition are linked quantitatively be-

[7] Because overall level of strength has an effect on recovery but not on sampling (see Equations 4 and 7), and because recognition performance is not dependent on overall level of strength, the strength level may be used to adjust the relative performance level of recall versus recognition. For this reason, it is possible to choose a value of v fairly arbitrarily, within certain limits, and adjust the other parameters to fit the combined recognition–recall data.

cause the familiarity value used in recognition for item and context cues is the denominator of the sampling ratio used during search when that item and context cue make up the probe set.

2. The Simulation Model and Qualitative Predictions

In Section 2 we first describe a particularly simple simulation model, its parameters, and the choices of parameter values. The model is then used to produce qualitative predictions for a number of findings in the recall and recognition domains.

The Simulation Model

The model's predictions are all generated with a computer simulation because analytic derivations are in most cases not feasible. The basic situation to be modeled assumes presentation of a list of items followed by a recognition or a recall test. The storage and retrieval assumptions follow.

Storage

A buffer rehearsal process is assumed (Atkinson & Shiffrin, 1968). In this version a particularly simple buffer is assumed. The buffer contains a maximum of r items. When the buffer is full, each new item replaces the oldest item in the buffer.[8] Based on earlier work, r was set equal to 4 (see Raaijmakers & Shiffrin, 1980).

We assume that the items in the buffer share rehearsal time—when n items are present, each gets $1/n$ of the available rehearsal time. (Thus primacy is predicted because the first item is alone and several items are needed to fill the buffer.) In particular, a mean of a^* units of strength are stored between an item and the context cue for each second of rehearsal; a mean of c^* units of strength are stored between an item when used as a cue and its own memory image, for each second of rehearsal; a mean of b^* units of strength are stored between two different items, for each second of individual rehearsal for one of the items that takes place during the period that both items are rehearsed together; and a mean of d units of strength are present in the retrieval structure between any two items that have never been rehearsed together. For studies using random word lists, the value of d is assumed to be small relative to b^* and represents preexperimental associations between the items. Note that a distractor as a cue is given mean interitem strength values of d, like a list-item cue that has not been rehearsed with any other list items.

Retrieval During Free Recall

The flow chart shown in Figure 7 is used to govern retrieval, but for simplicity no increments are utilized (that is, the increment parameters are set to 0). The two retrieval parameters that govern cessation of search are chosen fairly arbitrarily, although with previous fits in mind, and are as follows: $K_{max} = 30$, $L_{max} = 8$.

Retrieval During Recognition

The value of familiarity is calculated using Equation 8. A criterion, Cr, is chosen. If the familiarity value is higher than the criterion, then a yes recognition response is given, otherwise a no response is given. Sensible criterion choices are generally very near the point where the distributions of familiarity for targets and distractors cross. An example is given in Figure 9.

How the subject learns where to place the criterion is an interesting question that is discussed later. For present purposes it is simply assumed that the criterion is fixed in value from the start of the recognition test until the end. In every simulation in this section we assume that the criterion is placed at a point 0.455 times the mean value of the distractor distribution plus the mean value of the target distribution; that is, $Cr = 0.455\ (M_d + M_t)$. This criterion choice was used because for the present parameters it places the criterion

---

[8] An assumption of oldest item deletion predicts a recency effect of size r, which surely doesn't match the known data when $r = 4$. However, because in this article we predict only long-term performance (say, in situations where STS is cleared before test), our buffer assumption makes very little difference, and we use oldest item deletion for ease of calculation.

somewhere close to the crossing points of the target and distractor distributions.

Strengths in the Retrieval Structure

A given strength in the retrieval structure between a cue and an image depends not only on coding processes that take place at the time of study but also on the match of the encoding given to the cue at test to the encoding given that cue during storage. For example, increasing the delay of test usually causes the context cue at test to differ more from the context cue used at study. As another example, a change in the item preceding the test item could cause different meanings of the item to be chosen at test and study.

We can capture this factor in the model by setting the strength in the retrieval structure equal to the originally stored strength times a factor indicating the degree of match between the two encodings of the cue: $S(Q_i, I_j) = m(Q_i)S^*(Q_i, I_j)$, where S^* represents the originally stored strength and $m(Q_i)$ represents the degree of match of the encodings of the cue at test and study. Our storage assumption lets us produce the strength values by setting $a = m(C)a^*$, $b = m(I)b^*$, and $c = m(I)c^*$, and letting the retrieval strengths equal a, b, and c times the appropriate number of seconds of rehearsal (C is the context cue and I is an item cue).

It is convenient to couch the predictions of the model in terms of a, b, and c. When experimental manipulations are used that should shift the degree of match (e.g., delay of test or biased encoding), these can be described as shifts in a, b, or c rather than in the m parameters. This approach keeps the present treatment identical to that used in earlier papers on the SAM model.

Simulation Procedures

For the set of qualitative simulations, an arbitrary parameter set was chosen, roughly in accord with previous simulations of free-recall data. The values were $a = 0.25$, $b = 0.2$, $c = 0.15$, and $d = 0.075$. In addition, the variability parameter, v, was set equal to 0.5. The patterns of predictions that result are not dependent on the exact values chosen (although, of course, quantitative performance levels do depend on these values).

For a given application, simulated data is generated for 500 lists for both recall and rec-

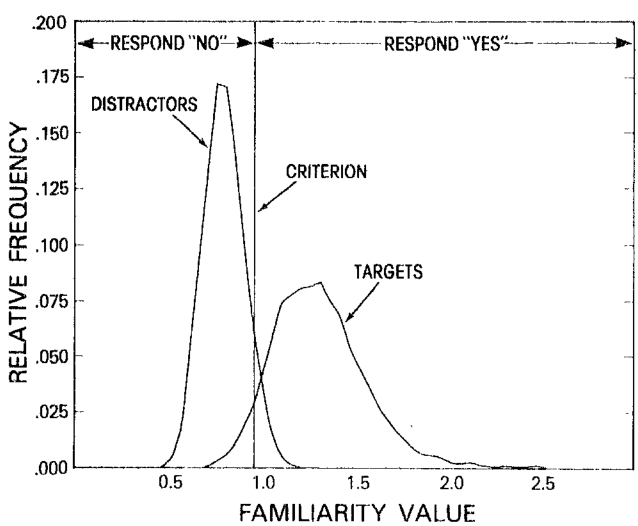

Figure 9. A target and distractor distribution. (If the value is greater [to the right of the criterion] a *yes* response is made. If the value is to the left of the criterion, a *no* response is made. The distributions are generated from the model applied to a 20-item list at 2 s per item. A buffer size of 4 is assumed, with oldest item deletion. Parameter values are $a = 0.25$, $b = 0.20$, $c = 0.15$, $d = 0.075$, and $v = 0.5$.)

ognition (using different randomizations each time). These pseudodata are averaged and the results are plotted as the predictions of the model. Because we are not trying to produce quantitative predictions, we do not plot actual data. Instead the literature is reviewed and the qualitative patterns that are observed in the literature are described. Recall predictions are given in terms of probability of correct recall and recognition predictions in terms of d'.[9] (Additional information is provided in Appendix A.)

Roles of the Parameters

The roles played by the various parameters become clear in the following parts of this article, where applications to various paradigms are discussed. It is useful nevertheless to summarize briefly the main functions of the parameters. To demonstrate the actions of the parameters most clearly, predictions were derived for a standard set of parameter values, and then each value was varied in turn, while holding the other values constant. The results are given in Figure 10 for recognition (d'), and in Figure 11 for free recall.

Parameter v. For recognition paradigms, decreases in the variability parameter, v, decrease the variability of the entries in the retrieval structure and hence decrease the variabilities of the target and distractor distributions, without changing the mean difference. Performance therefore improves, as indicated in the right-hand panel of Figure 10.

In free recall, changes in the v parameter have an effect on the recovery process. Because recovery probability is an exponential function of strength, moving from a smaller to a larger variance (increases in v) produces a lowering of the average recovery probability. The results are shown in the right-hand panel of Figure 11.

Parameter a. The context parameter, a, acts as a scale factor in recognition and does not affect performance. (For example, multiplying the value of a by a factor k causes all the entries in the top row of the retrieval structure of Figure 8 to be multiplied by k. In turn, this causes all entries in the adjacent matrix in Figure 8 to be multiplied by k and, hence, all the familiarity values to rise by a factor of k. Obviously the ordering of magnitudes of targets and distractors are not affected by multiplication by a constant.) The results are shown in the left-hand panel of Figure 10.

In recall tasks, increases in the a parameter cause increases in recovery probabilities (and no effects on sampling probabilities) and, hence, produce increases in recall, as shown in the first panel of Figure 11.

Parameter b. Increases in the interitem parameter, b, improve recognition performance by raising familiarity of the target distribution, but not the distractor distribution. Thus only the rows labeled I in the retrieval structure of Figure 8 are altered by changes in b. The results are shown in the second panel of Figure 10.

Increases in b also improve free recall by improving recovery and by altering sampling in subtle ways (e.g., decreasing self-sampling when an item cue is used—see next paragraph). The results are indicated in the second panel of Figure 11.

Parameter c. Increases in the self-coding parameter, c, improve recognition for similar reasons as in the case of the b parameter: Only the target distribution is raised (one of the entries in each of the rows labeled I in the retrieval structure of Figure 8 are raised). On the other hand, increases in the c parameter harm free recall by increasing self-sampling (self-sampling is the tendency for an item used as a cue to sample its own image). The opposing results are shown in the third panel of Figures 10 and 11.

Parameter d. Increases of the residual parameter, d, reduce recognition by increasing the familiarity of the distractors much more than the familiarity of targets (e.g., all the entries in the lower two rows of the retrieval structure of Figure 8 are increased, but only a few entries in the rows labeled I are increased) and by increasing the variance of both distributions. Increases in the d parameter im-

[9] Because the theoretical target and distractor familiarity distributions are of unequal variance, and because the target distribution is skewed, it is technically incorrect to analyze the data in terms of d'. Nevertheless, because most actual data are analyzed this way, we have converted the hit and the false alarm rates into d' scores. The patterns of predictions are unaltered by alternative measures of discriminability.

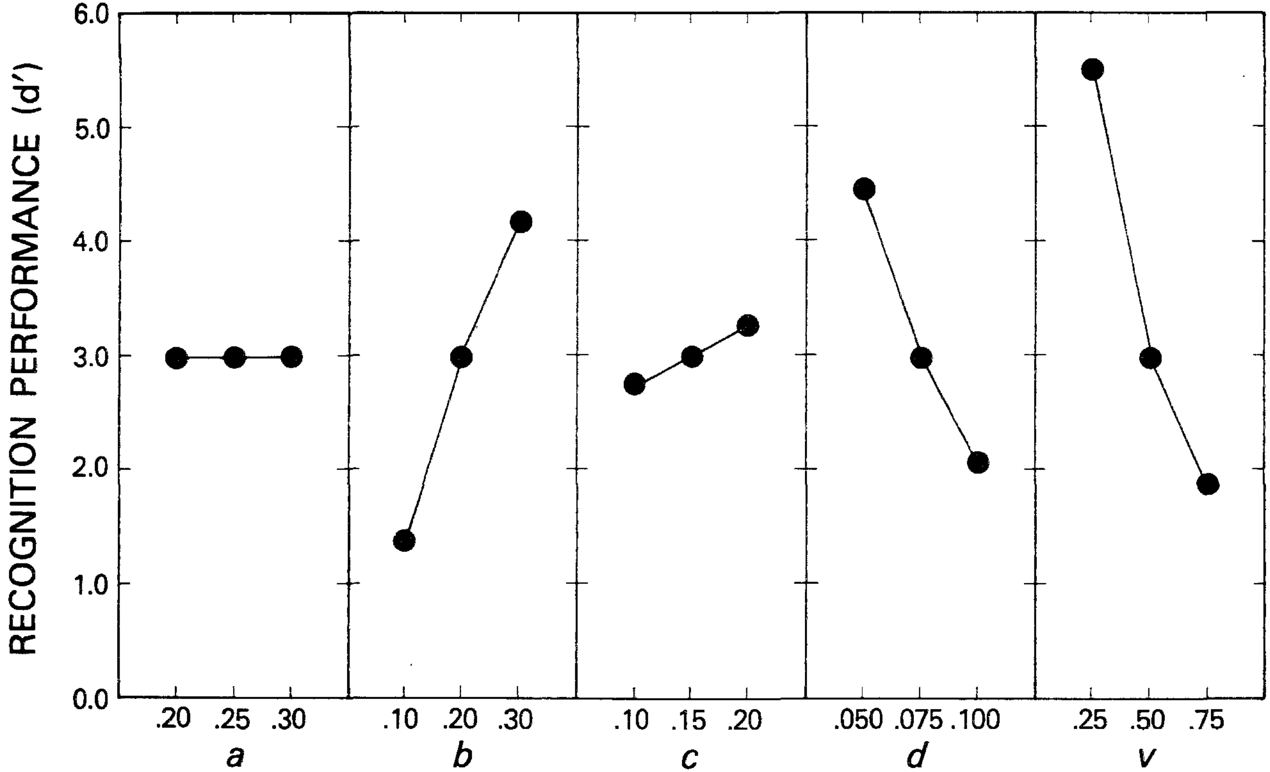

Figure 10. Changes in predicted recognition performance (d') for a 20-item list caused by changes in the a, b, c, d, and v parameters. (Unless otherwise indicated, t = 2 s; $r = 4$; $a = 0.25$; $b = 0.20$; $c = 0.15$; $d = 0.075$; $v = 0.5$.)

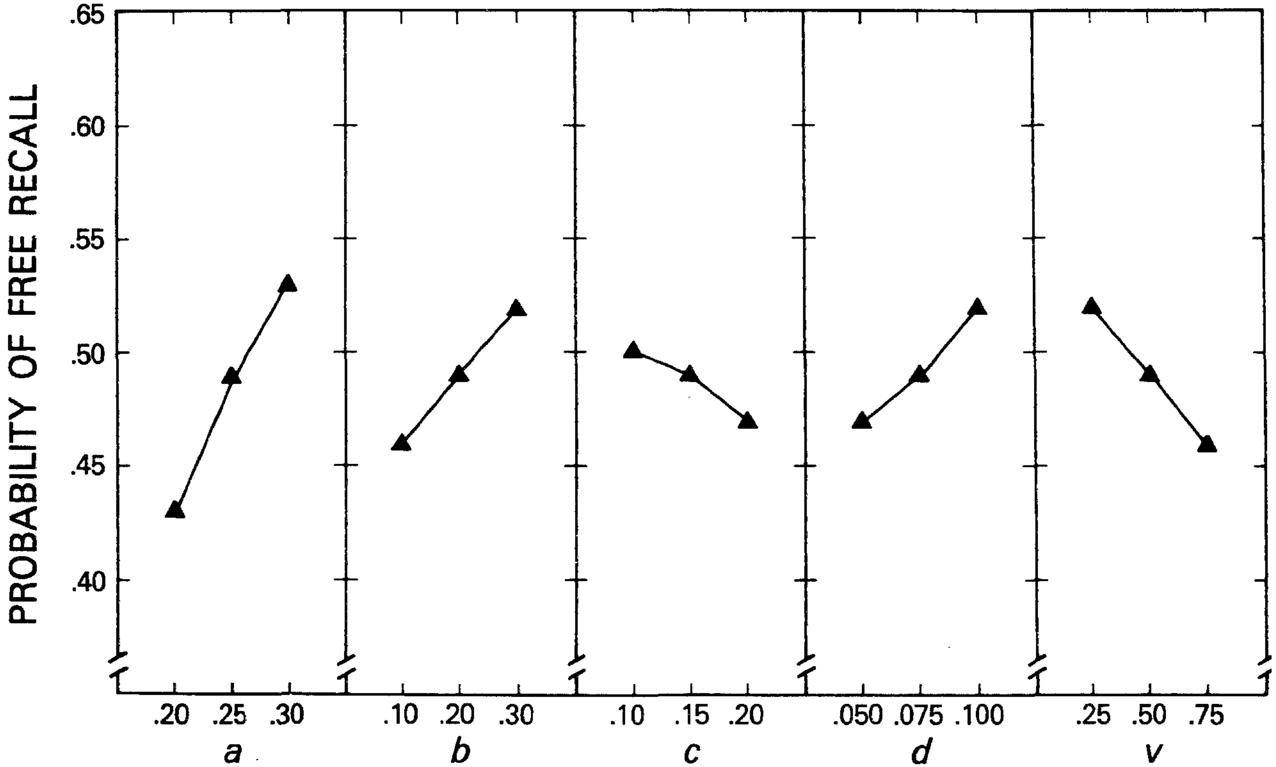

Figure 11. Changes in predicted recall performance for a 20-item list caused by changes in the *a*, *b*, *c*, *d*, and *v* parameters. (Unless otherwise indicated, t = 2 s; $r = 4$, $a = 0.25$, $b = 0.20$, $c = 0.15$, $d = 0.075$, $v = 0.5$; $K_{max} = 30$, $L_{max} = 8$.)

prove free recall mainly by increasing recovery probabilities and also through subtle effects on sampling (e.g., decreasing self-sampling). The predictions are shown in the fourth panels of Figures 10 and 11.

K_{max} and L_{max}. These parameters apply only to free recall, and only K_{max} has large effects: Increases in K_{max} are equivalent to extending the length of search and produce better free recall. We do not vary K_{max} in any of the applications in this article.

Applications of the Recall and Recognition Model

We illustrate the workings of the new model by applying it to a variety of results from the literature. In most of the cases we contrast the predictions of the model for a recall task with the predictions of the model with the same parameters and assumptions for a comparable recognition task. We are not interested in exact quantitative fits of data; rather we show that the model (usually with a standard parameter set) fits the qualitative patterns that are observed in the data. (Detailed quantitative fits are carried out in Section 3 of this article.) The figures plot recognition performance in terms of d'. The means of the target and distractor distributions and the hit and false alarm rates for each point are given in Appendix A.

The List-Length Effect

It is well known that the probability of recalling or recognizing a particular item from a list decreases as the list length increases. This is true for recall (e.g., Murdock, 1962; Roberts, 1972) as well as for recognition (e.g., Strong, 1912; see also Figure 2).

Figure 12 gives the results of the model's predictions for lists of words presented for 2 s each. Recall and recognition are plotted together, but we do not wish to imply that d' and the probability of recall are in any way directly comparable. We show the results in this fashion simply to demonstrate that performance decreases for both measures.

The list-length effect occurs in free recall as a result of the sampling rules. The probability of sampling a given item on a given search decreases as the list length increases. Even though more total samples are made for longer lists before the stopping criterion is reached (because the probability of resampling a previously sampled item goes down with list length), the sampling effect is the more powerful and recall per item decreases with list length.

In recognition, quite a different mechanism is responsible for the list-length effect. The mean difference in familiarity for targets and distractors remains constant as list length increases. This is so because only a few items are jointly rehearsed with any target item, regardless of list length (beyond some minimum). Therefore, adding list items plays the same role for target and distractor tests—in each case images are added to memory that are only residually associated to the test item. Why then does performance decrease? The answer is simple: The variance of the distributions increases with increasing list length (because each additional item in the list adds an independent variance component). As a result, there is more overlap of the target and distractor distributions for the longer lists and d' is decreased.

Presentation Time

Increasing the presentation time per item improves both recall and recognition performance (e.g., Ratcliff & Murdock, 1976; Roberts, 1972; Shiffrin, 1970). In SAM, increasing presentation time increases the strengths of associations built up during rehearsal: Context, interitem, and self-strengths all get larger. In recall these increases improve the probability of recovery (the ratio rule for sampling does not change with presentation time).

In recognition, the increase in d' arises because the self-strength and interitem strengths rise with presentation time. These increases cause only the target distribution to increase (the rise in mean far outweighs the concomitant rise in variance), thereby improving performance. Note that the increase in context strength does not affect d'—the increase in context strength acts like a scale factor increasing both mean familiarity and variances so that no net change in d' results (see the discussion of the a parameter). The results of the model's predictions for presentation time

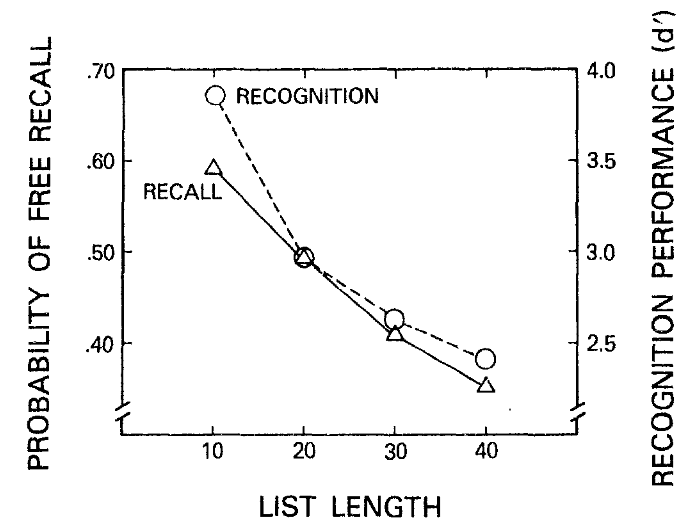

Figure 12. The model's predictions for recall and recognition as a function of list length. (t = 2 s; $r = 4$, $a = 0.25$, $b = 0.20$, $c = 0.15$, $d = 0.075$, $v = 0.5$; $K_{max} = 30$, $L_{max} = 8$.)

for a list of 30 words can be seen in Figure 13.

Encoding Effects

A large literature exists on the effects of different types of encoding operations on recall and recognition performance. Consider first the effects of item-specific, or maintenance, rehearsal. Woodward, Bjork, and Jongeward (1973) presented a series of words, each of which was followed by 0, 4, or 12 s of delay. After the delay there was a cue to remember or to forget the word. It was assumed that the delay would be used for maintenance rehearsal until the cue was given. At that time subjects

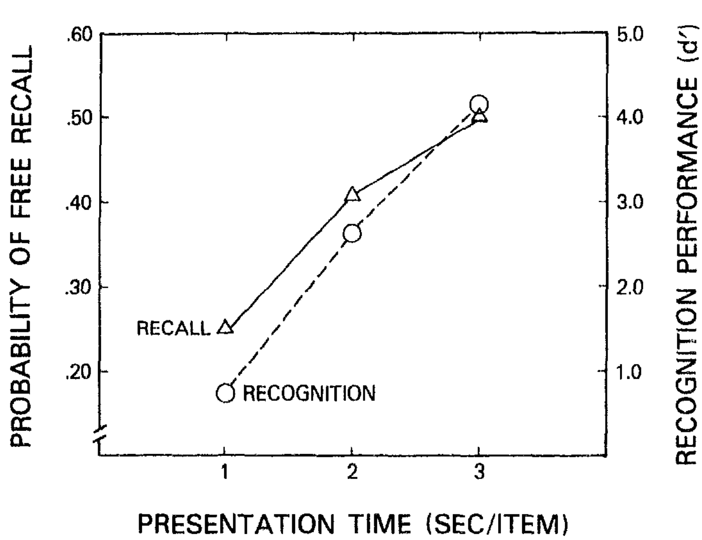

Figure 13. The model's predictions for increasing amounts of presentation time per word. (List length = 30; $r = 4$, $a = 0.25$, $b = 0.20$, $c = 0.15$, $d = 0.075$, $v = 0.5$; $K_{max} = 30$, $L_{max} = 8$.)

could either quit processing the word or give it more elaborate processing. Subjects were then given a test for the to-be-remembered words only. In a surprise end-of-session free-recall or recognition test for all words, Woodward et al. (1973) found no advantage of extra maintenance rehearsal for recall performance but a significant advantage for recognition (see Glenberg & Adams, 1978; Glenberg, Smith, & Green, 1977).

We assume that the effect of maintenance rehearsal is to cause repetitions of the features related to the physical aspects of the item, so self-coding, c, should increase with the amount of rehearsal. We also assume that extra context and interitem coding due to increased maintenance rehearsal are reduced to a minimum. These assumptions are quantified by beginning with the standard parameter set of the preceding figures and then modeling changes in maintenance rehearsal by changes in the c parameter. The results are depicted in the third panel of Figures 10 and 11, for the parameter values $c = 0.10$, 0.15, and 0.20, from left to right, respectively.

Figure 11 shows recall to fall with increases in c because self-sampling is increased: When an item is recalled and used as a cue it tends to sample its own image, reducing the chances of sampling something as yet unrecalled. Perhaps context coding does increase slightly with maintenance rehearsal, canceling this drop. The predictions when the c values are 0.10, 0.15, and 0.20, as before, but $a = 0.23$, 0.25, and 0.27, from left to right, are given in Figure 14.

An interesting result by Nairne (1983) seems superficially at odds with the predictions just given. He found in several conditions of his study that increasing amounts of maintenance rehearsal of pairs of items did not improve recognition performance on a later surprise test. The answer lies in the nature of the distractors used: The distractors consisted of two list items, one from each of two different studied pairs. In such a case, it turns out that the b parameter, rather than the c parameter, is the crucial determinant of performance. Because the b parameter is not assumed to rise in this type of study, little effect is predicted. The reasons underlying the predictions for these special types of stimuli and distractors are covered in greater detail in Section 4.

Consider next the experiments on intentionality of encoding and test expectancy. These paradigms may produce variations in

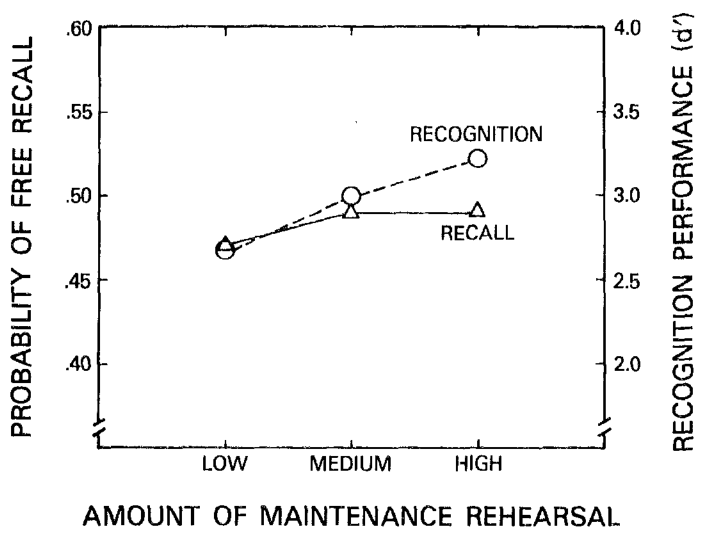

Figure 14. Predictions from the model for increasing amounts of maintenance rehearsal. (Increasing maintenance rehearsal is assumed to increase slightly context coding [a increases from 0.23 to 0.25 to 0.27, from left to right] and to increase substantially self-coding [c increases from 0.10 to 0.15 to 0.20, from left to right], whereas interitem encoding remains constant [$r = 4$, $b = 0.2$, $d = 0.075$, $v = 0.5$; $K_{max} = 30$, $L_{max} = 8$; 20-word list at 2 s/item].)

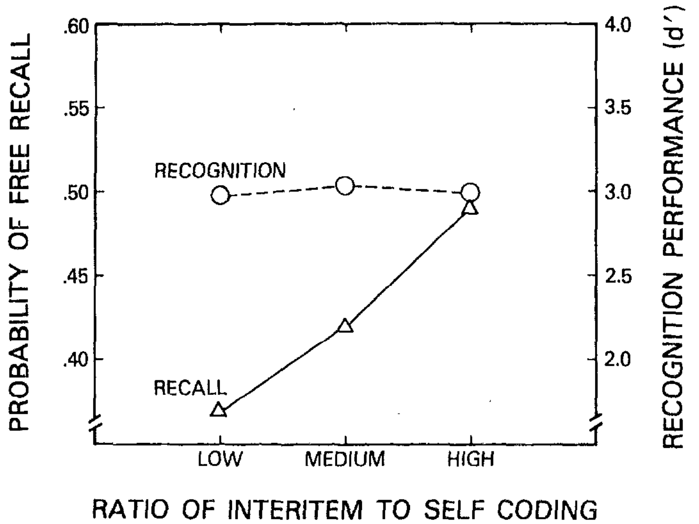

Figure 15. Predicted recall and recognition performance as interitem coding increases (b increases) and self-coding decreases (c decreases). (From left to right, $b = 0.1$, 0.15, and 0.2; and $c = 0.51$, 0.33, and 0.15. $r = 4$, $a = 0.25$, $d = 0.075$, $v = 0.5$; $K_{\max} = 30$; $L_{\max} = 8$; 20-word list at 2 s/item.)

the amount of interitem rehearsal. (a) Increasing interitem processing by instructional set results in increased recall performance but little change in recognition performance (e.g., Estes & DaPolito, 1967). (b) Changing from incidental to intentional instructions increases recall performance but may slightly lower recognition performance (e.g., Eagle & Leiter, 1964). (c) Varying the test expectancy from recognition to recall generally increases both recall and recognition performance (e.g., Balota & Neely, 1980; Neely & Balota, 1981; Hall, Grossman, & Elwood, 1976).

We suggest that increased intentionality, or increased expectancy of recall, leads to increased interitem coding, possibly at the expense of decreased self-coding, that is, increased b and possibly decreased c.[10] The second panel of Figures 10 and 11 show the predictions of the model when the b parameter takes on the values 0.10, 0.20, and 0.30, and the other parameters are unchanged. In this case, both recall and recognition performance are predicted to rise as a result of increased interitem coding.

Figure 15 shows predictions of the model when the b parameter increases from 0.10 to 0.15 to 0.20, but in addition, the c parameter decreases: $c = 0.51$, 0.33, and 0.15, from left to right. In this case recall increases but recognition remains relatively constant because subjects are trading interitem coding for self-coding.

Why does the model predict the patterns seen in Figure 15? For recall, predictions are clearcut: Both increases in the b parameter and decreases in the c parameter improve recall. Increases in b improve both sampling and recovery and decreases in c reduce self-sampling. For recognition, the situation is more complicated: Increases in b improve recognition performance (due to increased associations to items rehearsed with the test item) but decreases in c harm recognition (due to lower associations of the image of the test item

[10] Neely and Balota (1981) came to the opposite conclusion, that interitem coding was *not* affected, because the test-expectancy effect (better performance for recall expectancy) did not interact with the semantic organization effect (better performance for related words). Their argument was predicated on an assumption derived from the theory of J. R. Anderson (1972) and of J. R. Anderson and Bower (1972) that interitem coding should not only be higher under recall expectancy but should be differentially higher for related words. We see no basis for such a prediction in our theory and, hence, we do not agree with the conclusion. Furthermore, we think our theory could handle the Neely and Balota (1981) results, although we have not yet attempted to fit the findings.

itself). In this case, the precise nature of the trade-off in rehearsal is crucial to the predicted outcome. Just such a sensitivity to trade-offs may explain the variability of observed results.

Match Between Encodings at Study and Test

There is a well-tested principle stating that memory improves to the degree that codings at storage and test match (e.g., the *encoding-specificity* principle; Tulving & Thomson, 1973). For example, a number of studies show that words recognized in one semantic context may not be recognized in another. Light and Carter-Sobell (1970) presented subjects with homographic nouns to study. Each was presented with an adjective to bias a specific meaning of the word (e.g., *traffic-jam*). Subjects were then tested for recognition of the homographic nouns in the presence of the same or a different biasing adjective (e.g., *strawberry-jam*). Recognition performance was higher in the same-adjective tests (see also Tulving & Thomson, 1971; Winograd & Conn, 1971).

The present model handles such coding effects through the match parameters, m, which determine the relationship between encoding of a cue at study and test. Changes in m are captured in our present simulation by changes in a, b, and c as appropriate.[11] This encoding variability approach makes it easy to predict a variety of semantic context effects. The reasoning is similar to that put forth by Tulving and Thomson (1973) and Reder, Anderson, and Bjork (1974), among others. By the same reasoning, when a word has one meaning (e.g., *rhinoceros*) effects of semantic context should be lowered (Reder et al., 1974; Muter, in press).

It may be asked, how does the biasing item or context affect the encoding of a study or test item? Because a variety of subject controlled coding and rehearsal processes occur during most study settings, the effects of semantic context are easy to incorporate. On the other hand, the initial encoding of a test item is generally thought to be a relatively automatic process (e.g., Shiffrin & Schneider, 1977). However, even though the initial encoding may be automatic, it is determined by the general stimulus environment, not just by the nominal word itself. Therefore, context determines encoding at both study and test, and the possible different encodings then produce the cited effects of semantic context.

There is a variety of ways to demonstrate encoding match effects in our present simulation. Perhaps simplest is the assumption that the match of encoding at test to that at storage is lowered, thereby lowering the values of b and c. The resulting predictions (for recognition only) are given in Figure 16. Decreases in the b and c parameters each cause performance to drop as shown.

Although changes in meaning are undoubtedly responsible for many semantic context effects, another factor may also contribute. When the test consists of both a critical word and a word used to produce a bias, both words may be used in the probe set, along with context. As we discuss in Section 4 the biasing word affects the resulting familiarity value. These changes in familiarity may produce semantic context effects similar to those observed, even if the coding of the critical word is assumed to be the same in all contexts. For example, if biasing items at study and test are different, recognition of the critical item is harmed: An identical bias item is more strongly linked to both the test item (due to rehearsal) and itself (due to identity) than is a shifted bias item, so the familiarity of the pair of items is higher in the identical case. If the bias item is present only at test, it produces more familiarity when it matches the dominant meaning, because the dominant meaning was more likely coded at study and the residual association to the image of the test item is higher in such a case. These examples make it clear that the SAM model has two complementary mechanisms that can help explain context effects in recognition, one based on the coding of the test item and the other based on the familiarity of a test pair.

Related results concerning encoding matches have been found in a slightly different paradigm. The paradigm consists of three basic phases: (a) the study of A–B pairs of words, of which the B members are the critical words; (b) a recognition test for the B members. The recognition test is given in the absence of the

[11] An alternative treatment consistent with the SAM model would involve treating each new and different encoding of a nominally identical item (whether presented at study or at test) as a new item with appropriate interitem strengths. This approach would probably be difficult to distinguish from the approach suggested in the text.

A members and, in addition, is usually given in a situation where a different meaning or sense of the word is indicated by the surrounding test items; and (c) a cued-recall test in which the A members are presented as cues for recall of the B members. The standard result is that there are a number of words that are recalled but not recognized. Tulving and Thomson (1973) were the first researchers to utilize the procedure, and they termed the result "recognition failure of recallable words" (p. 39). Since that study, the results and the implications of the results have been a topic of much debate and experimentation (e.g., Antognini, 1975; Rabinowitz, Mandler, & Barsalou, 1977; Reder et al., 1974; Salzberg, 1976; Tulving & Watkins, 1975; Watkins & Tulving, 1975; Wiseman & Tulving, 1975).

The finding that cues can sometimes lead to recall of items that cannot be recognized is quite consistent with the SAM model. The fact that recall operates as a stochastic search process guarantees that some searches succeed even though the familiarity value for the recalled items may not reach criterion. However, if recall occurs first, recall of an item would probably produce learning (an increment in context strength) that would raise the familiarity value for that item. Thus a later recognition test of that item would tend to be above threshold. Most recognition failures in this case (and many other cases, as well), then, are probably due to a biasing item or to other context changes that induce a meaning during the recognition test that is different from the meaning accessed during the recall test. For similar reasons, it is predicted that cued recall can fail when free recall succeeds, or vice versa. Such results are expected whenever differing test and study contexts induce different meanings of a word (and may even occur when the same meaning is involved due to the random nature of the search process). All such effects can be expected to decrease when the target words have only one meaning or encoding (e.g., Muter, in press).

Context Shifts and Test Delays

It is well known that interposing a delay or new items between presentation and test decreases both recall and recognition performance (e.g., Shepard, 1967; Strong, 1913; Underwood, 1957). In the SAM model two factors are mainly responsible for such decreases in performance. One factor is the storage in memory of new information that occurs during the delay. Indeed, the context cue used at test may well be more closely associated to the intervening material (items, stimuli, external

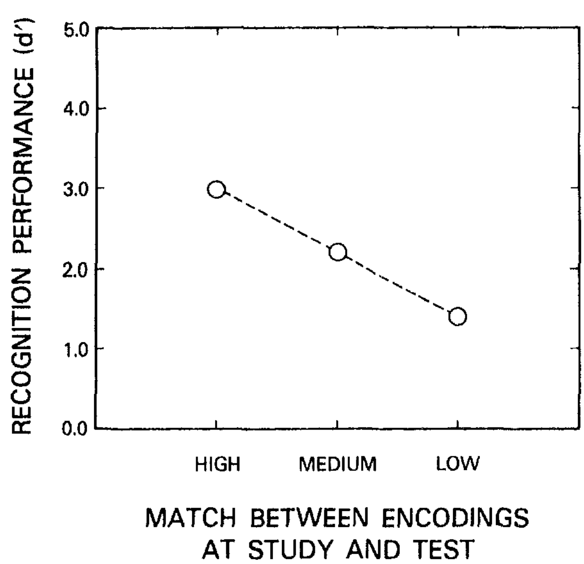

Figure 16. Predictions for recognition as the match between the studied items' encodings at study and test decreases. (From left to right, b = 0.20, 0.16, and 0.12; and c = 0.15, 0.12, and 0.09. r = 4, a = 0.25, d = 0.075, v = 0.5, K_{max} = 30, L_{max} = 8; 20-word list at 2 s/item.)

events, thoughts, etc.) than to the list items themselves, because of the closer temporal proximity of the test setting to the intervening material than to the list. On the other hand, the context cue surely includes some sort of list cue that tends to access only the list items, thereby offsetting the factor of temporal proximity. Regardless of the balance of effects, there is some tendency at test to sample information stored during the delay, and the tendency increases with delay duration. The effect of this factor is much the same as that of increased list length—performance drops for both recall and recognition for the reasons discussed in connection with the list-length effect.

The second factor contributing to the delay effect is the change in context from storage to test (see Bower, 1972). Normally the test context is increasingly changed from the storage context, as delay increases. Remember that the a parameter combines both a storage parameter and a degree of cue matching; thus it represents the degree of context similarity between study and test. As a result, the effect of context shift can be captured simply by lowering the value of a. Lowering a lowers recall directly, because recovery probabilities are less. However, lowering a has no effect on recognition, because a acts as a scale factor.

In an experiment varying delay, these two factors operate jointly, and respective effects are difficult to disentangle. However, some studies have altered context while holding delay constant. For example, Smith et al., (1978) presented words for study in a particular room (context). Subjects were tested for the words by recall or recognition in either the same context or in another room that was very different in setting than the first. The results showed decrements in recall with context change but no change in recognition performance (see also Godden & Baddeley, 1975, 1980). Similar findings have sometimes been been reported in the state-dependent learning literature (see Eich, 1980; but see also Birnbaum & Parker, 1977). In a typical state-dependent learning task, subjects learn some material under the influence of a drug (e.g., caffeine, alcohol) and then are tested later either under the same or different drug state. It is sometimes found that performance decreases with state changes when measured by recall but remains relatively constant when measured by recognition tests.

We model the context-shift paradigm (with fixed delay) by letting the a parameter decrease with context shift. The results for no delay, given in Figure 17, show the expected pattern—recall drops but recognition is unaffected.

We model the delay situation by retaining the assumption that the a parameter drops with delay but by adding the assumption that new information is added to memory. In particular, the retrieval structure is increased in size through the addition of several columns in the matrix that are meant to represent new information that has been stored. We refer to this new information as *junk*. Item-to-junk strengths are assigned a value of j_i and context-to-junk strengths are assigned a value j_c. Increases in delay are modeled by increases in the number of such junk items (i.e., columns in the matrix).

Figure 18 shows predictions derived by letting $a = 0.25$, 0.20, and 0.15 and letting the number of junk columns be 0, 10, or 20, from left to right, respectively. The junk parameters were $j_c = 0.1125$ and $j_i = 0.375$. Both the decrease in the a parameter and the increase in the number of junk columns cause recall to decrease. Only the increase in the number of junk columns causes recognition to decrease.

Each of the patterns in Figures 17 and 18 is of course consistent with the results in the literature. The notion of junk information is actually worth a deeper analysis. In all applications of SAM thus far, it has been assumed that the context cue is sufficiently precise to limit search to the list just presented (or at least it has been assumed that the contributions of nonlist information is negligibly small). For recall, such assumptions do not seem to be crucial, failing mainly because intrusions of nonlist items are not predicted. We have examined the standard recall predictions (in previous papers as well as in this article) with the added assumption that several junk columns exist in the retrieval matrix (possibly representing events that occurred prior to the list). Recall is always lowered by this alteration, but the patterns of recall are unchanged. In fact, an adjustment of parameter values can be made to reinstate the original predictions with little change.

For recognition, the presence of junk col-

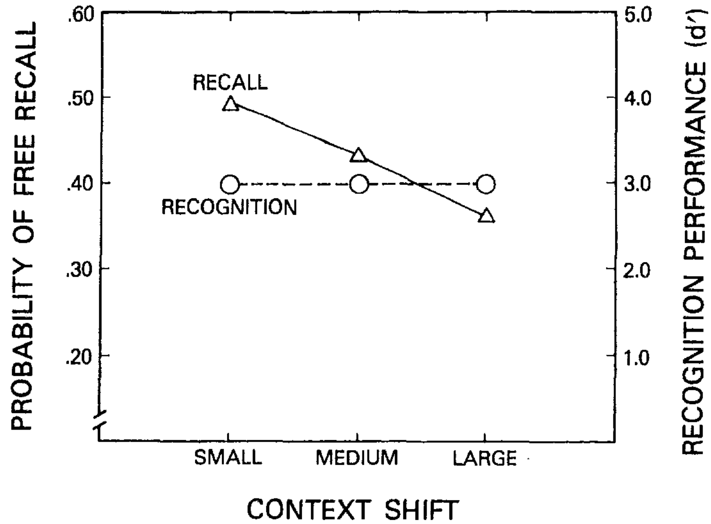

Figure 17. Predictions for context shifts between presentation and test, with delay held small and constant. (From left to right, a = 0.25, 0.20, and 0.15. r = 4, b = 0.2, c = 0.15, d = 0.075, v = 0.5; $K_{\max}$ = 30, $L_{\max}$ = 8; 20-item list at 2 s/item.)

umns (whether due to information occurring during a study-test delay or to information preceding the list) can make an important difference. For example, the lack of effect of context shift shown in Figure 17 (in the absence of substantial delay) no longer obtains if junk columns are assumed; this is shown in Figure 19. Figure 19 indicates that the absence of an effect (due to context shift) occurs only when no junk is present; as the junk increases, the effect appears and gets larger. By this logic, it can be weakly argued in the context of the

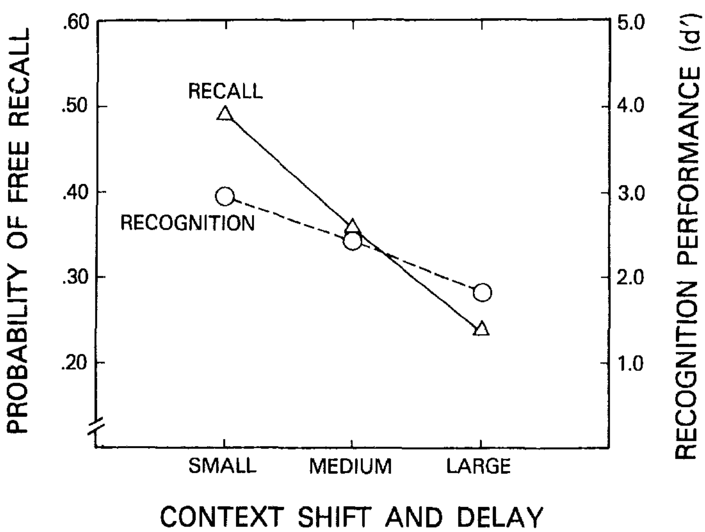

Figure 18. Predictions for increasing amounts of delay in conjunction with increasing amounts of context shift. (From left to right, a = 0.25, 0.2, and 0.15; and number of junk items = 0, 10, and 20. r = 4, b = 0.2, c = 0.15, d = 0.075, v = 0.5; j_c = 0.1125, j_i = 0.375; 20-word list at 2 s/item.)

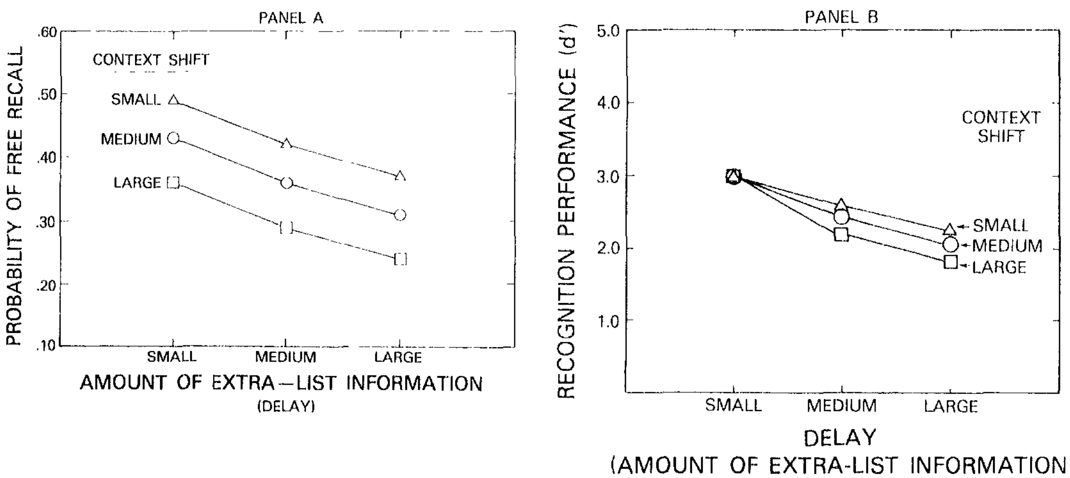

Figure 19. The model's predictions for different amounts of context shift as a function of the amount of extra-list information (delay) present at test. Panel A shows the results for recall. Panel B shows the results for recognition. (All parameters in each case are identical to Figure 18, except number of junk items = 0, 10, and 20 from left to right; and a = 0.25, 0.20, and 0.15 from top to bottom.)

present model that only a small amount of junk exists in certain types of list experiments, because the observed effect of context shift on recognition is so small (where *small amount* refers either to few junk items or to weak junk strengths).

Test Position Effects

In previous recall applications we have assumed that learning can take place during the testing procedure. Specifically, we have assumed that any time a probe set successfully recovers an item from memory, the strength of association from the cues in the probe set to the recovered item is given an increment. The result of strengthening is to reduce the probability of recall of a new item as the number of items previously recalled increases. This occurs because self-sampling is increased: If any of the cues that have been given an increment are used again, they are more to likely sample an old (previously recovered) item, and thus fewer new items are recalled (see Raaijmakers & Shiffrin, 1981a, 1981b, for a more detailed discussion of learning during recall).

The justification for learning during recall is straightforward: Learning occurs when items are together in STS, and is especially strong if those items are attended, coded, or rehearsed. If an image is recovered in the presence of a set of cues, then that image and those cues are together in STS, probably under attended conditions. The situation in recognition is quite different. In this case we have assumed that a specific image is not recovered; instead, access is gained to the accumulated activation of all images. It might be possible to imagine that the cue strengths to all images are slightly increased, but we think that this is an implausible assumption and is likely to have negligible effects on performance in any event.

Instead, we propose that learning in recognition testing takes a slightly different form—each item tested produces a new image in the retrieval structure with some new self-strength and some new context strength. (It is possible that these strengths could differ for items given positive and negative responses, because they might receive differential attention; see Rabinowitz, Mandler, & Patterson, 1977. However, as an initial simplification the strength for positives and negatives is assumed equal.) The effect of this assumption is quite simple: Recognition testing produces an increase in the effective list length that applies for subsequent test items. The effect is much the same as that seen when list length increases (Figure 12), or when junk items are added to a list due to delay of test (Figure 18). Performance is predicted to decrease across test positions.

Simulation predictions are shown in Figure 20. Self-strength is not relevant when items are not retested. Therefore only one parameter is needed: The context strength for the new images of tested items was set to 0.5. The changes in d' across quadrants of the test list (all 20 targets plus 20 distractors, mixed) is calculated by comparing hit rates and false alarm rates for items in similar test positions, with the results shown in the Figure 20.

Do such predictions correspond to the observed data? It is difficult to answer this question on the basis of previous research. Many studies report decreasing performance with test position (usually confounded with lag effects for targets), but the situations are ones in which short-term memory contributions to performance are not controlled or eliminated (e.g., Murdock & Anderson, 1975; Ratcliff & Murdock, 1976). It seems sensible that an item still in STS when tested will give rise to an extremely rapid and accurate *yes* response. Because the probability of this event decreases with test position, exaggerated position effects are to be expected. (Furthermore, the common practice of restricting consideration to high confidence responses only accentuates this problem.) We obtained test position effects from an experiment in which intervening arithmetic was used to clear STS before testing.

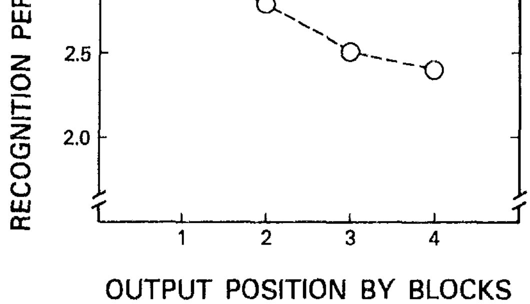

Figure 20. Predictions for recognition performance in a 20-item list (40-item test list) as a function of test position (blocked in groups of 10 items) when each test causes storage of a new image with total context strength (increment) = 0.5, and residual interitem strengths (d) of 0.075. (v = 0.5, r = 4, a = 0.25, b = 0.20, c = 0.15, d = 0.075; 2 s/item.)

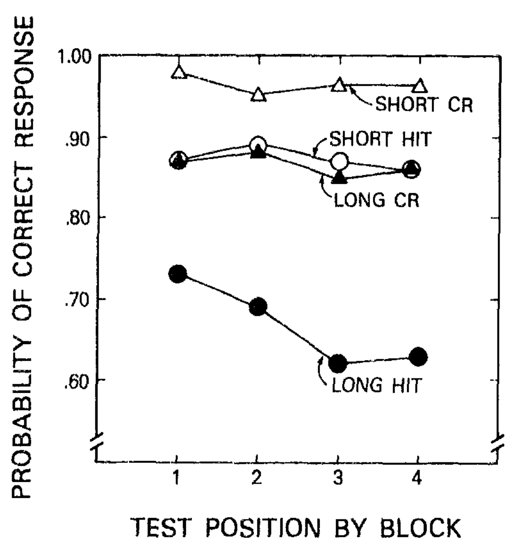

Figure 21. Probability of correct rejections and hits for 24-item (short) lists and 240-item (long) lists, averaged across quarters of the 48- or 480-item test lists. (Arithmetic intervened between presentation and test.)

The probability of hits, and of correct rejections, averaged over quarters of the test list, is shown in Figure 21, for both long (240 items) and short (24 items) lists.

Obviously the effects are not very large, appearing only in the case of hits for the long list, and showing only about a 9% falloff in that case. (Other partitions of test positions gave rise to similar results.) It seems likely that an effect of this magnitude could be handled by assuming that a small amount of learning takes place during testing, primarily for positive responses.[12]

Distractor Type and Similarity

It is well known that the choice of distractor affects recognition performance (e.g., Anisfeld

[12] Todres and Watkins (1981) found a small decrement in recognition performance for categorized items when subjects were given additional exposure to some of the presented or nonpresented items before test (as compared to subjects not receiving such experience). Neely, Schmidt, and Roediger (1983) showed that six successive test items from a category slowed reaction time to another test item from that category (compared to two successive test items before the critical item). In both of these cases, learning during retrieval or during an intervening task could account for the findings, although a quantitative demonstration of this claim would require a simulation of the particular procedures in these studies.

& Knapp, 1968; Hall, 1979; Underwood & Freund, 1968; Walter, 1977). The point seems clear: If a distractor is highly confusable with the list items, discriminations are difficult. An interesting question concerns the nature of the relationship necessary to produce such an effect: Must the distractor be similar to many (or all) of the list items, or may the distractor be similar to just one list item? Our model suggests that a very large decrement occurs when the distractor is similar to a large proportion of the list items, because familiarity is a sum of familiarity values over all of the list items. For example, if a study list consisted of flower names and the test list consisted of all flower names (targets mixed with distractors), surely performance would be lower than in the case where all of the distractors were animal names. Evidence that this actually occurs is difficult to find, perhaps because experimenters believe the outcome is so obvious that experiments have not been done. There is evidence from reaction time studies that noncategory distractors are rejected faster than category distractors (distractors from the same category as the presented items; Herrmann, McLaughlin, & Nelson, 1975; Homa, 1973; Okada & Burrows, 1973).

Of greater interest for some purposes is the effect of familiarity of a distractor to a single, matched list item (see Anisfeld & Knapp 1968; Underwood & Freund, 1968). Our model predicts an effect of distractor similarity on the following basis: The residual interitem strength between the similar distractor and the image of the matched list item will be higher than the usual residual value. If this value is called d_s, then we suggest that d_s is greater than d to the degree that the similarity between the items is high. Let us suppose, moreover, that the similarity of a distractor to a list item does not cause that distractor to share in the associative linkages that had been formed during study between the similar list item and other jointly rehearsed list items. That is, d_s for the distractor would apply only to the single, similar list item. The predictions under these assumptions are depicted in Figure 22.

The lower curve in Figure 22 shows a sharp performance drop when similarity of the distractors to all of the list items is increased. In this case the value of d is increased from left to right (d = 0.075, 0.1, and 0.125), but otherwise the parameters remain the same in the previous model applications. The upper curve shows a much smaller decrease when distractor similarity to a single list item is increased. In this case, all parameters remain the same as for previous applications, but for just a single list item, the value of d_s is increased from left to right (d_s = 0.75, 0.1, and 0.125). For these predictions, d' is calculated from hit rates for the list item with the similar distractor, and false alarm rates for the similar distractor.

Data concerning similarity effects are somewhat mixed. Some studies report that similarity to a matched list item reduces performance (e.g., Anisfeld & Knapp, 1968; Underwood & Freund, 1968), but others do not (at least not for some related distractors or under some encoding instructions—Buschke & Lenon, 1969; Cramer & Eagle, 1972; Elias & Perfetti, 1973; Underwood & Freund, 1968). Furthermore, it is often the case that physical similarity (phonemic or graphemic) produces more false alarms (and lower performance) than does semantic similarity (Cramer & Eagle, 1972; see also Davies & Cubbage, 1976; Eagle & Ortof, 1967; Juola et al., 1971). We have carried out several studies of our own on this subject, with mixed results. Although one study showed associative similarity to increase false alarms, the others did not. A typical result is shown in Figure 4: Synonyms did not lead to increased false alarms, although phonemic and graphemic similarity did. A similar result was obtained in another study for associative relatedness that was not based on synonyms.

These results are generally consistent with the model's predictions that distractor similarity to a single list item should produce small performance decrements. Nevertheless, it is not easy for the model to predict that semantic similarity produces smaller effects than physical similarity or that semantic similarity sometimes produces no increase in false alarms at all. A possible explanation for such findings is discussed in Section 5.

3. Word-Frequency Effects

The predictions in Section 2 make it clear that our model for recognition and recall has the potential to predict many of the observed findings in the literature that involve accuracy of performance. In the present section we ad-

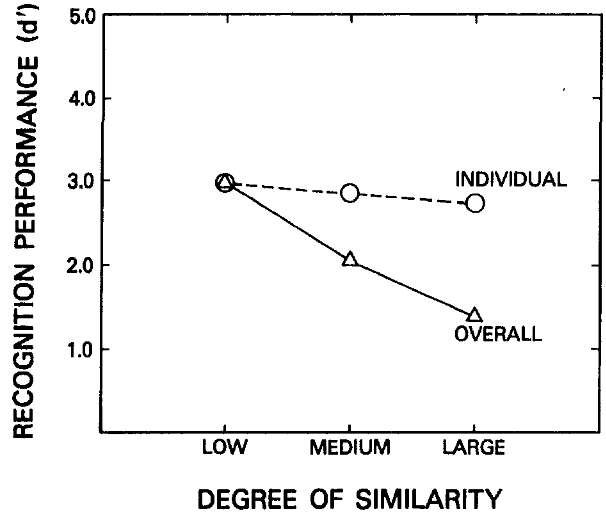

Figure 22. Predictions for recognition when the similarity between distractors and targets is varied. (The top curve shows the predictions for the case when a distractor is similar to only a single-list item [d_s = 0.075, 0.1, and 0.125 from left to right, all other d = 0.075]. The lower curve shows the predictions when distractors are related to all items. [d = 0.075, 0.1, and 0.125 from left to right; r = 4, a = 0.25, b = 0.2, c = 0.15, v = 0.5; 20-item list at 2 s/item].)

dress the issue of word-frequency effects in some detail. We present the main findings on word-frequency effects and show how our model predicts the observed patterns of results. We then present new experiments on the word-frequency effect and give a quantitative fit of the model to the findings. Finally, we discuss some additional findings on word-frequency effects and related topics.

The Word-Frequency Effect

Word frequency, or more precisely, natural-language word frequency refers to the relative frequency of occurrence of words in normal usage.[13] Words like *boy, chair,* and *blue* are words that occur with high frequency (HF) in the language whereas words like *tandem, ocelot,* and *pagoda,* occur with low frequency (LF). The interesting property of HF and of LF words for our present discussion is that they influence recall and recognition quite differently (see Gregg, 1976, for a review). The most commonly cited finding with regard to frequency effects is that subjects recall HF words better than LF words (e.g., Deese, 1960; Hall, 1954; Sumby, 1963) but recognize LF words better than HF words (e.g., Gorman, 1961; McCormack & Swenson, 1972; Schulman & Lovelace, 1970). This effect is found when pure lists are studied. A pure list consists of either all HF or all LF words. The advantage of LF words in recognition is largely unaffected by mixing HF and LF words in a single study list (Allen & Garton, 1968; Garton & Allen, 1968; Gregg, 1976; Schulman, 1967; Shepard, 1967). However, when mixing is used, the standard finding that HF words are better recalled no longer obtains: Although the results are somewhat variable, recall is close to equal for the two types of words. In most studies, LF words are recalled as well as, or even slightly better than, HF words (Duncan, 1974; Gregg, 1976; May & Tryk, 1970), but Balota and Neely (1980) found an advantage for HF words. Gregg, Montgomery, and Castano (1980) extended these recall findings by studying both pure and mixed lists (in different groups); to the standard conditions they added two conditions in which a distracting task was used during presentation to eliminate inter-item coding. The standard findings in the normal pure and mixed lists were replicated. With distraction, recall levels went down, and in

[13] We hope that there is no confusion between the word-frequency effects discussed here and the frequency-discrimination paradigm in which episodic frequency of presentation in a given list is manipulated.

pure lists the advantage of HF words was reduced.

When considering the word-frequency effect for recognition, we note that subjects show a frequency effect for distractors as well as for targets. The correct-rejection rate for LF distractors is higher than the correct-rejection rate for HF distractors (McCormack & Swenson, 1972). Even when distractors are not used, however, frequency effects are found. Wallace, Sawyer, and Robertson (1978) presented subjects with either HF or LF words and then gave a recognition test that contained no distractors (only the target items were tested). Subjects were informed of this fact and were directed to respond positively to words that they actually remembered seeing on the list. Subjects showed better recognition for LF words than for HF words.

Predictions for the Word-Frequency Effect

The model we use to deal with frequency effects is identical with that presented in Sections 1 and 2, with one slight addition to take word frequency into account. We assume that the tendency for HF items to access images in memory is larger than that for LF items. That is, the retrieval strength between a HF cue and any image is higher than that between a LF cue and any image. This assumption is instantiated in two ways: First, the residual association between any two items not rehearsed together, or between any distractor and any list item, is made frequency dependent. Thus when a HF word is a cue, the value of the d parameter is set equal to d_H; when a LF item is a cue, $d = d_L$; $d_H > d_L$. (The frequency of the image in memory is not taken into account in this approach.) Similarly, for words that are rehearsed together, it is assumed that HF cues are stronger than LF cues. Thus we set $b = b_H$ for a HF word and $b = b_L$ for a LF word; $b_H > b_L$. In short, HF words are assumed to be stronger cues than are LF words.

In our approach to word frequency, we make the assumption that the a and c parameters are not frequency dependent. Thus the strengths to context and to a word's self-strength (tendency to sample a word's own image) are assumed equal for HF and LF words. These assumptions can be justified by the reasoning that context coding is based on the time for episodic coding and rehearsal, and self-coding is based on time for rehearsal of interitem features, whereas interitem and residual interitem strengths are based, at least in part, on extraexperimental semantic factors that are sensitive to word frequency.

We assume that two criteria are selected by the subject, one for HF words and one for LF words. This assumption is sensible for pure lists, but we assume it holds for mixed lists and mixed tests, as well. Such an assumption could be justified on the basis that subjects can ascertain the word frequency of the tested word and can use this knowledge to set a criterion.

The predictions of the simulation model are shown in Figures 23 and 24. The b_H and d_H parameters were set equal to the b and d parameters used in earlier simulations ($b_H = 0.2$; $d_H = 0.075$). For LF words we set b_L equal to 0.1 and d_L equal to 0.035. The remaining parameters, a, c, r, and v are assumed to be constant for both HF and LF words and were set equal to the same values used in earlier simulations ($a = 0.25$; $c = 0.15$; $r = 4$; $v = 0.5$). The HF criterion was set to 1.3 and the LF criterion to 0.66. Figure 23 shows the model's predictions for uniform frequency lists. Recognition performance is higher for LF words, whereas recall performance is higher for HF words. Figure 24 shows the results for mixed lists. It is assumed that the test list contains both HF and LF targets and HF and LF distractors. We follow the usual procedure in such cases by calculating d' from a comparison of HF hits to HF false alarms or from a comparison of LF hits to LF false alarms. Recognition performance shows the same pattern seen in the pure lists: LF words show better recognition than do HF words. Note that recall in mixed lists is predicted to be equal for LF and HF words.

The model predicts a pure-list HF advantage for recall for two reasons. First, a given word (of any frequency) has a higher probability of recovery if it is sampled by a HF word as a cue than if it is sampled by a LF word, due to the higher strength values. In a pure HF list, only HF words are used as cues, of necessity, so recall is superior to that in a pure LF list, where all cues are LF items, of necessity. Second, equal c values for HF and LF words cause more self-sampling for pure lists

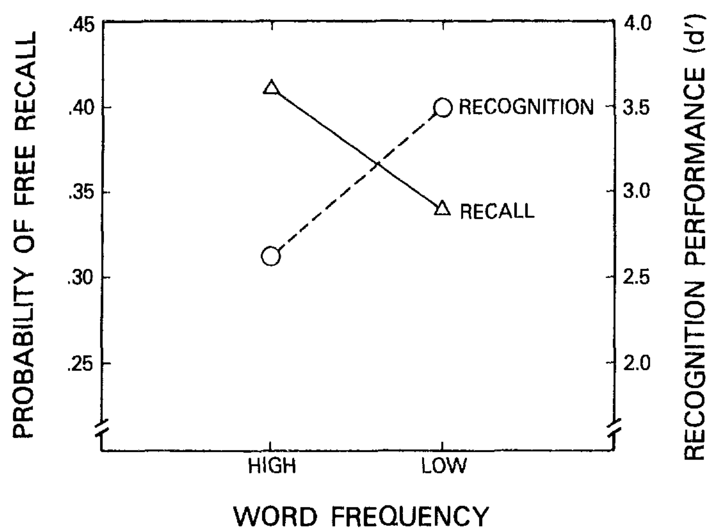

Figure 23. The model's predictions for recall and recognition as a function of word frequency for uniform frequency lists (30 words total). ($r = 4$, $a = 0.25$, $b_H = 0.2$, $b_L = 0.1$, $c = 0.15$, $d_H = 0.075$, $d_L = 0.035$, $v = 5$; $K_{max} = 30$, $L_{max} = 8$; 2 s/item.)

of LF words because the interitem strengths are lower for LF cues. Self-sampling reduces recall because sampling the cue itself, which has just been recalled, is of no use and reduces the opportunities for recovering a new word.

In mixed lists, however, a HF word is just as likely to sample a LF word as a HF word and then either type is equally likely to be recovered. Similarly, when a LF word is used as a cue, it is equally likely to sample either a HF or a LF word, and they have the same probability of recovery (although this is lower than the probability of recovery when a HF word is used as a cue). As a result HF and LF words are sampled and recovered equally often and, hence, used as cues equally often, pro-

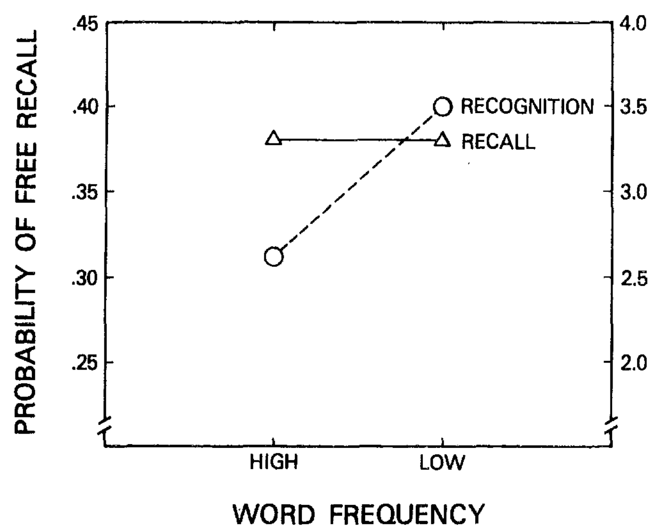

Figure 24. The model's predictions for recall and recognition as a function of word frequency, for mixed-frequency lists (15 high frequency and 15 low-frequency words). ($r = 4$, $a = 0.25$, $b_H = 0.2$, $b_L = 0.1$, $c = 0.15$, $d_H = 0.075$, $d_L = 0.035$, $v = 0.5$; $K_{max} = 30$, $L_{max} = 8$; 2 s/item).

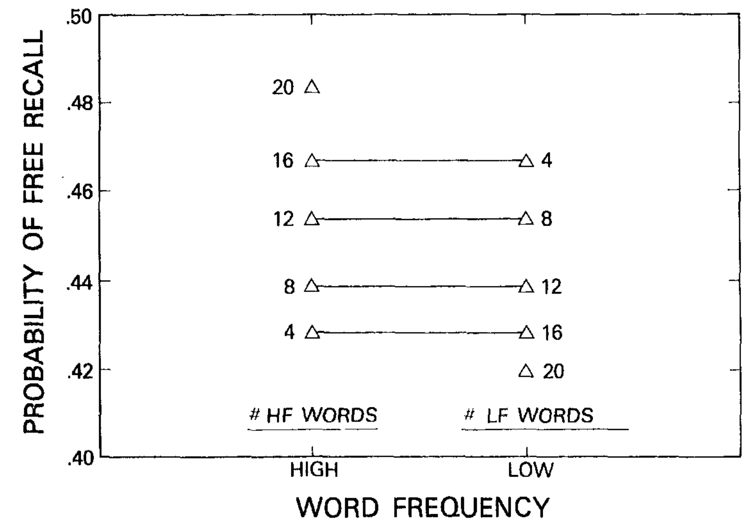

Figure 25. Predictions for free recall of a 20-item list as the composition of the list changes from being one of all high-frequency words to one of all low-frequency words. (Parameters are equal to those in Figures 23 and 24).

ducing equal recall. (Due to self-sampling, more failures take place whenever LF words are cues, but this does not cause any differential recall of HF and LF items).

These somewhat mysterious recall predictions may be clarified by considering how recall probability changes with the proportion of HF words in a list. Figure 25 gives the relevant predictions, as the number of HF items in a 20-item list is varied from 0 to 20. In any list, recall probability is equal for HF and LF words, but the proportion of HF words determines the level of recall for both types.

The model predicts recognition superiority for LF words because the self-strength parameter, c, is equal for HF and LF items. Because the residual interitem parameters are lower for LF items, the increase from distractor to target distributions is larger for LF items, relative to the corresponding variances of the distributions. (Higher values of b_H than b_L actually help the HF items more than the LF items, but for most reasonable combinations of parameter values, the net effect of the b parameters is outweighed by the other factors that cause LF superiority.)[14] Note that the predictions are identical for mixed lists and for pure lists, because the distribution shapes and placements are determined by the frequency of the test item rather than the frequency composition of the list.

An Experiment on Word Frequency and Test Type

Most of the experiments we have discussed on word frequency have examined recall or recognition, but not both (although see, e.g., Balota & Neely, 1980; Neely & Balota, 1981). It is thus possible that some of the differences that exist between recall performance and recognition performance are due to encoding differences for recall and recognition. It may be that subjects who study for recall tests tend to interrelate items and that subjects who study for recognition attend to structural features of the word as Tversky (1973) and Mandler (1980), among others, have suggested. To draw proper comparisons between recall and recognition, therefore, it seems important to hold back knowledge concerning whether a recall or a recognition test is given until after list presentation. This is especially important in cases where the data are to be fit by a quan-

[14] All of the predictions in this paper concerning word frequency would be obtained qualitatively if only the d parameter (and not b) varied with frequency. When we fit such a model quantitatively, we found that in order to predict a large enough pure list recall effect, so large a difference between d_H and d_L was needed that recognition predictions were seriously in error (and vice versa). However, we are reluctant to conclude that the difference between b_H and b_L is essential, without additional analysis and research.

titative model. In the present experiment subjects received several different types of tests (both recall and recognition) but were not informed prior to the test which type they would receive. Presumably, subjects studied all lists in the same fashion. This presumption has the useful property that all encoding parameters in our simulations may be held equal for recall and recognition tests. This provides a good test for our retrieval assumptions and for the model's ability to handle recall and recognition together.

In the present experiment, we wished to extend the word-frequency findings to different types of presentation modes and to different types of tests than are normally used in frequency experiments. Specifically, in each list, we presented subjects with 16 pairs of words to study. The pairs consisted of either HF–HF, HF–LF, LF–HF, or LF–LF word pairs. Subjects were told to associate the words of a pair together. We hoped that such a procedure and instructions would give us better control of the encoding operations used by the subjects. Each list was followed by an arithmetic task to clear short-term memory.

Five different types of memory tests were given: (a) a single-recognition test where one item of each pair was tested as a target along with an equal number of never-presented distractors, (b) a pair-recognition test in which targets consisted of both members of a presented pair (tested in the order they were presented) and in which distractors consisted of two items never presented, (c) a cued-recognition test where one member of each pair was presented as a cue (and so indicated) along with either its original member or a never-presented distractor, (d) a free-recall test, and (e) a cued-recall test where subjects were given one member of each pair as a cue and were asked to recall the corresponding members.

We hoped that such a rich array of test types carried out within subjects would provide a powerful-enough data base to allow us to fit our model quantitatively and to provide a fair test of the model's ability to handle both recall and recognition. The details of the experimental procedures are given in Appendix B.

Results and Discussion

There was no difference in performance for items presented as the first member of a pair or for items presented as the second item of a pair. This is usually the case when neither word is designated as the cue during study and when, in cued tests, either member may serve as the cue (see Raaijmaker & Shiffrin, 1981a). Therefore, all analyses were completed by summing over position within pair.

The primary results can be seen in Figure 26. The probability of correct recall is plotted for the cued data and for the free-recall data. The hit rate minus the false alarm rate is plotted for the recognition tests. This measure was used because it takes into account the hit rate and the false alarm rate simultaneously, much like d'. However, d' is sensitive to the assumptions of normality and homogeneity of variances (and cannot be computed when either the hit rate is 1 or the false alarm rate is 0). We have analyzed the recognition data for hit rates minus false alarm rates, for hit rates and false alarm rates separately, and for d'. The patterns of the data are the same for all methods of analysis. In fact the model is fit to the hit rates and to the false alarm rates separately, as can be seen in the upcoming figures.

The main findings are as follows. Cued-recall performance is higher than free-recall performance, but in neither case is there a sizeable effect of frequency on performance. In the recognition results the most obvious findings show that performance is higher for LF than for HF words and that paired recognition is superior to single recognition, which is superior to cued recognition. The details of the statistical analyses are given in Appendix C.

In general, the results accord well with the existing data. Recall performance showed little effect of frequency, which is consistent with the finding that in mixed lists the advantage of HF words over LF words is greatly reduced or eliminated (Duncan, 1974; Gregg et al., 1980; May & Tryk, 1970). The findings of LF advantages in recognition are certainly consistent with the literature, as is discussed at the start of Section 3.

A Quantitative Model for Frequency Effects for Different Test Types

The model, with a few exceptions, is identical to the model outlined before in this section. However, some special assumptions are made for the different presentation lists and

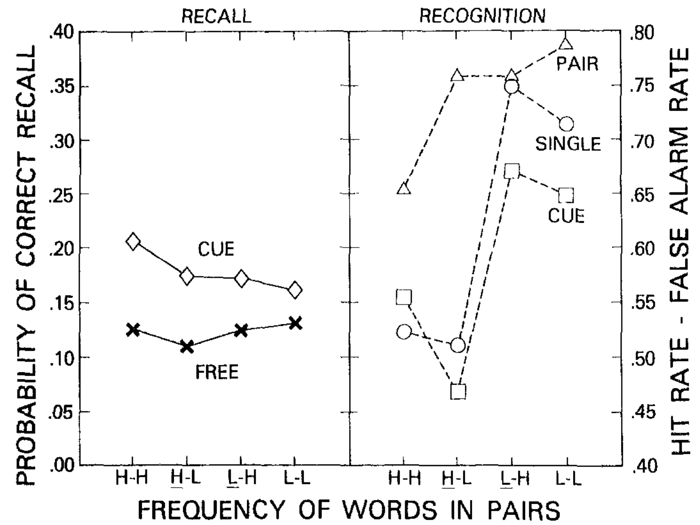

Figure 26. Data from the five test conditions of the experiment reported in Section 3. (The left panel gives the recall results and the right panel gives the recognition results, both as a function of word frequency. In mixed pairs, the underlined letter designates whether performance is being measured for the high-frequency (H) or the low-frequency (L) member.)

the different types of tests. Specifically, we assume that subjects only rehearse items that are presented as pairs at any one time—they never rehearse items from different pairs together. Thus the buffer-storage assumption is gone (as is the r parameter). We simply assume that each presentation of a pair produces a total context strength a, for each item in the pair and a total self-strength c, for each item in the pair. We assume that the strength of association of a HF word to its paired member (whether it be a HF word or a LF word) is b_H and from a LF word to its paired member is b_L. We assume that the residual associations between items that are not presented together as pairs follow the usual assumption, giving parameters d_H and d_L. Finally, we assume that the context and self-strengths are not only equal to each other but are equal for HF and LF words ($a = c$, equal for both HF and LF words).

Free recall is modeled in the same manner as was discussed earlier. The subject is assumed to begin search with context alone as a cue, and if an item is recovered then the item and context cues are combined and used until a new item is recovered or until L_{max} failures accrue. The entire process continues until K_{max} total failures are counted. K_{max} was set to 30 as it has been in all simulations and in many of the previous applications (Raaijmakers & Shiffrin, 1980; 1981b).

For cued recall the subject is assumed always to use both context and the cue given by the experimenter to probe memory. The subject is assumed to use the cue until M_{max} failures accumulate. There are no other stopping rules in the cued-recall model. (Once the paired member is recovered, it is assumed that the subject correctly realizes that that item had been paired with the cue. Furthermore, in our simple model, no provision for intrusions is made. Natural and very minor extensions of the model could be made that would deal with intrusions, but we prefer to keep the assumptions as simple as possible.)

The recognition models we propose are very similar to those used to generate predictions earlier in this section. However, for the paired- and cued-recognition conditions some assumptions must be made concerning how the two words available at test are combined to form a familiarity value. One might assume that the two words at test are independent and thus both could be used with context as a three-cue probe. Unfortunately, for any parameter values that fit the other conditions, such a model results in predicted paired-recognition performance that is far higher than that actually observed. The problem lies in the fact that two items that make up a pair have a strong interitem strength, b, as well as strong self-strengths, c, and context strengths, a. When the two items of the pair are used together as cues, the cross product terms between the two items increase familiarity disproportionately.

The solution we suggest to this problem involves fractional weights assigned to multiple cues. In the present case, we decided to preserve a special role for the context cue and assign it always a weight of 1.0. We then decided to limit the total weights, and attentional capacity, by letting all the remaining cues share a weight of 1.0. That is, the weights assigned to all cues other than context were made to sum to 1.0.

Under these assumptions, the modified recognition model is straightforward. Single-item recognition is handled just as in the simulation. Paired recognition is handled by assuming that the subject probes memory with three cues, context with a weight of 1.0, and the two items, each with weights of 0.5. Cued recognition is handled by assuming that the subject weights the test item very heavily and the cue item less heavily. In particular, the cue weight is represented by a parameter, W_c, and then the item weight is $1.0 - W_c$.

Note that the recall and single-recognition models need not be modified at all by these weighting assumptions. Because at most two cues are used in any probe during free or cued recall or during single recognition, the weights will still be 1.0 even under the new assumptions.

The model was fit to the hit rates and the false alarm rates in the various recognition conditions, and to the recall probabilities, by choosing parameter values that minimized the sum of squared deviations between the observed and predicted values. Actually, the time needed to produce predictions by Monte Carlo methods was too large to carry out a complete search of the parameter space, so certain simplifications were adopted. First, we arbitrarily set $v = 0.5$, to match the value used in the previous simulations. (Actually, the model seems fairly insensitive to the value of the v parameter over small ranges because fits about as good were obtained with $v = 0.25$ and $v = 0.75$, although other parameter values, of course, had to be adjusted accordingly.) The recall-stopping parameters were set to values estimated in previous research (Raaijmakers & Shiffrin, 1980, 1981): $K_{max} = 30$; $L_{max} = 3$. Also, following previous work, we set $a = c$. The remaining parameters were estimated, with the following results: $a = 0.16$, $b_H = 0.27$, $b_L = 0.19$, $d_H = 0.053$, $d_L = 0.0262$; $M_{max} = 10$; and $W_c = 0.1$. The recognition criteria were set individually for each condition and the values are given in the figure legends.

The number of parameters may seem excessive, even considering the different tasks and variety of conditions that are being predicted, but we do not feel that this is a proper assessment. First of all, the parameters used here are essentially the parameters already used by Raaijmakers and Shiffrin (1980) to predict a wide variety of recall data. A few additional parameters are needed to handle the different tasks and conditions, but the additions are quite natural: high and low values of the b and d parameters to take word frequency into account; M_{max} to replace K_{max} and L_{max} for the stopping rule for cued recall; and W_c to govern the weighting of the cue in cued recognition. Second, the values of the estimated parameters are close to the estimates obtained in previous fits to recall data and are about what one might expect given the new stimuli and tasks in the present study.

The predictions of the model for the parameter values indicated above are given along with the observed data in Figures 27, 28, 29, and 30. A few words of explanation are needed to explain the genesis of the predicted effects.

For free recall, for reasons similar to those discussed earlier, not much effect of frequency is predicted—this is typical of predictions for mixed lists. Actually, there are some slight predicted differences that result from trade-offs between sampling and recovery probabilities when one member of a pair has been recovered and is being used as a cue. In such a case, slight differences in recallability of the paired member occurs depending on the frequency of the cue.

For cued recall, the frequency differences are somewhat magnified because all samples from memory are made with the cue of specified frequency. Even here, however, the differences are small. Thus for mixed lists of pairs of items varying in frequency composition, only very small differences are predicted in

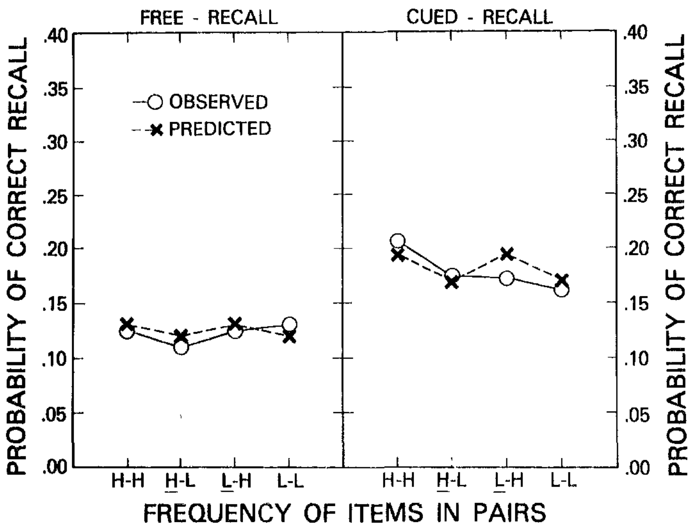

Figure 27. Predictions and data for cued and free recall from the experiment of Section 3. ($a = 0.16$, $b_H = 0.27$, $b_L = 0.19$, $c = 0.16$, $d_H = 0.053$, $d_L = 0.0262$, $v = 0.5$; $K_{max} = 30$, $L_{max} = 3$, $M_{max} = 10$.)

both free and cued recall. Such predictions conform quite well to the findings.

For single recognition, substantial effects of frequency are predicted. It is interesting that the frequency of the tested item, rather than the frequency of the paired member, determines performance. This is not surprising when one realizes that the tested item is the cue used to probe memory, so that its frequency is the determining factor according to the model. LF items give superior performance for essentially the reasons mentioned earlier in this section: The self-strength is higher relative to the residuals for LF items, because the c parameter does not change with frequency, but $d_H > d_L$ (this effect being much larger than an opposing factor; because $b_H > b_L$, a HF test item gains some strength due to its extra interitem strength to its paired member).

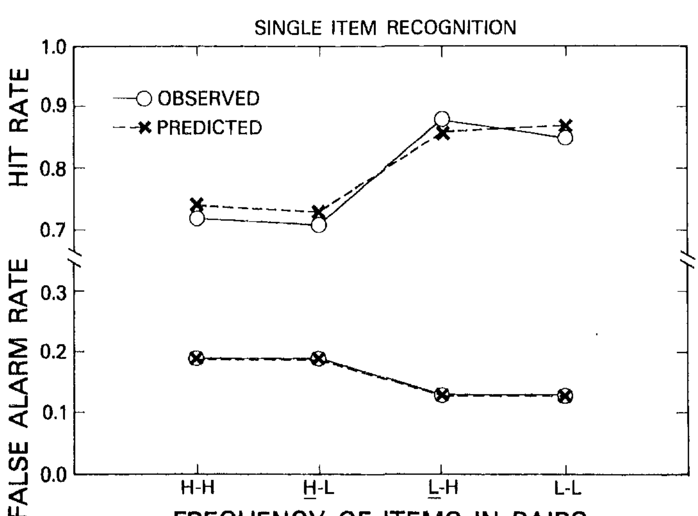

Figure 28. Predictions for the single-item recognition test. (Parameters are equal to those in Figure 27. The criterion for high-frequency (HF) items is 0.295 and for low-frequency (LF) items is 0.150.)

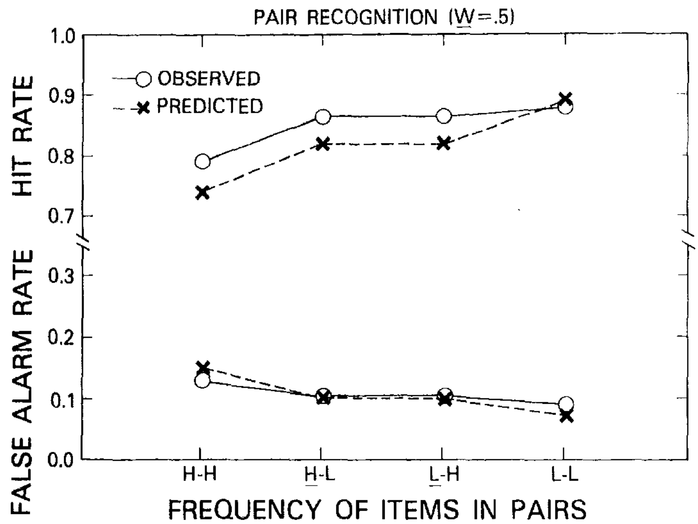

Figure 29. Predictions for the paired-recognition test using a shared-weight model for multiple cues. (The weights for the two test items are $W = 0.5$. Other parameters are equal to those in Figure 27. The criterion for HF–HF pairs is 0.285, for HF–LF or LF–HF pairs is 0.205, and for LF–LF pairs is 0.146. HF = high frequency; LF = low frequency.)

For pair recognition, the predicted effects are slightly different, because the strengths for each cue member, at their respective frequencies, are combined multiplicatively and then are summed. In effect, this produces an average of two different frequency effects in mixed-frequency pairs. Paired performance is predicted to be somewhat higher than single performance because the product rule for the two cues increases the signal-to-noise ratio, due to the interitem strength between members of a studied pair. In fact, the predicted improvement would be far larger were it not for the weights of 0.5 assigned to the two item cues.

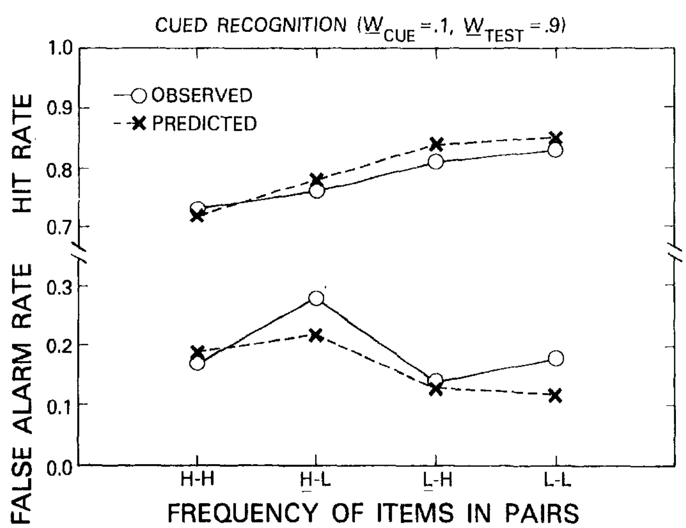

Figure 30. Predictions for the cued-recognition test, using a shared-weight model for multiple cues. (The weights are $W_{\text{cue}} = 0.1$, $W_{\text{test}} = 0.9$. The other parameters are equal to those in Figure 27. The criterion for the pairs as indicated from left to right are 0.292, 0.270, 0.160, and 0.150.)

It should be noted that quite different predictions would result if the distractor pairs contained either one or two items that had been presented on the list. Then performance would be lowered. This can be seen, in part, by examining the predictions for the cued-recognition case.

For cued recognition, the cued member is always from the list, whether a distractor or a target is tested. Therefore, its cue effect is common to both targets and distractors, so that discrimination of targets from distractors is not helped. In fact, the effect is to add noise to both target and distractor distributions, thereby reducing cued-recognition performance relative to the single-recognition case.

One might ask why subjects would assign any weight at all to the cue if this would lower performance. There are three plausible answers. First, the subjects are instructed that the cue item is present as a potential aid. Second, they may not realize that assigning any attention to the cue will harm performance. Third, the process by which the cue is assigned attention (and weight) may be automatic, occurring without conscious choice by the subject, simply as a byproduct of the subject's attempt to utilize the cue in some fashion.

Note that the predictions are not as close to the cued and paired data as to the single-recognition data. It may be that our method of weighting cues does not correspond exactly to the manner in which subjects assign their attention.

Discussion

Cue Weights

The paired-recognition condition provides the first evidence pointing to the need for sharing of cue weights. If weights are set equal to 1.0, then the use of context plus both test items as cues produces predictions of performance levels that are much too high. In particular, for parameter values that allow the single-recognition data to be predicted, paired-recognition predictions are essentially at ceiling. We are not arguing that the present results provide definitive empirical evidence in favor of shared cue weights. It is clear that models exist that can predict the paired- and cued-recognition data even if weights are all 1.0. For example, we have been able to show this to be the case for a model in which separate familiarity values are calculated for each of the test items and then are added together before a decision is made. This model and others like it, however, do not really resolve the basic problem: If cue weights are 1.0 then subjects can always improve their paired-recognition performance to ceiling by utilizing both test items as cues in the probe set. Because this option would always be available according to the logic of the SAM model, and because the expected results are not found, it seems reasonable to conclude that the model is wrong and that shared weights are used for multiple item cues.

Such reasoning is made even more plausible when generalized for multiple cues in recognition. Imagine a list of single items followed by a typical yes–no test. The model assumes that the subject probes memory with the test item and context as cues. Suppose, however, that the subject adds extra cues to the probe set, cues that are closely related to the test item. For example, if "table" is the test item, the subject might choose a probe set consisting of "context," "table," "food," "eating," "furniture," "chair," and so forth. Each of these related cues is only residually associated to list items other than "table," but is more strongly associated to "table" than to these other items. Thus the product rule for multiple cues insures that the discriminability increases with the number of these added cues. In effect, the familiarity assigned to "table" by the product rule, assuming "table" is on the list, becomes increasingly independent of the other items on the list, and performance continues to increase to ceiling, as the number of related cues increases.

A general solution to problems of this sort involves limits on cue usage that are produced by appropriate choices of weights. One possibility is a limit on the number of cues that can be utilized (perhaps due to a capacity limit on short-term memory). In this case, when the limit is exceeded, all cue weights for additional cues would be zero. A second possibility is to assign cue weights to new cues in accord with the new information they provide that is not already encapsulated in previous cues. In this case, the weights assigned to new cues continues to decrease as related cues are added. A final possibility is an attention limit on cue utilization. In this approach, all cue weights sum to no more than some constant (1.0 in a pure attention sharing approach). If the cue weights sum to 1.0, then adding cues generally does not help and will hurt if the added cues are less strongly associated to the target image than the target name itself, which is usually the case. In fact, in a model of this type, it is best to use just the cues that are most strongly associated to the target and least strongly linked to other list items. Usually the best cue is the test item itself.

This line of reasoning and the results of the present experiment convince us that some appropriate cue-weighting scheme is needed. Note that none of the qualitative predictions shown in Section 2 are altered by the use of fractional cue weights. Thus, although the use of fractional weights seems highly desirable, it was not necessary to adopt such an approach until the experimental results of the present section were encountered. Finally, although it seems clear that a shared-weight assumption is needed to handle the experimental results, it is probably the case that a variety of sharing rules could handle the present findings quite well, so that additional research is needed before firm conclusions can be reached.

Frequency-Dependent Cue Strengths

The SAM model explicates many results concerning word frequency by positing higher cue strengths to other LTS images for HF words than for LF words; in other words, the b and d parameters are higher for HF words than for LF words, although the a and c parameters do not vary with word frequency. This approach to word-frequency effects is far from novel; similar approaches, albeit stated in somewhat different terms, have been suggested or mentioned by Deese (1960), Gregg (1976), Sumby (1963), and Earhard (1982), to name just a few.

The Earhard (1982) results are particularly interesting. He set out to test the notion that HF words have more meanings available at study and test, so the chances of identical encoding at these two times would be lower; without identical encoding, performance would be worse. Although this hypothesis seems compelling, in the sense that different encodings would surely reduce performance, and in the sense that HF words surely have more meanings, the degree to which different encodings actually occur may be fairly small, unless specific attempts are made to change the context and bias different meanings (see Section 2). In fact, when Earhard insured identical encodings at study and test, the word-frequency effect was not reduced in size, which suggests that the encodings at study and test are normally similar.

In a final experiment, Earhard tested the hypothesis that both HF and LF items cause implicit rehearsal of (largely) HF items during study. If so, false alarms to these rehearsed HF items would tend to occur during test, lowering performance. Indeed, a study phase in which subjects were forced to generate three free-associates to each study item did produce a sizeable word-frequency effect, and errors on the forced-choice test often consisted of choices of previously generated words (mostly HF words). Unfortunately, the word-frequency effect was, if anything, even larger for the control condition in which the study word was just repeated three times. Although subjects may have implicitly generated even more associates in the repeat condition than they were instructed to produce overtly in the experimental condition, it seems to us more likely that a different but related explanation is called for.

In our model, it is not necessary that an implicit or an explicit HF associate be generated at the time of study. After all, the result of such a generation in Earhard's (1982) model is to make the generated item more familiar at test. In our model, such items would be more familiar when tested anyway, because they have higher residual strengths to the study items. These higher residuals not only produce the extra familiarity of the generated items (as needed in Earhard's model), but also the higher familiarity accruing to HF distractors when these are not rehearsed at study time (thus explaining why the repetition condition also produces a large word-frequency effect).

Although our explanation of the word-frequency effect is quite powerful, there is undoubtedly more to the story, considering the many ways that HF and LF items differ (other than frequency). Furthermore, subjects are generally capable of assessing the frequency of presented items and may vary their rehearsal and coding operations accordingly (see May & Tryk, 1970).

Directional Retrieval Strengths

Previous articles on the SAM model allowed for the possibility of asymmetrical retrieval strengths between two items (one a cue, one an image), but in practice always adopted a symmetry assumption. It certainly seems plausible that asymmetrical strengths should exist. Examples like *air-port* and *port-air* provide illustrations. Within the frequency domain, illustrations are not so obvious and sometimes seem to go in the direction opposite to our frequency assumption: for example, *stirrup-horse* and *horse-stirrup.*

This example and others like it are somewhat misleading for the following reasons. First, one's intuitions about strengths are probably based on free association results—undoubtedly *horse* is a more likely response to *stirrup* than the reverse. However there are probably only a limited number of such instances. In general, there are at most a few high likelihood HF responses to a given LF cue, and these responses are unlikely to have been included on the presentation list or test list. In typical studies, virtually all the items on the list are unrelated to a given LF item, in which case it seems quite plausible that the residual strengths should have the relationship $b_H > b_L$.

The second reason why a higher free-association probability for a HF response to a LF cue should not cause problems for our model involves the absolute levels of strength. If one extends the SAM model to the semantic domain and substitutes a *semantic* cue for a *context* cue, the response probabilities do not directly reflect absolute levels of strength. They reflect relative strength values as well, because these determine sampling. Thus *stirrup* may cause sampling of *horse,* not because that connection is strong but because the other competing associations (like *stirrup-table*) are so weak. *Horse* could have a stronger connection to *stirrup* than the reverse, yet produce sampling of *stirrup* less often because *horse* has strong connections to so many other words as well.

In the SAM model for episodic recall, Equations 7, 8, and 9 give the recovery probabilities that follow sampling; these do depend on the absolute levels of strength. If applied directly to the semantic situation, these equations predict the counterintuitive result that a LF response to a HF cue, once sampled, would be recovered with a higher probability than the reverse. It is possible that these equations do not apply in their present form to the semantic recall case or in any situation involving asymmetrical retrieval strengths. Perhaps the retrieval strength from the response to the cue should determine recovery in such cases, or perhaps recovery should depend on other factors not yet incorporated in Equations 7, 8, and 9.

Very Low Frequency Words and Nonwords

Pronounceable nonwords may be equivalent to very low frequency (VLF) words, words that are probably unknown to the subjects and therefore are functional nonwords. These items are recalled worse than HF or LF words. The situation in recognition is more complex: Performance is worse for VLF words than for LF words and may be as poor or poorer than HF words (e.g., Mandler, Goodman, & Wilkes-Gibbs, 1982; Rao, 1983; Schulman, 1976; Zechmeister, Curt, & Sebastian, 1978). The difficulty of forming interitem associations between VLF words is certainly one reason for their low recallability (e.g., it might not be easy to associate *abbatial* and *fluviatile* in a meaningful sentence or image). The recovery process might also be exceptionally poor when the image sampled does not incorporate a unitized lexical entity.

How could our model deal with recognition of VLF words or of nonwords? It is probably premature to speculate considering the paucity of relevant data. On the surface, it seems likely that for items without semantic or lexical representations, orthographic and phonological codes and similarities would take on primary importance. On the other hand, it might not be difficult to encode many VLF items in terms of a word that is suggested by the VLF item (probably a HF word). This would lend to semantic encoding in terms of the suggested word. Thus it is not difficult to come up with patterns of relationships among the parameters b, c, and d that would give rise to intermediate recognition performance, but it is difficult to select a hypothesis with any confidence, without further data.

Perhaps the simplest hypothesis holds that the *b* and *d* parameters are low for VLF items, but so is the *c* parameter. This pattern of parameter values would be especially likely if subjects were told to code for meaning (e.g., Mandler et al., 1982), because a self-coding for meaning would not be easy for VLF words. However, if subjects were to attend to orthographic features of the stimuli, then a higher *c* value might result, and recognition performance for VLF items would improve (e.g., Rao, 1983).

Forced-Choice Recognition

Several interesting studies concerning word frequency have utilized forced-choice rather than yes–no testing. Before considering these studies, the SAM model must be extended to forced-choice settings. The simplest model for forced-choice performance would have the subject compute separate familiarity values for the two (or more) alternatives and choose the item with the highest value. However, the items being compared might consist of classes that differ widely in familiarity values. One way to handle cases of this kind involves selecting a criterion for each item type, forming an estimate of the standard deviation of the distractor distribution for each item type, and constructing a standard score for each item of that type. The standard score would be the difference from the criterion divided by the standard deviation estimate; the largest standard score could mark the subject's choice. (Such a scheme could also be used to generate confidence ratings; these could be determined by cutoffs on the standard score dimension. Also, it is possible that some automatic process carries out the standardization by, say, shrinking or expanding the different familiarity scores to a common scale before a criterion is chosen.)

Interactions of the Word-Frequency Effect With Distractor Similarity

An interesting word-frequency study using forced choices has been carried out by Hall (1979). He was interested in demonstrating that recognition performance is crucially dependent on choice of distractor, a point about which we are in full agreement. He used two groups of subjects. One group replicated the usual word-frequency effect: After a list of mixed HF and LF items, forced-choice tests were given comparing HF targets with HF distractors chosen from the same populations, and LF targets with LF distractors. As usual, recognition was superior for the LF items. A second group received mixed lists of HF and LF items but were given forced-choice tests with distractors chosen for orthographic similarity. The HF targets were paired with HF distractors that differed in one letter position. The LF targets were paired with nonwords that differed in one letter position (the LF words were low enough in frequency that they may well have been functional nonwords—in any event lexicality could not have been of much help in making a decision in light of the results). In this group, HF tests were superior to LF tests; in addition HF performance was higher than in the control group, and LF performance was worse. The results were replicated in a subsequent study.

We offer the following somewhat speculative hypothesis to explain these data. Suppose that a forced-choice test with an orthographically matched distractor leads the subject to encode the test items in a way that deemphasizes orthographic features. This might be done because such features are common to both choices and therefore might be felt to be of no help in making a discrimination (in the same way that the common cue in our word-frequency study was no help and may even have harmed recognition performance). The result of deemphasis of orthographic encoding might be improved HF performance, if semantic coding is thereby emphasized, and if semantic coding is the best discriminator for HF items (highest *c* to *d* ratio). On the other hand, elimination of orthographic encoding might harm LF performance because these items are not well discriminated by semantic features (and would not have been easy to code semantically at study).

This hypothesis requires a deeper analysis to justify the stipulated discrimination differences. An item cue can be encoded in terms of a variety of features, containing two important subclasses, *orthographic* (O) and *semantic* (S). The overall strength relating an item cue to an image can be viewed as an appropriate combination of the individual strengths for O, S, and any other classes of features. At the time of study it seems reasonable that LF items are coded less in terms of S and more in terms of O because these items are harder to code semantically and, in addition, tend to be more orthographically distinct. The result would be a higher c to d ratio for the O component than for the S component. Just the reverse reasoning should apply to HF items, so that a higher c to d ratio would be expected for the S component. For either item type, then, it would be most helpful for recognition performance to encode the item at test only in terms of the component that is the best discriminator. Subjects may not do this most efficiently unless they are led to do so by fairly obvious test manipulations like orthographically matched forced choices. (Also note that this analysis suggests that it is not necessarily optimal to match the study encoding at test because the d values are independent of study encoding, and it is the c (and b) to d ratio that determines recognition performance).

Forced Choices Between Two Distractors or Between Two Targets

Another very interesting set of findings on word-frequency effects were collected by Glanzer and Bowles (1976). Subjects were given mixed lists of LF and HF items. Six different types of forced-choice tests were given as follows: (a) HF target–HF distractor, (b) LF target–HF distractor, (c) HF target–LF distractor, (d) LF target–LF distractor, (e) LF target–HF target, and (f) LF distractor–HF distractor. Subjects were not told that the last two "trick" choices would be used and, presumably, chose the item that they were more sure was on the list. The probabilities of correct choices for the first four conditions were 0.71, 0.82, 0.83, and 0.91, respectively. In addition, a HF distractor was chosen over a LF distractor with probability 0.65, and a LF target was chosen over a HF target with probability 0.64.

Our model can be applied to this situation directly, assuming that scaled familiarity distances from a criterion are used to make forced choices. In particular, assume that an evaluation of the familiarity of each of the forced-choice test items is carried out independently. Two criteria are chosen, one for HF items and one for LF items. The subject assesses the frequency of a test word, chooses the appropriate criterion, subtracts the criterion from the familiarity value attained, and then scales the resulting difference by dividing by an estimate of the standard deviation of the distractor distribution for that frequency item. After this process is carried out for each test item, the subject compares the scaled values and chooses the item with the highest value to be the target.

The essentials of this situation can be graphed as follows: We take the familiarity distributions for HF targets and distractors (the same distributions that underlie the predictions in Figure 9) and scale them by dividing by the standard deviation of the distractor distribution; an analogous scaling is carried out for the LF items. These two sets of distributions are then aligned so that the respective criteria are in the same location. The result is the four distributions graphed in Figure 31. From left to right these distributions are for LF distractors, HF distractors, HF targets, and LF targets. If we now make the mandated assumption that in a forced choice of items from any two of these distributions the subject will choose the higher (scaled) value, then it is easy to see that the findings of Glanzer and Bowles (1976) will be predicted in qualitative fashion. (For quantitative fits of the data, we would need to estimate parameters).

We can hardly claim much insight for the approach shown in Figure 31 because this is exactly the framework proposed by Glanzer and Bowles (1976). Glanzer and Bowles proposed a particular model according to which the various distributions were generated—and that underlying model had few similarities to ours—but the general approach is identical. We report these findings because our model does predict the ordering of the four distributions necessary to produce the correct predictions.

We should note finally that a number of researchers have obtained results that are consistent with those obtained by Glanzer and Bowles (e.g., Shepard, 1967). Our model would handle these findings in a manner analogous to that described above.

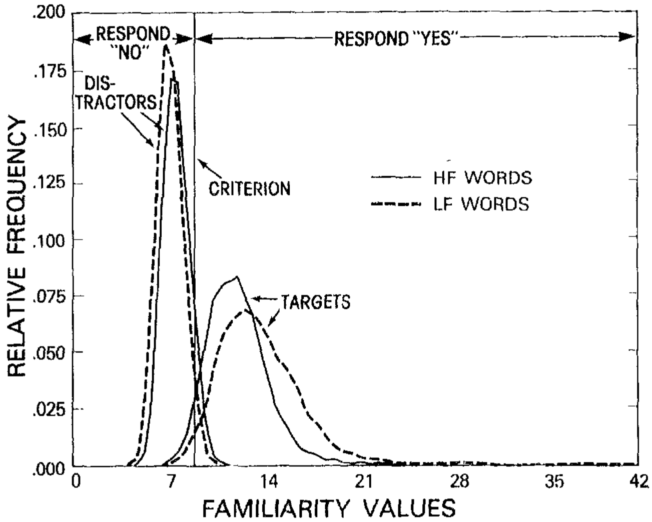

Figure 31. Scaled distributions for high-frequency (HF) and low-frequency (LF) words. (The distributions are the ones that give rise to the predictions in Figures 23 and 24.)

Final Remarks

The quantitative applications of the model presented in Section 3 amplify considerably the qualitative applications of the Section 2. It seems clear that the model can predict a great deal of data from recognition and recall studies that measure levels of accuracy, including studies manipulating word frequency.

Although there are a fairly large number of parameters in the model, the essentials of the predictions are due to four working parameters, *a*, *b*, *c*, and *d*, with *a*, *b*, and *c* reflecting coding at study and the match of coding at test to that at study, and *d* reflecting preexperimental factors and coding at test. Each of these parameters has a well-defined role, is quite distinct in its effects, and should respond appropriately to experimental manipulations. The context strength is the *a* parameter, the interitem strength is the *b* parameter, the self-strength is the *c* parameter, and the residual strength (based on preexperimental factors) is the *d* parameter. We have seen in Sections 2 and 3 that the results of experimental manipulations can usually be handled in the model by making natural and appropriate adjustments in the values of the parameters as required by the changes in the experimental designs. Nevertheless, the progress thus far must be considered very preliminary. At the least, quantitative applications to many of the paradigms are required. Beyond this point, it is essential that the model be extended and/or modified to deal with reaction time and confidence-rating data. We consider this possibility and other extensions in the Section 4 of this article.

4. Extensions

In Section 4 we briefly discuss possible extensions of the SAM model of recognition to other types of paradigms and response measures.

Effect of Number and Spacing of Target Presentations

Increasing the number of target presentations causes both recall and recognition performance to rise (e.g., Atkinson & Juola, 1973; Murdock, 1962; Ratcliff and Murdock, 1976). Increasing the spacing of such presentations also causes recall and recognition performance to rise (e.g., Hintzman, 1974; Hintzman, Block, & Summers, 1973; Madigan, 1969).

The simplest case occurs when presentations are massed. In this case it seems plausible to treat *n* successive presentations each of time t as equivalent to one presentation of time *nt*. When this is done, the predictions for increasing numbers of presentations are identical to those shown in Figure 13 for increasing presentation time (at least for the case when all of the list items are repeated n times in massed fashion).

When presentations are spaced, the situation is much more complex—so much so that a detailed treatment is not possible in this article. Nevertheless, a tentative conclusion can be reached. It must be realized that the equivalence assumption made in the case of massed presentations does not hold. It seems more likely that n spaced presentations would give rise to n separate images, each with a (somewhat) independent variability contribution and each with differing interitem strengths to different items. For recognition, all of these changes produce a performance increase compared to the massed presentation case.

Frequency Discrimination

It ought to be possible to apply our model to the paradigm of frequency discrimination, in which the subject attempts to discriminate between an item presented m times and some other item presented n times. This can be viewed as a generalization of recognition (for which $m = 0$ and $n = 1$). For m and n that are both greater than 0, both test types are familiar, but more presentations should result in greater familiarity on the average, other factors being equal.

The considerations mentioned above during the discussion of multiple presentations also come into play when predicting frequency discrimination. Matters are especially difficult because most studies utilize a spaced presentation method. For a few reasonable sets of assumptions we have looked at, performance in discriminating n occurrences from $(n + 1)$ occurrences is a decreasing function of n. This prediction matches the findings in the literature (e.g., Hintzman, 1974; Rao, 1983). Other interesting issues in this area, such as (a) the automatism of frequency knowledge (with which the model seems to be consistent: The model predicts that the frequency discriminations should take place under incidental instructions, as shown by Zacks, Hasher, & Sanft, 1982), (b) spacing effects (which may be due in part to differential coding at storage; see Hintzman, 1974), and (c) variations in presentation time for items of different frequency (the model may predict a choice of the LF item if the times for this item are longer) are beyond the scope of this article.

Multiple Test Items

In the experiment we reported, two items were presented for test in the paired and in the cued conditions. In the paired condition, both items were relevant, but the decision could have been made on the basis of either item alone, because distractor pairs consisted of two new items. In many studies, this possibility is eliminated by using distractor pairs consisting of one new and one old item or two old items from different study pairs. Many experiments of this type involve sentence memory, with distractor sentences consisting of studied words in new arrangements (e.g., J. R. Anderson, 1976). What is relevant for present purposes is the fact that the SAM recognition model produces a familiarity value for an intact group of items that is higher on the average than the familiarity value for a rearranged group of items. This fact provides a basis for discriminating intact pairs from rearranged pairs in recognition tests, a feat that subjects are quite capable of accomplishing (e.g., Humphreys, 1976, 1978).

The basis for the extra familiarity is the high interitem strength between members of a studied pair; the interitem strength between members of a rearranged test pair is usually lower and often at the residual value. It is easiest to follow the argument by way of an example, as indicated in Table 2.

Assume for the example that the study list contains just two pairs, A–B and D–E. Recognition tests of pairs could be of four types: A–B (intact), A–D (rearranged), A–F (old–new), and F–G (new–new). Table 2 gives possible retrieval structures for this case under two assumptions: Structure 1 assumes that the b parameter for the items studied in a pair is greater than the residual value, whereas Structure 2 assumes that $b = d$. The familiarity for the different types of test pairs is indicated at the bottom of the table. Note that for $b > d$, all types are discriminable (Structure 1), but for $b = d$, intact and rearranged pairs are not discriminable. Thus the interitem parameter, b, controls this type of discrimination. These

Table 2
Retrieval Structures and Pair Familiarity for a List of Two Word Pairs, A–B and D–E

Structure 1: Interitem strength, *b*, within pair is higher than residual					Structure 2: Interitem strength, *b*, within pair equals residual value				
Image	A	B	D	E	Image	A	B	D	E
C	10	10	10	10	C	10	10	10	10
A	10	10	1	1	A	10	1	1	1
B	10	10	1	1	B	1	10	1	1
D	1	1	10	10	D	1	1	10	1
E	1	1	10	10	E	1	1	1	10
F	1	1	1	1	F	1	1	1	1
G	1	1	1	1	G	1	1	1	1
Familiarity					Familiarity				
F(AB) = 2020 (intact)					F(AB) = 220				
F(AD) = 400 (rearranged)					F(AD) = 220				
F(AF) = 220					F(AF) = 130				
F(FG) = 40					F(FG) = 40				

qualitative conclusions hold true even if fractional weights are assigned to the cues, a point that is important in light of the results of Section 3.

In his work on this topic, Humphreys (1976, 1978) contrasted *item information,* used mainly to discriminate old from new items, and *relational information,* used mainly to discriminate intact from rearranged pairs. This contrast is captured in our model by the contrast between the *c* and *b* parameters. To this degree our model is consonant with Humphreys's, although our detailed assumptions certainly do not match his, and we do not yet know whether our model can fit his data in quantitative fashion.[15]

We conclude Section 4 by mentioning one other factor in our model that can lead to discrimination of intact from rearranged pairs. This factor is the match of the encoding of a given item at study to the encoding of that item at test. The paired member also presented in these two cases can be viewed as an *encoding context* (see Section 2). When an intact pair is tested, the encoding of both items might match the study encoding better than would a rearranged pair. This would be reflected in higher values of both the *c* and *b* parameters. Note that this encoding factor would be especially important if either the test or the study pairs were chosen to lead to an unusual coding for an item. In cases with random word pairings, the dominant encoding might tend to occur at both study and test for both the intact and the rearranged cases, and therefore the match of encodings might not play an important role.

Recognition of Items From Categorized Lists

Consider a paradigm in which items from *n* categories are presented on a list, followed by a yes–no recognition test for both list items and distractors from each of those categories (e.g., Rabinowitz, 1978). This situation is not difficult to model: Assume that an item has within-category residual strengths that are higher than residual strengths to items in other categories (i.e., within-category strengths = d_w; between-category residual strengths = d_b; $d_w > d_b$). Assume also that when category size varies, different criteria are chosen for the different categories. Ignoring for the moment the

[15] Mandler (1980) has suggested that an analogy can be drawn between Humphreys's item information and the familiarity phase of recognition, and Humphreys's relational information and the search phase of recognition. Obviously, our familiarity model can predict the basic phenomena in question with a familiarity approach alone. Note also that whereas we share Humphreys's emphasis on the distinction between relational and item information, and his emphasis on cuing effects, the present model does not incorporate a search phase in cued-item recognition, a key component in Humphreys's model (see Humphreys & Bain, in press).

possibility of using a category cue in addition to item and context cues, predictions are relatively straightforward.

Category size effects are seen for reasons similar to those that explain list-length effects, although the relative size of the decrement as category size increases is reduced in a multiple-category list compared to a one-category list. For example, the decrease from a 5-flower list to a 10-flower list is lessened if 40 words from other categories are added to each of these lists. The reason is simple: d_b is not zero, so added items in other categories make each category appear to have extra items, the number extra depending on the ratio of d_w to d_b; the added number does not depend on initial category size. If the total list length is held constant, and the categories making up the list are varied in size, as in Rabinowitz (1978), then performance is still predicted to drop as category size increases (because d_w is larger than d_b), though at a somewhat different rate. These predictions are confirmed by reanalysis of Rabinowitz's data.

An interesting question arises if one compares performance for a list of n items all from one category with a list of n unrelated items (each tested with distractors similar to the list items). Because residuals between the category items are larger, performance should be lower in this case. Although we do not know of a direct test of this point, Kintsch (1968) compared recognition of a random 30-word list with a list composed of six 5-word categories. These gave roughly equal performance, possibly because the improvement due to using a category size of 5 (rather than 30) offset the harm caused by extra-high within-category residual strengths. (By way of comparison, note that our recall model predicts recall of n words from one category to be superior to recall of n random words, for several different reasons—see Raaijmakers & Shiffrin, 1980).

What would be the effect of adding a category cue to the probe set in addition to the item and context cues? Because the category cue would have higher strengths to images of items in that category than to images of other items, and because this effect would multiply the same effect caused by item cue strengths, the noncategory images would contribute proportionally less familiarity. Although a related effect occurrs for distractors, it is much smaller, so that performance would increase for the category in question. In addition, category performance would depend more on the size of the tested category and less on the size of the whole list. (Note that fractional cue weights might reduce or eliminate the effect of adding a category cue to the probe set).

When all list items and distractors are from one category, a good case can be made for eliminating a category cue, because it increases familiarity for both targets and distractors, perhaps in a fashion that would reduce performance. It could even be argued that it is wise to encode an item cue in such a case so that category information would be de-emphasized as much as possible. The residuals and the match parameters, $m(b)$ and $m(c)$, could both be reduced by such a de-emphasis, perhaps leading to an overall improvement. Such a possibility depends on a subject's ability to exercise control over encoding of items and cues. In any event, performance on a list and a test composed all of items from one category might well be improved by de-emphasis of category information, so that comparisons to performance levels for unrelated-item lists might be closer than otherwise expected.

Continuous Recognition Paradigms

Virtually all of the discussion in this article is directed toward study–test paradigms. We have done this because the previous recall work (Raaijmakers & Shiffrin, 1980) suggested that context changes during the course of a given list either do not occur or may be assumed to have negligible impact. A very different situation obtains when a continuous recognition paradigm is used: Words are presented in a continuous sequence and tests are alternated with presentations throughout. In this paradigm we must assume that context is changing throughout the experimental session, so that as study–test lag increases, study context differs more from test context and performance decreases (e.g., see Hockley, 1982, for a continuous recognition paradigm and Shiffrin, 1970, for a continuous recall paradigm).

We do not analyze continuous recognition paradigms in this article, but one important point can be made. Most of the effects that are seen as a function of study–test lag can be explained by a change in context. This would be handled in the model by changes in the parameter $m(a)$, which can be instantiated in the simulation by decreases in the value of the context strength in the retrieval structure, as a function of increasing lags. The familiarity would be based on a sum over all of the images of items up to that point in the session. As a target is tested at increasing delays, it contributes less and less of the total familiarity despite large self-strength, because the self-strength is multiplied by decreasing context strength.

Short-Term Recognition

In the applications of the model to our data and to the results from the literature, care has been taken to exclude possible contributions to performance that might be due to tests of items in STS. It seems likely that an item still in STS (say, undergoing rehearsal) at the time of a recognition test is given a very rapid and accurate response, regardless of long-term factors that would produce other results. For example, if the last item presented in a long list had been the first tested, it is very likely that both the accuracy and latency would be well off the distribution expected for other items in such a long list.

How might short-term recognition be incorporated in the model? For recall, we have made the simple assumption that the short-term rehearsal buffer is dumped. For recognition accuracy, a related assumption could be made that a correct response would be given if a tested item is currently in STS.

It is interesting that a number of theorists have put forth models to handle recognition in both the long-term and short-term spheres (e.g., Murdock, 1974; Ratcliff, 1978). The debate between those who prefer two (or more) systems and those who prefer one has been highly active for 20 years without any definitive resolution. It is even conceivable that our familiarity model could be extended to short-term data with good success. Nevertheless, we prefer to retain a two-store system so as to maintain the correspondence with the recall part of the model.

Response Times

A great deal of the modern research into recognition memory involves reaction time measures (e.g., Atkinson & Juola, 1973; Hockley, 1982; McNicol & Stewart, 1980; Murdock & Dufty, 1972; Ratcliff & Murdock, 1976). The present model remains incomplete until a mechanism for the production of response times is included.

A Descriptive Approach to Response Times

Within the context of the present model, the simplest way to generate predictions for response time involves assignment of times in accord with the distance from the criterion. It seems reasonable that familiarity values far from the criterion should be associated with fast reactions and low variability; values close to the criterion should be associated with slow and variable reaction times.

For the reasons discussed in Section 3, it is not sufficient to use only the distance from the criterion; because the variances can differ greatly with item type, the distance must be scaled. Perhaps it is simplest to divide the distance from the criterion by the standard deviation of the distractor distribution. The resulting scaled distance may be called D. Then a model for reaction times might propose that a reaction time distribution be associated with each value of D, with mean $\mu(D)$ and variance $\sigma^2(D)$. Most likely, we should set $\mu(D) = f_1(D)$ and $\sigma^2(D) = f_2(D)$, where both f_1 and f_2 are monotonically decreasing functions of the absolute value of D.

A model of just this sort was fit quantitatively to the reaction time results of our experiment reported in Section 3. The means and variances of the correct response times were fit quite accurately, as were the patterns of the error-response times (though the absolute values of the means and the variances of the error-response times were overpredicted). We do not present these results and predictions because the model is purely descriptive. Although the approach fit the data reasonably well, all that this demonstrates is a regular and lawful relationship between scaled criterion distance and response time, and hence a regular relationship between accuracy and latency measures.

Toward a Process Model of Response Times

A descriptive model of the type proposed in Section 3 has many drawbacks and limitations. To give just one example, there is no obvious way to treat data concerning trade-offs between speed and accuracy (e.g., Corbett, 1977; Dosher, 1981; Pachella, 1974; Reed, 1973, 1976; Wickelgren, 1977; Wickelgren & Corbett, 1977).

We have not yet developed a process model for response time generation, but a random walk model of the type proposed by Link (1975) seems a reasonable candidate. One way to implement such a model follows: A scaled familiarity value is obtained for any test item, as in the present model. This scaled value is used to generate a drift rate (and distribution) for a random walk that takes place until either a positive response barrier or a negative response barrier is reached. The barrier reached determines the response accuracy, and the response time is determined by the number of steps needed to reach the barrier.

Note that this model is not equivalent to the model put forth in earlier sections of this article. It is possible to have a positive scaled value of familiarity (which in the model of the earlier sections always produced a positive response) and yet have the random walk proceed to the negative response barrier (and vice versa). The closer the barriers, the more often such errors occur, and the faster are the response times. In fact, it is in just this fashion that speed–accuracy trade-off results are dealt with in a random walk approach—barriers are moved in to speed response time and lower accuracy or moved out to slow response time and improve accuracy. We suspect that an appropriate choice of barriers and drift rates would allow the random walk model to mimic closely the accuracy predictions of the present model. Whether the accuracy and latency data could be fit simultaneously by a random walk model is a question that must be left for future work.

Criticisms of Criterion Models for Response Times

Although we have not extended the SAM model to predict response times, we feel confident that such an extension is possible, along the lines suggested above. Even if the random walk version of the model is excluded from consideration, the descriptive version can handle most of the results. Our suggested descriptive version is similar to a number of strength models that have postulated that reaction time is determined by distance from a criterion (e.g., Norman & Wickelgren, 1969). This approach has usually been used to predict short-term recognition data. It has been dismissed as a viable candidate to explicate long-term recognition because it (a) predicts error variances to be smaller than correct variances, when the reverse is actually found (Murdock & Dufty, 1972); (b) is not a dynamic theory; that is, there is nothing than can change in the model as variables change (such as test position—Ratcliff & Murdock, 1976); and (c) makes no provision for response bias (Pike, 1973). None of these criticisms hold for an extension of the SAM model to the reaction time domain.

Confidence Ratings

Many of the same concerns that were discussed with respect to response times apply to confidence ratings as well. Descriptively, one could suppose that subjects utilize a number of criteria along the familiarity dimension, assigning confidence ratings in accord with the criteria between which a given observation of familiarity falls. Alternatively, in a random walk model, the confidence ratings could be made a function either of the drift rate, the number of steps to reach a barrier, the placement of the barrier, or some combination of all three (see Ratcliff, 1978).

5. Assessment and Generalization

Comparisons of the SAM Model to Other Theories

Our model postulates a deep-rooted relationship between recall and recognition. Both processes make use of the same types of information during retrieval and both are cue-dependent retrieval processes. In fact, the same quantitative expression serves as the basis for the recall and recognition models. The difference between the two paradigms lies in how the information is used. In recognition, the strength of the total amount of information activated is used to make a familiarity decision. In recall that same information is used to determine sampling and recovery during an extended search.

The relationship we have specified is quite different from the relationship specified by *embedded two-process theories* (Anderson & Bower, 1972; Kintsch, 1970), and *two-phase theories* (Atkinson & Juola, 1973, 1974; Mandler, 1980), as discussed in the introduction.

Moreover, even the familiarity component of such two-phase models (Model 3 of Table 1) differs significantly from our model of familiarity. Atkinson and Juola (see also Atkinson & Wescourt, 1975) postulate that the familiarity process is based on the level of activation of the word (that was directly accessed) in the lexicon (i.e., semantic memory)—the decision is thus context independent. Mandler (1980) also argues that the familiarity process is context independent, being based on *intra*item aspects of the stored word rather than *inter*item connections (which are used primarily in search). Our conceptualization of a familiarity process is diametrically opposed, because the decision is based on a sum of familiarity values across all the episodic images in the list and is not based on the familiarity value of a particular item. Whereas it is true that intraitem factors have an important role in our model, because the self-coding of targets is one of the two crucial factors in raising the familiarity value for targets above that of distractors, such intraitem effects are embedded in a general *interitem episodic* sum (see Equations 7 and 8). In fact the activation by a distractor of its own semantic image does not even enter into the familiarity value on which a decision is based.

Our conceptualization of the familiarity process is closer to that used in models that propose simultaneous access to multiple memory images or to a single, holistic memory store. In such models the memory decision is based on an activation of many items rather than on direct access to a specific memory image. Such models have been proposed for perceptual tasks (Pike, 1973), short-term memory tasks (Anderson, 1973; Cavanagh, 1976; Pike, Dalgleish, & Wright, 1977), free recall (Metcalfe & Murdock, 1981; Murdock, 1982), paired-associate learning and forgetting (Eich, 1982), and recognition memory (Ratcliff, 1978).[16] Ratcliff's model, although it emphasizes latencies and is restricted to recognition, comes closest to predicting the types of data in which we are interested, and is worth a closer examination.

Ratcliff assumes that multiple searches occur in parallel. A test probe activates each image in memory and causes it to begin a random walk between a *yes* barrier and a *no* barrier. If any of the walks reaches the yes boundary before all reach the no boundary then a positive response is given; otherwise a negative response is given. Negative responses can be rapid because the negative boundary is much closer than the positive one, although the positive drift rate (if a target is tested) may be faster than all of the negative drift rates. Drift rates are determined by similarity of the test item to each image in memory.

The model is similar to ours in that recognition does not involve direct access to one image but rather involves the activation of a large portion of memory. The details of the models differ quite a bit because Ratcliff's model is a process model of activation in real time, whereas we have been more interested in asymptotic familiarity. (Furthermore, Ratcliff makes no distinction between short- and long-term memory and applies his model to results in both domains.) We do not necessarily regard Ratcliff's approach as inconsistent with ours in a global sense. In fact, as we discussed in Section 4, a real time–process version of our model may well be conceptualized as a random walk. (We prefer a random walk model with one rather than multiple walks because we are not enamored of a model in which negative responses must wait until all of some very large but unspecified number of random walks reach a negative barrier.)

Next consider the work of Tulving (1974, 1976; Tulving & Watkins, 1973), who suggested that recall and recognition involve the same cue-dependent processes and information. All that distinguishes them is the amount and the type of information available at test

[16] The Glanzer and Bowles (1976) model is, in effect, a model of this type. Although only the test item's features are contacted, all the other list items have an effect because they caused tagging of some of the features of the test item when they were first presented.

that serve as cues. Our model is like Tulving's in that we also posit cue-dependent retrieval systems and place a great deal of importance on the types of cues available at test. We agree with Tulving that recall and recognition make use of basically the same types of information. However, we differ in that we postulate quite a different process for recall and recognition, whereas Tulving does not.[17]

Strengths and Weaknesses of the SAM Model for Recognition

We are reasonably satisfied with the predictions for accuracy measures in recall and recognition. The model has been able to fit a wide variety of data in the domain of episodic memory, in the paradigms of recall and recognition, with a common set of assumptions and parameters and with sensible values of the parameters. That is, when experimental manipulations are carried out that ought to change the values of certain parameters, the changes in those parameter values actually do predict the new data.

Another strength is the relative simplicity of the model. We have tried not to encumber the model with all sorts of special processes designed to handle particular, troublesome results. It is also gratifying that the model has produced a number of accurate predictions that we did not anticipate prior to carrying out the simulations. Perhaps the greatest strength of the present model is the tight link between recall and recognition processes that is produced by the use of the denominator of the sampling equation as the decision variable for recognition judgments. Even though the familiarity mechanism for recognition is very different from the search process for recall, there is a common basis that underlies both processes. It is this fact that allows us to produce joint predictions for recall and recognition with the same model using the same parameters.

The weakest part of the present model is its preliminary nature: Predictions for response times and confidence ratings have not yet been produced. Of the applications in the accuracy domain, the main weakness is the lack of a model for choosing criteria. Another potential problem is the failure to include a search component during recognition. We consider these problems in the following sections.

Criterion Placement

In the simulations of Section 2 an arbitrary rule was used to choose criteria; in each case a fixed point was chosen between the means of the target and distractor distributions. This rule was used only as a demonstration device; in fact, the subject is assumed to choose a criterion appropriate for each test setting, as was assumed in the quantitative simulations of Section 3. The question then arises: How is an appropriate criterion chosen? This issue is an important one that cannot be addressed in detail in the present work, but we have some suggestions.

First, it is clear that a subject with some knowledge of the expected proportions of targets and distractors can adjust the criterion during the course of testing until a reasonable setting is reached. This solution has several problems. First, the setting of the criterion on the initial test trials is likely to be very imprecise. Second, subjects can produce sensible recognition results when tested with targets only, and no distractors whatever. In this case, it is possible that a criterion could be chosen (over test trials) to produce some arbitrary level of misses, but it is difficult to see why any particular level should be chosen.

These considerations suggest a second solution, which may appear more plausible. During the list presentation, the subject is constantly being presented with new items (in effect, distractors) for each of which a familiarity value can be calculated. Although these familiarity values change during the course of the presentations, the subject can nevertheless use the last several presentations of the list to form an idea of the mean and variance of the distractor distribution. (It may also be possible to get an idea of the characteristics of the target distribution through recall of earlier items in the list). Given this knowledge of the famil-

[17] Note that Tulving has recently modified his view of the relation between recall and recognition to allow different processes to occur in recall and recognition (Tulving, 1982a, 1982b).

iarity distribution for distractors, it should be fairly easy to choose a sensible guess for a criterion setting, a setting that could be fine tuned as needed during the test sequence.

A somewhat different problem concerns the placement of criteria for different item types as was needed for HF and LF items in Section 3. One problem involves memory load—if many item types are used in a list, how can all the criteria be kept in memory and used appropriately? A second problem concerns speed of response—when a test item is presented how can the item type be ascertained, and the appropriate criteria retrieved, rapidly enough to allow fast recognition responses to be given? An answer to both of these questions might involve an automatic adjustment mechanism that utilizes nonepisodic characteristics of the test item to scale the values of familiarity accordingly. Such a solution is attractive (although not yet implemented) but would not work if the need for different criteria is episodically induced. For example, a list made up of differing numbers of members of different categories requires different criteria for each category. The basis for the choice, however, would not be inherent in the category but instead would be the experimenter's assignment of presentation numbers to categories. In this case, an automatic mechanism seems unlikely, and an appropriate adjustment in criterion would have to be made by the subject as the category membership of each test item is ascertained.

Is There a Search Component to Recognition?

Several types of data tend to suggest the presence of search processes and hence the insufficiency of the familiarity-only model (Model 4 of Table 1). One source of evidence concerns false alarms to distractors that are similar to list items. A similar distractor would have a higher than normal familiarity and hence should give rise to increased false alarm rates. However, this is not always the case. If enhanced false alarms are not seen, then it may well be that information recovery from the image of the associated list item is occurring. The recovered information might then be used to inhibit a false alarm. Perhaps such a process explains the failure to find excess false alarms to synonyms in the study with results shown in Figure 4: Locating the list image paired with the synonym could allow the subject to discount the familiarity of the cue.

A different example of this kind is found in a study by Tulving (1982a). After presentation of a list of items, a test list was presented consisting of 25% list items as targets, 25% control distractors (random), 25% rhyme distractors, and 25% associated distractors. Group 1 subjects performed a yes–no recognition judgment. Group 2 used the test items as cues and wrote down next to each a list word that the cue suggested (including the cue itself, if it was a target). An analysis showed that distractors that were better cues for Group 2 (they led to more recalls of their matched-list members) were given fewer false alarms by subjects in Group 1. Such a result is hard to explain by a familiarity model alone. It seems likely that good cues used as distractors for Group 1 led on occasion to recovery of the list item matched to that cue; having found a similar list item, the subject might then be able to decide that the cue was a distractor, despite its level of familiarity.

These results suggest that sampling of an image may sometimes occur and, thus, inhibit false alarms, but other results suggest that such sampling of the relevant image occurs only occasionally. For example, in a continuous-recognition experiment, false alarms to similar distractors are a nonmonotonic function of lag, with the peak somewhere between one and five intervening items. (MacLeod & Nelson, 1976, tested lags of zero, five, and higher numbers and found a peak at five. In a similar study, we found a maximum false alarm rate at a lag of one). Presumably at zero lag, the item is not given many false alarms because the relevant list item is still in STS at test. At larger lags, though, the increased familiarity caused by a test at a short lag apparently overcomes whatever tendency there is for the matched-list item to be sampled by the similar distractor and apparently overcomes residual contamination from STS retrieval.

Results like those obtained by Tulving suggest that a search component is at least sometimes used during recognition. Whether the search component occurs only when familiarity does not give a clear answer, as in Figure 1, or whether the image-recovery process occurs quickly (on some trials) and precludes use of the familiarity system altogether, is an open question.

Evidence of a somewhat different kind has been reported by Mandler et al., (1969). After sorting a list of words into *n* categories, a subject attempts either recall or recognition, either immediately or after a delay. Larger values of *n* (up to about 7) lead to better recall at all test delays, but recognition does not depend on *n* until long delays occur. Mandler argues that search is utilized relatively more and familiarity relatively less as recognition-test delay increases, thereby explaining the result if two additional assumptions are true: (a) Search is better after sorting into more categories, and (b) familiarity judgment is unaffected by sorting into more categories. This sort of argument is a bit indirect but may provide evidence for search processes in recognition. Whereas data concerning a lessening of false alarm rates for similar distractors or for good recall cues seem to suggest automatic sampling and recovery occuring in parallel with the familiarity process, Mandler et al.'s (1969) data seem more compatible with a search carried out after the familiarity process fails.

There may be other data that indicate the presence of search during recognition, but the appropriate analyses have not yet been carried out. One paradigm of this kind is that of multiple-item recognition. Both Humphreys (1978; Humphreys & Bain, in press) and Mandler (1980; Mandler, Rabinowitz, & Simon, 1981) have analyzed multiple-item recognition by collecting data on the recognition, the cuing ability, and the recall recoverability of each individual item being tested. Through appropriate analysis, it may be possible to argue that joint recognizability is a function of both familiarity and search in combination. We must reserve judgment on this issue until quantitative fits of our model to such data are undertaken, because it is possible that our model could fit the same data, although for different reasons.

It is important to emphasize that we are not opposed to the concept of a search component in recognition. To the contrary, the fundamental nature of the SAM approach requires that subjects may carry out an extended search during recognition paradigms, if they so desire. Furthermore, it is entirely consistent with our approach that a sample of an image from memory may occur automatically, in parallel with the evaluation of familiarity. Thus, the question is really not whether search processes occur, but the degree to which such processes occur. Perhaps it is surprising that the results reported in this article have shown that most of the experimental findings are explicable with a global familiarity model with no search component at all. This finding should not be used to come to strong conclusions concerning the use of search. Because our global familiarity model is designed to predict many findings that are normally thought to indicate the presence of search, it is no easy matter to discriminate our pure familiarity model from one that relies to some degree on search.

Toward a General Model for Recall and Recognition

Regardless of the status of the data concerning evidence in favor of search during recognition, the underlying logic of the SAM model requires that the subject be able to utilize search if he or she so chooses. We begin therefore by describing a generalization of the SAM model that incorporates a search component.

The generalized model is depicted in Figure 32. As shown, there are two points at which search processes can contribute: an initial, perhaps automatic, single sample of an image that occurs in parallel with familiarity generation, and an extended search that may follow familiarity generation.

Next we wish to place the model in the larger context of a general information-processing framework that covers the memory system from initial feature encoding to final retrieval decisions.

On initial presentation of an item, a series of encoding and retrieval processes takes place which is largely automatic in the early stages. The general framework is depicted in Figure 33. The set of automatically encoded information is termed the *autofield.* We use the terms *encoded* and *retrieved* interchangeably here, because initial encoding is a form of re-

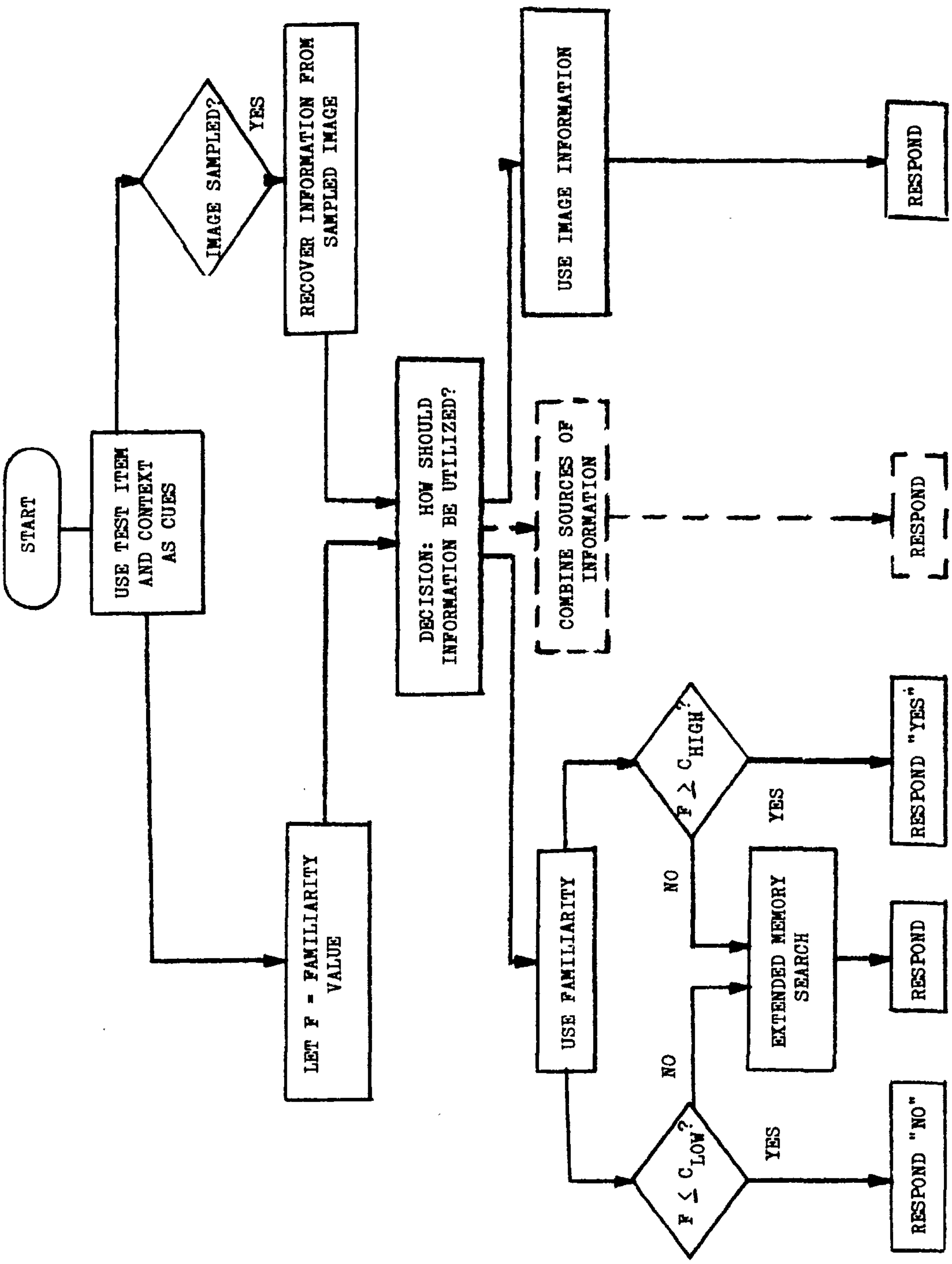

Figure 32. The extended SAM model for recognition, modified to take into account the possibility of image sampling and/or extended search, in addition to familiarity judgments.

trieval from LTS. When we say, for example, that a straight-line feature is encoded, we mean that the system has processed the physical input to the point where a particular straight-line feature in LTS has been activated and made available to the system so that further processing can take place.

The information that drives the encoding process includes not only the sensory input from the external environment, including the contextual surround in addition to the nominal stimulus defined by the experimenter, but also internal context deriving from the subject's physical state and mental activities at the time of input. The system is assumed to act continuously to encode the changing environment, but we consider only the stages following the occurrence of some event of particular interest, such as the presentation of a word in a recognition experiment.

It is our view that the automatic encoding process produces (over a short period of time, perhaps several hundred milliseconds) a large number of different types of informational

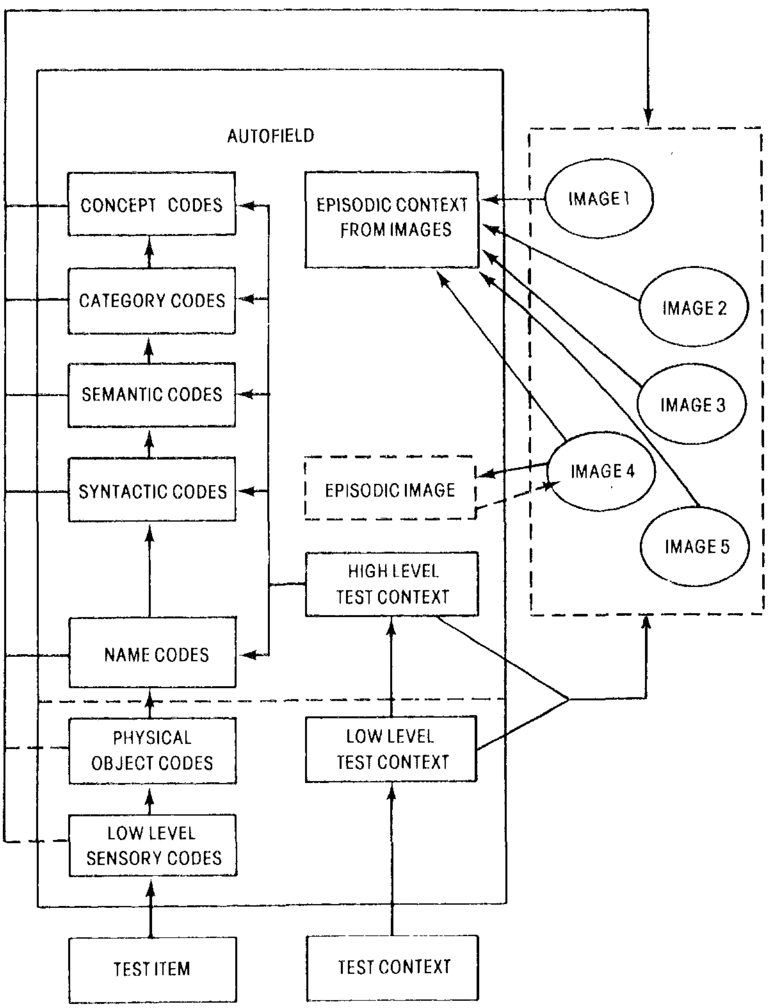

Figure 33. A framework for initial encoding according to the SAM model. (The autofield represents the result of automatic encoding processes, based on the test cues in the probe set. The autofield contains a region of temporal-contextual information activated from recent episodic images. This information is used both for recognition decisions and as a basis for sampling images during an extended memory search.)

features, organized by content, and not by episodic images. Occasionally an episodic image may be activated as part of the automatic encoding process (see Figure 32), but with this exception, the images of list items and recent events are not in the autofield. However, context from these images is in the autofield, in a localized region called *episodic context.* It is this episodic context, integrated together in appropriate fashion, that is used to make familiarity decisions in recognition. The same episodic context governs search, in the following manner: An element of context is sampled at random from the contextual region of the autofield and is traced back to the image from whence it arose. This is the sampling process in the memory search.

Consider next the way in which the various codes are extracted and placed in the autofield. The typical assumption is made that processing takes place in an ordered sequence, though strict seriality is not implied. Some of the codes that might be retrieved are indicated in Figure 33, ranging from low-level sensory features to high-level concepts. Of course the processing depends on the context as well as on previous features in the chain, so as upper stages are reached the features retrieved may become biased by the context as much as they are determined by the item itself. Thus one meaning of a word might be retrieved ahead of a more usual meaning, given the presence of a biasing word in the context. Presumably, the later in the stage of processing, the larger is the potential influence of context. The arrows leading from context to the episodic images indicate that the similarity of the test context to the context in the stored images causes that context to be activated (in conjunction with the extracted content features that are similar to features in the images; this other path of activation is indicated by the arrows leading from the content features to the episodic images). The arrows leading from the episodic images to the autofield (actually to the episodic context area) show the genesis of the context activation from recent images, that serves as the basis for both familiarity judgments and sampling. The short arrow leading from one of the images to the autofield allows for the possibility that one of the images may be sampled as part of the automatic encoding process in parallel with the other types of encoding. When this occurs, it may be thought of as the first step in the search process, assuming that an extended search is carried out.[18]

Why should the context from all episodic images appear in the autofield, but not the images themselves? The answer is based on the associative basis by which the automatic encoding system operates. The autofield is generated through associations to the probe cues (item and context). Because the probe cues contain context, the similar context from a recent image tends to be activated and placed in the autofield. On the other hand, if a word is presented as a distractor, so that none of its images is recent, then the context from its images will not match that used at test, but the content features of the word will match many stored content features in LTS, and a variety of physical and semantic features are placed in the autofield. When both the context and the item in a sorted image match the test cues (i.e., a target or a similar distractor is presented), then both types of information enter the autofield, each in its appropriate region. (Occasionally, in addition, the image itself is activated and is placed in the autofield.)

It is assumed that the subject can utilize the information in the autofield once it has been retrieved and can do so according to type of information, depending on the requirements of the task. Thus a lexical decision task would be accomplished through use of the lexical information in the autofield, a physical match task through use of the low-level sensory codes, and so forth.

The possibility that our model can be used for tasks other than episodic recognition might eventually prove quite important. As one example, consider the learning of categories. Suppose N_i items are input and assigned to the *i*th category. At test either old or new items might be given to the subject to assign to categories. One approach within the SAM model would utilize search. The test item and the

[18] Where do the episodic images in LTS come from? We propose that each image represents a sample of the information in the autofield that arises due to some event occurrence. (We make no proposal at this time concerning how many of, and how often, these are stored.)

context would be used as cues, and some item would be sampled and recovered (eventually). The category of the first item recovered could be given as the response. This model is consistent with versions of the exemplar model of Medin and Schaffer (1978) and is essentially isomorphic to the *complete-set* exemplar model discussed by Homa, Sterling, and Trepel (1981) and Medin, Busemeyer, and Dewey (in press).

Of course this simple version of a SAM search model is not the only approach within the SAM framework, even if attention is restricted to search models only. A simple generalization would let the subject attempt longer searches. For example, N samples could be made, with a decision in favor of the category that is most often recovered. A further generalization would have the subject use a more intelligent decision rule, assigning a category to the test item on the basis of a similarity decision that results from examining the recovered images. (Note also that the SAM model is perfectly consistent with a mixed-exemplar–prototype model. If the subject should store a prototype separate from the individual images, this could be found during a memory search and could be used to generate a decision).

All of these search models for categorization are subject to the same criticisms that were applied to search models for recognition. That is, fast, accurate, and high-confidence category judgments might be difficult to explain on the basis of search alone. One solution is available by utilizing a variant of the familiarity model for recognition. For example, memory could be probed with three cues: context, test item, and category name. A decision in favor of a match would be made when the total system response (i.e., the denominator of the sampling fraction) exceeds some criterion.

This brief discussion of category learning is meant to provide an example of the type of generalization that may be possible for the present model. Such generalizations are left for future research.

References

Allen, L. R., & Garton, R. F. (1968). The influence of word-knowledge on the word-frequency effect in recognition memory. *Psychonomic Science, 10,* 401–402.

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review, 80,* 417–438.

Anderson, J. R. (1972). FRAN: A simulation model of free recall. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 5). New York: Academic Press.

Anderson, J. R. (1976). *Language, memory, & thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123.

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory and Cognition, 2,* 406–412.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity and directionality in false recognition. *Journal of Experimental Psychology, 77,* 171–179.

Antognini, J. (1975). The role of extralist associate cues in retrieval from memory. Unpublished doctoral dissertation, Yale University.

Atkinson, R. C., Herrmann, D. J., & Wescourt, K. T. (1974). Search processes in recognition memory. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium.* Hillsdale, NJ: Erlbaum.

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance* (Vol. 4). New York: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology (Vol. 1): Learning, memory, & thinking.* San Francisco: Freeman.

Atkinson, R. C., & Shiffrin, R. M., (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2). New York: Academic Press.

Atkinson, R. C., & Wescourt, K. T. (1975). Some remarks on a theory of memory. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance* (Vol. 5). London: Academic Press.

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

Birnbaum, I. M., & Parker, E. S. (Eds.) (1977). *Alcohol and human memory.* Hillsdale, NJ: Erlbaum. (1977).

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory.* Washington, DC: Winston.

Buschke, H., & Lenon, R. (1969). Encoding homophones and synonyms for verbal discrimination and recognition. *Psychonomic Science, 14,* 269–270.

Cavanagh, P. (1976). Holographic and trace strength models of rehearsal effects in the item recognition task. *Memory and Cognition, 4,* 186–199.

Cermak, G., Schnorr, J., Buschke, H., & Atkinson, R. C. (1970). Recognition memory as influenced by differential attention to semantic and acoustic properties of words. *Psychonomic Science, 19,* 79–81.

Corbett, A. T. Retrieval dynamics for rote and visual image mnemonics. (1977). *Journal of Verbal Learning and Verbal Behavior, 16,* 233–246.

Cramer, P., & Eagle, M. (1972). Relationship between conditions of CrS presentation and the category of false recognition errors. *Journal of Experimental Psychology, 94*, 1–5.

Davies, G., & Cubbage, A. (1976). Attribute coding at different levels of processing. *Quarterly Journal of Experimental Psychology, 28*, 653–660.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports, 7*, 337–344.

Dosher, B. A. (1981). The effects of delay and interference: A speed-accuracy study. *Cognitive Psychology, 13*, 551–582.

Duncan, C. P. (1974). Retrieval of low-frequency words from mixed lists. *Bulletin of the Psychonomic Society, 4*, 137–138.

Eagle, M., & Leiter, E. (1964). Recall and recognition in intentional and incidental learning. *Journal of Experimental Psychology, 68*, 58–63.

Eagle, M., & Ortof, E. (1967). The effect of level of attention upon "phonetic" recognition errors. *Journal of Verbal Learning and Verbal Behavior, 6*, 226–231.

Earhard, B. (1982). Determinants of the word-frequency effect in recognition memory. *Memory and Cognition, 10*, 115–124.

Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval. *Memory and Cognition, 8*, 157–173.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627–661.

Elias, C. S., & Perfetti, C. A. (1973). Encoding task and recognition memory: The importance of semantic encoding. *Journal of Experimental Psychology, 99*, 151–156.

Estes, W. K., & DaPolito, F. (1967). Independent variation of information storage and retrieval processes in paired-associate learning. *Journal of Experimental Psychology, 75*, 18–26.

Fischler, I., & Juola, J. F. (1971). Effects of repeated tests on recognition time for information in long-term memory. *Journal of Experimental Psychology, 91*, 54–58.

Garton, R. F., & Allen, L. R. (1968). Familiarity and word recognition. *Quarterly Journal of Experimental Psychology, 20*, 385–389.

Gillund, G., & Shiffrin, R. M. (1981). Free recall of complex pictures and abstract words. *Journal of Verbal Learning and Verbal Behavior, 20*, 575–592.

Glanzer, M., & Bowles, N. (1976). Analysis of the word frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 21–31.

Glenberg, A., & Adams, F. (1978). Type I rehearsal and recognition. *Journal of Verbal Learning and Verbal Behavior, 17*, 455–463.

Glenberg, A., Smith, S. M., & Green, C. (1977). Type I rehearsal: Maintenance and more. *Journal of Verbal Learning and Verbal Behavior, 16*, 339–352.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 311–325.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology, 66*, 325–331.

Godden, D. R., & Baddeley, A. D. (1980). When does context influence recognition memory? *British Journal of Psychology, 71*, 99–104.

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61*, 23–29.

Gregg, V. H. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition*. London: Wiley.

Gregg, V. H., Montgomery, D. C., & Castano, D. (1980). Recall of common and uncommon words from pure and mixed lists. *Journal of Verbal Learning and Verbal Behavior, 19*, 240–245.

Grossman, L., & Eagle, M. (1970). Synonymity, antonymity, and association in false recognition responses. *Journal of Experimental Psychology, 83*, 244–248.

Hall, J. F. (1954). Learning as a function of word-frequency. *American Journal of Psychology, 67*, 138–140.

Hall, J. F. (1979). Recognition as a function of word frequency. *American Journal of Psychology, 92*, 497–505.

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory and Cognition, 4*, 507–513.

Herrmann, D., Frisina, R. D., & Conti, G. (1978). Categorization and familiarity in recognition involving a well-memorized list. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 428–440.

Herrmann, D. G., McLaughlin, J. P., & Nelson, B. C. (1975). Visual and semantic factors in recognition from long-term memory. *Memory and Cognition, 3*, 381–384.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium*. Hillsdale, N.J.: Erlbaum.

Hintzman, D. L., Block, R. A., & Summers, J. J. (1973). Modality tags and memory for repetitions: Locus of the spacing effect. *Journal of Verbal Learning and Verbal Behavior, 12*, 229–238.

Hockley, W. E. (1982). Retrieval processes in continuous recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 497–512.

Homa, D. (1973). Organization and long-term memory search. *Memory and Cognition, 3*, 369–379.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorized information. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 418–439.

Humphreys, M. S. (1976). Relational information and the context effect in recognition memory. *Memory and Cognition, 4*, 221–232.

Humphreys, M. S. (1978). Item and relational information: A case for context independent retrieval. *Journal of Verbal Learning and Verbal Behavior, 17*, 175–187.

Humphreys, M. S., & Bain, J. D. (in press). Recognition memory: A cue and information analysis. *Memory and Cognition*.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 3*, 306–340.

Juola, J. F., Fischler, I., Wood, C. T., & Atkinson, R. C. (1971). Recognition time for information stored in long-term memory. *Perception & Psychophysics, 10*, 8–14.

Kintsch, W. (1968). Recognition and free recall of organized lists. *Journal of Experimental Psychology, 78*, 481–487.

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Eds.), *Models of human memory.* New York: Academic Press.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum.

Light, L. L., & Carter-Sobell, L. (1970). Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 1–11.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology, 12,* 114–135.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Macleod, C. M., & Nelson, T. O. (1976). A nonmonotonic lag function for false alarms to associates. *American Journal of Psychology, 89,* 127–135.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8,* 828–835.

Mandler, G. (1972). Organization and recognition. In E. Tulving & W. Donaldson (Eds.), *Organization of memory.* New York: Academic Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Mandler, G., & Boeck, W. (1974). Retrieval processes in recognition. *Memory and Cognition, 2,* 613–615.

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

Mandler, G., Pearlstone, Z., & Koopmans, H. J. (1969). Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 8,* 410–423.

Mandler, G., Rabinowitz, J. C., & Simon, R. A. (1981). Coordinate organization: The holistic representation of word pairs. *American Journal of Psychology, 94,* 209–222.

May, R. B., & Tryk, H. E. (1970). Word sequence, word frequency, and free recall. *Canadian Journal of Psychology, 24,* 299–304.

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

McNicol, D., & Stewart, G. W. (1980). Reaction time and the study of memory. In A. T. Welford (Ed.), *Reaction times.* London: Academic Press.

Medin, D. L., Busemeyer, J. R., & Dewey, G. I. (in press). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Metcalfe, J., & Murdock, B. B., Jr. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 20,* 161–189.

Murdock, B. B., Jr. (1962). The serial position effect in free recall. *Journal of Experimental Psychology, 64,* 482–488.

Murdock, B. B., Jr. (1974). *Human memory: Theory and data.* Potomac, MD: Erlbaum.

Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626.

Murdock, B. B., Jr., & Anderson, R. E. (1975). Encoding, storage and retrieval of item information. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium.* Hillsdale, NJ: Erlbaum.

Murdock, B. B., Jr., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology, 94,* 284–290.

Muter, P. (in press). Recognition and recall of words with a single meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Nairne, J. S. (1983). Associative processing during rote rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 3–20.

Neely, J. H., & Balota, D. A. (1981). Test expectancy and semantic organization effects in recall and recognition. *Memory and Cognition, 9,* 283–306.

Neely, J. H., Schmidt, S. R., & Roediger, H. L. III. (1983). Inhibition from related primes in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 196–211.

Nelson, D. L., & Davis, M. J. (1972). Transfer and false recognitions based on phonetic identities of words. *Journal of Experimental Psychology, 92,* 347–353.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in short-term memory. *Journal of Mathematical Psychology, 6,* 192–208.

Okada, R., & Burrows, D. (1973). Organizational factors in high-speed scanning. *Journal of Experimental Psychology, 101,* 77–81.

Pachella, R. G. (1974). An interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition.* Hillsdale, NJ: Erlbaum.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review, 80,* 53–68.

Pike, R., Dalgleish, L., & Wright, J. (1977). A multiple-observations model for response latency and the latencies of correct and incorrect responses in recognition memory. *Memory and Cognition, 5,* 580–589.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 14). New York: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981a). Order effects in recall. In A. Long & A. Baddeley (Eds.), *Attention and performance* (Vol. 9). Hillsdale, NJ: Erlbaum.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981b). Search of Associative Memory. *Psychological Review, 88,* 93–134.

Rabinowitz, J. C. (1978). Recognition retrieval processes: The function of category size. Unpublished doctoral dissertation, University of California, San Diego.

Rabinowitz, J. C., Mandler, G., & Barsalou, L. W. (1977). Recognition failure: Another case of retrieval failure. *Journal of Verbal Learning and Verbal Behavior, 16,* 639–663.

Rabinowitz, J. C., Mandler, G., & Patterson, K. E. (1977). Determinants of recognition and recall: Accessibility and generation. *Journal of Experimental Psychology: General, 106,* 302–329.

Rao, K. V. (1983). The word frequency effect in situational frequency estimation. *Journal of Experimental Psychology: Leaning, Memory, and Cognition, 9,* 73–81.

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

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

Reder, L. M., Anderson, J. R., & Bjork, R. A. (1974). A semantic interpretation of encoding specificity. *Journal of Experimental Psychology, 107,* 648–656.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science, 181,* 574–576.

Reed, A. V. (1976). List length and the time course of recognition in immediate memory. *Memory and Cognition, 4,* 16–30.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology, 92,* 365–372.

Rundus, D. (1973). Negative effects of using list items as recall cues. *Journal of Verbal Learning and Verbal Behavior, 12,* 43–50.

Salzburg, P. M. (1976). On the generality of encoding specificity. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 586–596.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science, 9,* 211–212.

Schulman, A. I. (1976). Memory for rare words previously rated for familiarity. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 301–307.

Schulman, A. I., & Lovelace, E. A. (1970) Recognition memory for words presented at a slow or rapid rate. *Psychonomic Science, 21,* 99–100.

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

Shiffrin, R. M. (1970). Memory search. In D. A. Norman (Ed.), *Models of human memory.* New York: Academic Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: 2. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6,* 342–353.

Strong, E. K., Jr. (1912). The effect of length of series upon recognition memory. *Psychological Review, 19,* 447–462.

Strong, E. K., Jr. (1913). The effect of time-interval upon recognition memory. *Psychological Review, 20,* 339–372.

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior, 1,* 443–450.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Columbia University Press.

Todres, A. K., & Watkins, M. J. (1981). A part-set cuing effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 91–99.

Townsend, J. T. (1971). A note on the identifiability of parallel and serial processes. *Perception & Psychophysics, 10,* 161–163.

Tulving, E. (1968). When is recall higher than recognition? *Psychonomic Science, 10,* 53–54.

Tulving, E. (1974). Cue-dependent forgetting. *American Scientist, 62,* 74–82.

Tulving, E. (1976). Ecphoric processes in recall and recognition. In J. Brown (Ed.), *Recall and Recognition.* London: Wiley.

Tulving, E. (1982a). *Elements of episodic memory.* Oxford: Oxford University Press.

Tulving, E. (1982b). Synergistic ecphory in recall and recognition. *Canadian Journal of Psychology, 36,* 130–147.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87,* 175–184.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373.

Tulving, E., & Watkins, M. J. (1973). Continuity between recall and recognition. *American Journal of Psychology, 86,* 739–748.

Tulving, E., & Watkins, M. J. (1975). Structure of memory traces. *Psychological Review, 82,* 261–275.

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive Psychology, 5,* 275–287.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review, 64,* 49–60.

Underwood, B. J., & Freund, J. S. (1968). Errors in recognition learning and retention. *Journal of Experimental Psychology, 78,* 55–63.

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

Walter, D. A. (1977). Discrete events in word encoding: The locus of elaboration. *American Journal of Psychology, 90,* 655–662.

Watkins, M. J., & Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General, 1,* 5–29.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85.

Wickelgren, W. A., & Corbett, A. T. (1977). Associative interference and retrieval dynamics in yes-no recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 189–202.

Winograd, E., & Conn, C. P. (1971). Evidence from recognition memory for specific encoding of unmodified homographs. *Journal of Verbal Learning and Verbal Behavior, 10,* 702–706.

Wiseman, S., & Tulving, E. (1975). A test of confusion theory of encoding specificity. *Journal of Verbal Learning and Verbal Behavior, 14,* 370–381.

Woodward, A. E., Bjork, R. A., & Jongeward, R. H., Jr. (1973). Recall and recognition as a function of primary rehearsal. *Journal of Verbal Learning and Verbal Behavior, 12,* 608–612.

Zacks, R. T., Hasher, L., & Sanft, H. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 106–116.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are U-shaped function of word-frequency. *Bulletin of the Psychonomic Society, 11,* 371–373.

(Appendixes follow on next page)

Appendix A

Table A-1
Means, Criteria, Hit Rates, and False Alarm Rates for the Simulations in Sections 2 and 3

Variable		Target mean	Distractor mean	Criterion	Hit rate	False alarm rate
		Figure 10				
Parameter						
a = 0.20		1.064	0.647	0.779	.942	.077
0.25		1.330	0.808	0.973	.942	.077
0.30		1.596	0.970	1.168	.942	.077
b = 0.10		1.033	0.808	0.838	.861	.385
0.20		1.330	0.808	0.973	.942	.077
0.30		1.627	0.808	1.108	.958	.007
c = 0.10		1.270	0.808	0.946	.937	.114
0.15		1.330	0.808	0.973	.942	.077
0.20		1.391	0.808	1.001	.945	.050
d = 0.050		1.145	0.539	0.766	.962	.003
0.075		1.330	0.808	0.973	.942	.077
0.100		1.515	1.078	1.180	.915	.244
v = 0.25		1.329	0.807	0.972	.997	.003
0.50		1.330	0.808	0.973	.942	.077
0.75		1.331	0.810	0.974	.830	.176
		Figure 12				
List length						
10		0.997	0.435	0.651	.916	.007
20		1.330	0.808	0.973	.942	.077
30		1.686	1.176	1.302	.954	.171
40		2.067	1.560	1.650	.965	.264
		Figure 13 (LL = 30)				
Time						
1 s		0.652	0.588	0.564	.864	.634
2 s		1.686	1.176	1.302	.954	.171
3 s		3.102	1.763	2.214	.975	.014
		Figure 14				
Parameter						
a = 0.23	c = 0.10	1.168	0.743	0.869	.937	.122
0.25	0.15	1.330	0.808	0.973	.942	.077
0.27	0.20	1.502	0.872	1.080	.943	.051
		Figure 15				
Parameter						
b = 0.10	c = 0.51	1.470	0.808	1.036	.848	.026
0.15	0.33	1.400	0.808	1.004	.913	.047
0.20	0.15	1.330	0.808	0.973	.942	.077
		Figure 16				
Parameter						
b = 0.20	c = 0.15	1.330	0.808	0.973	.942	.077
0.16	0.12	1.175	0.808	0.902	.921	.205
0.12	0.09	1.020	0.808	0.831	.874	.400

Table A-1 (*continued*)

Variable	Target mean	Distractor mean	Criterion	Hit rate	False alarm rate
		Figure 17			
Parameter					
$a = 0.25$	1.330	0.808	0.973	.942	.077
0.20	1.064	0.647	0.779	.942	.077
0.15	0.798	0.485	0.584	.942	.077
		Figure 18			
Parameter and junk					
$a = 0.25$; 0	1.330	0.808	0.973	.942	.077
0.20; 10	1.488	1.069	1.163	.951	.216
0.15; 20	1.639	1.325	1.348	.948	.424
		Figure 19			
Parameter and junk					
$a = 0.25$; 0	1.330	0.808	0.973	.942	.077
10	1.755	1.230	1.358	.952	.179
20	2.170	1.647	1.737	.956	.290
$a = 0.20$; 0	1.064	0.647	0.779	.942	.077
10	1.488	1.069	1.163	.951	.216
20	1.904	1.486	1.543	.954	.350
$a = 0.15$; 0	0.798	0.485	0.584	.942	.077
10	1.222	0.907	0.969	.947	.278
20	1.639	1.325	1.348	.948	.424
		Figure 20			
Test block					
1	1.595	1.011	1.186	.953	.081
2	1.979	1.385	1.531	.961	.157
3	2.348	1.760	1.869	.965	.242
4	2.726	2.133	2.211	.973	.314
		Figure 22			
Parameter					
Individual $d = 0.075$	1.330	0.808	0.973	.942	.077
0.100	1.330	0.822	0.979	.937	.089
0.125	1.330	0.835	0.985	.934	.104
Overall $d =$ 0.075	1.330	0.808	0.973	.942	.077
0.100	1.515	1.078	1.180	.915	.244
0.125	1.700	1.347	1.387	.878	.405
		Figure 23 Uniform lists (LL = 30)			
Parameters b and d					
$b_H = 0.2$; $d_H = 0.075$	1.686	1.176	1.302	.954	.171
$b_L = 0.1$; $d_L = 0.035$	0.897	0.549	0.658	.963	.043
		Figure 24 Mixed lists (LL = 30)			
Parameters b and d					
$b_H = 0.2$; $d_H = 0.075$	1.686	1.176	1.302	.954	.171
$b_L = 0.1$; $d_L = 0.035$	0.897	0.549	0.658	.963	.043

Note. Except as noted, the predictions shown are based on a list of 20 items presented at 2 s/item with the following parameters: $a = 0.25$; $b = 0.20$; $c = 0.15$; $d = 0.075$; $r = 4$; $v = 0.5$. In each case the criterion equals the sum of the target and distractor means multiplied by 0.455. LL = list length.

Appendix B

Method and Procedures for an Experiment on Word Frequency and Test Type

Subjects

Subjects were 80 undergraduates from Indiana University.

Materials

We selected from the Thorndike and Lorge (1944) general count, 520 HF (A or AA) and 520 LF (1–4 per million) words of between five and eight letters in length. All stimulus generation and presentation and response collection were controlled by a PDP 11/34 computer.

Design

The experiment consisted of 10 experimental trials, 2 trials of each type of memory test. Each trial contained a study list, an addition task (to clear the STS), and a memory test.

Each experimental study list consisted of 16 pairs of words. Four of each of the following combination of words formed the pairs: HF–HF, HF–LF, LF–HF, and LF–LF. The order of presentation of the pairs was randomized for each study list.

Two recall tests and three recognition tests were used. During the free-recall tests subjects were given a blank piece of paper and asked to write down all the words they could remember from the previous study list. They were asked to write down words together (on the same line) that had been presented together, if they were able to do so. For the cued-recall test one word from each pair was selected to be a cue. The subjects were asked to provide the paired member as a response. One half of the cues were HF words and the other half were LF words. One half of the HF cues were originally presented in the study list as the first member of a pair, and the other half were originally presented as the second member of a pair. The same selection was used for the LF words. The order of presentation of the cues was randomized, and the entire test list was printed on a sheet of paper for each subject.

The single-recognition test was composed of 32 words. One word from each of the study pairs was selected to form the 16 targets. Eight of the targets were HF words, eight were LF words, eight (four HF and four LF) words were originally presented as the first member of a pair, and eight were originally presented as the second member of a pair. In addition, eight HF words and eight LF words that were never presented to that subject as study items were selected from the master lists to be used as the 16 distractors. The ordering of the 32 words within the test list was randomized.

The paired-recognition test was composed of 32 pairs of words. Sixteen of the pairs were the original 16 target pairs. The remaining pairs were never presented items paired as distractors. The sixteen distractor pairs were of the same frequency combination as the targets. The ordering of the target and distractor pairs was randomized.

The cued-recognition test was composed of 16 pairs of items. One member of each pair was designated as a cue by placing an asterisk to the left of the first member of a pair or to the right of the second member of a pair. Each cue was selected from the study list in the same fashion that the cues were chosen in the cued-recall task. The 16 test items were made up of eight targets and eight distractors. Each target was paired with the cue that had been presented with the target on the study list. One half of the targets were HF and one half, LF. For each cue-distractor combination, a HF or a LF word that had never been presented during study was chosen at random to serve as the distractor.

Every group of four subjects received a different test order and a different ordering of words within a trial. Thus there were 20 different orders used in the experiment.

Between the presentation list and the test list, an arithmetic test was given. Each arithmetic test consisted of the mental addition of seven single-digit numbers, presented one at a time for 2 s each.

Procedure

Subjects were told that they would be presented with a series of words to study and that they would later be tested. They were told that the words would be presented in pairs and that they should try to associate the pairs together using a mnemonic of some type. Subjects then had the procedure for each type of test explained to them and were given practice trials for each type of test.

Each trial was initiated with the words GET READY printed on the screen for 2.5 s followed by 16 pairs of words presented for 4 s each. There were 100 ms of blank time between each presentation. Immediately after the last pair of words was presented, a series of single digits was presented to the screen, one digit at a time. Seven digits were presented for 2 s each. There were 100 ms of blank time between the end of one digit and the beginning of the next digit. After all digits were presented, WRITE THE

SUM ON THE ANSWER SHEET was presented on the screen. Subjects were then given 5 s to write their sum on an answer booklet. The type of test the subjects were about to receive was then indicated on the screen and remained there for 5 s. For the free- and cued-recall tests, subjects were given 1.5 min to complete the task.

For each of the recognition tasks the single word (for the single-recognition task) or the pair of words (for the paired-recognition and the cued-recognition tasks) was presented on the screen for 2.5 s. For each type of test the subjects were to respond yes on the keyboard if the test item had been studied and no if the test item had not been studied. Subjects were told to respond as quickly as possible without sacrificing accuracy, and response times were collected. One half of the subjects responded yes with their left hands and no with their right hands, whereas for the remaining subjects the order of responding was reversed.

Appendix C

Statistical Analyses of the Recall-Recognition Study

A series of analyses of variance (ANOVAS) was carried out that confirm the observations given in the text. A separate ANOVA was completed for each type of test. A 2 × 2 (Frequency, HF or LF, × Pair Type, same or different frequency) analysis was carried out on the number of items correctly recalled on the cued and free-recall results, separately. Neither the pair type nor the frequency main effects nor their interaction (in all cases $p > .1$) was significant. An additional ANOVA was done on the arcsin transformations of the proportion of words correctly recalled, combining the cued- and free-recall tests. The 2 × 2 × 2 (Frequency × Test × Pair Type) analyses revealed only one significant factor, namely test type. That is, cued-recall performance was better than free-recall performance.

A 2 × 2 (Frequency × Pair Type) ANOVA was completed for the single-recognition results and only the main effect of frequency was significant, $F(1, 79) = 100.53$, $p < .001$. A similar analysis for the cued-recognition test showed that again only the frequency effect was significant, $F(1, 79) = 15.52$, $p < .001$.

For the paired-recognition test a simple one-way ANOVA was completed with three levels of frequency (high, mixed, low). This was done because the HF–LF and LF–HF pairs are indistinguishable with regard to pair type. Therefore, they were combined into a single mixed-frequency point. The ANOVA again showed that the frequency effect was significant, $F(1, 79) = 15.43$, $p < .001$.

Matched-group t tests were carried out on the different types of tests to compare their overall level of performance. As expected, the paired-test performance was better than the single-test performance, $t\ (78) = 3.18$, $p < .005$. Because of the variability in the cued tests, the single-recognition performance was not significantly better than the cued-recognition performance ($t = .869$). We have chosen, however, to fit this difference in means because the results have a logical explanation in our model.

Received October 4, 1982
Revision received July 29, 1983 ■