Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.). *Rational models of cognition.* (pp. 73-95), Oxford, England: Oxford University Press.

4 *The effectiveness of retrieval from memory*

Richard M. Shiffrin and Mark Steyvers

Introduction

The term 'rational models' has connotations almost too numerous to list, including internal consistency of a model's assumptions, consistency of data with a model's predictions, matching of the model's predictions to the demands imposed by the environment and its probabilistic structure and pay-offs, and logical consistency of a model's decision structure. In this chapter we focus on 'optimality', in particular, optimality of retrieval from memory. In an absolute sense, we ask whether retrieval is as optimal, efficient, and effective as possible, given the way in which information has been stored in memory. This question may not be well defined or easy to answer. In a relative sense, we ask whether retrieval in different tasks and of different types of information operates with different levels of optimality, effectiveness, and efficiency. Such questions are of course highly model dependent, but by keeping the assumptions as simple as possible, it may be possible to obtain answers that generalize to quite a wide range of potential models. In addition, considerations of optimality may help point the way to consideration and adoption of certain classes of models. Our discussion will be couched in terms of a particularly simple and straightforward probabilistically based model called REM, standing for retrieving effectively from memory (Shiffrin and Steyvers, 1997). We will discuss a variety of memory paradigms including several types of episodic recognition, and episodic recall; we will also comment briefly upon access to generic memory, and the interaction of episodic and generic memory, usually termed implicit memory.

Questions of optimality depend upon a host of explicit and implicit assumptions that underlie a model, so definitive answers and general conclusions are unlikely. Our aim in this chapter is to raise some of the relevant issues, provide a few illustrations, and demonstrate that thinking about optimality can provide some insights that might otherwise be missed. We shall use the REM model as a basis for discussion, as this model was partly motivated by an attempt to think of simple, explicit, recognition memory as a Bayesian optimal decision, and because the probabilistic nature of the model makes consideration of optimality somewhat easier to codify, but our intent is not to support REM over other models (at least not in this chapter).

Retrieving effectively from memory: REM

To make our conclusions as transparent and generalizable as possible, we present a very simple version of the REM model. It is only in this case that the optimal Bayesian solution for explicit single-item recognition memory can be derived in a simple enough form that simulations are possible. More complex variants of REM with more plausible and realistic assumptions are presented in Shiffrin and Steyvers (1997), and it is shown there that the basic patterns of predictions hold up through the increasing complexities, when one applies the formulas that are optimal for the simplest case (even though they are no longer strictly optimal for the more complicated cases).

Representation and storage

Separate memory images are stored for different events. Each memory image is represented as a vector of feature values (including both content and context features); the values are positive integers, with the most environmentally probable values being the lowest integers. For convenience, assume that the distribution of feature values is geometric, as illustrated in Equation 4.1. V represents a feature value and g is a REM parameter. The number 0 also appears in certain positions of an image vector, and denotes no information stored.

$$P[V=j] = (1-g)^{j-1}g, \; j = 1, \ldots\ldots, \infty \qquad 4.1$$

It is convenient to divide memory images into two classes: very incomplete and error prone images representing recent events, called episodic, and relatively complete and accurate images representing accumulated knowledge, called lexical/semantic. Thus the lexical/semantic image for a presented word might look like: <3, 5, 3, 2, 1, 2, 3, ... 5, 1>. Upon presentation this image is retrieved and rehearsed and an incom-plete and error prone copy stored as an episodic image: <0, 0, 3, 0, 1, 1, 0, ... 4, 0>.

Note that only a few of the many feature values comprising an event actually get stored in an episodic image, and the values that do get stored may not always be correct. Let us assume there is some probability u^* of attempting to store a feature value for each unit of coding/rehearsal time. If an attempt is made to store, assume there is a probability c of copying the feature value correctly, and a probability $1-c$ of storing a value selected randomly according to the geometric distribution representing the environmental base-rates (Equation 4.1).

There are several rules governing which memory images receive the newly stored information:

1. When a new event occurs it may call to mind an already-stored lexical/semantic image; a typical example occurs when presentation of a known word causes contact with that word's lexical/semantic representation. Features of the event not already stored in the lexical/semantic vector may be stored there. As such vectors are relatively complete, not too much new information may be added, and what is added may be largely current context; none the less such additions to the lexical/semantic images are used in the theory to account for most implicit memory effects.
2. An episodic image may be retrieved that is extremely similar to the current event; in this case features will be added to the retrieved episodic image, and no new image is stored.
3. An episodic image may be retrieved that is similar to the current event, but distinguishably different; in this case current event information is stored both in the retrieved image and in a newly formed episodic image.
4. Finally, if no similar enough episodic image is retrieved, then current event information is stored in a new episodic image.

It is rather important to note that these rules allow the build-up of increasingly complete lexical/semantic images from a succession of episodic events over developmental time, simultaneously with the laying down of numerous separate incomplete episodic traces.

To tie these ideas to a particular memory paradigm, suppose a list of pairs of different words is studied. Assume that the lexical/semantic vectors representing different words have m feature values that are generated independently according to the geometric distribution given earlier (Equation 4.1). Let there be an (incomplete and error prone) episodic image stored for each pair, and let it be represented as a concatenated vector with the first half (m feature values) representing word 1 and the second half (m feature values) representing word 2. To keep things simple, assume that each different pair studied produces a different episodic image, but that repeated pairs within a list are stored in the same episodic image. Note that n pairs are therefore represented as $2n$ word vectors grouped by twos. In our simulations, we set $m = 20$.

Retrieval: explicit recognition of single words

Let us begin with single-word old–new recognition: half the test words are from the list and half are new. The simplest version of REM assumes that retrieval is as good as it can be given the storage constraints. That is, a Bayesian probability calculation is used to determine the probability that a test item is old. To be more precise, the lexical/semantic vector representing the test word is compared in parallel with each of the $2n$ word vectors in episodic memory for the studied list. Each comparison consists of a list of matching and mismatching feature values, the j-th such list being termed D_j, and the set of $2n$ D_j being termed the data, D. Under these assumptions, one can derive the odds (Φ) of the test item being old versus new, given the data, D. It equals the expression given in Equation 4.2, where the λ_j are likelihood ratios. λ_j is actually the probability of D_j, given that image j was produced by the word being tested (in which case it is termed an s-image), divided by the probability of D_j given that D_j was produced by presentation of some other word (in which case it is termed a d-image). Equation 4.3 gives one form of the expression for λ_j. In Equation 4.3, n_{jq} is the number of non-zero feature values in the j-th image that match the corresponding value in the test word, and n_{ijm} is the number of non-zero feature

values in the j-th image that have value i and mismatch the corresponding value in the test word.

$$\Phi = \frac{1}{n}\sum_{j=1}^{n}\lambda_j \qquad 4.2$$

$$\lambda_j = (1-c)^{n_{jq}}\prod_{i=1}^{\infty}\left[\frac{c+(1-c)g(1-g)^{i-1}}{g(1-g)^{i-1}}\right]^{n_{ijm}} \qquad 4.3$$

In the absence of differential pay-offs, the optimal decision rule is to respond old if the odds of 'old' are greater than 1.0. This assumption produces a two-parameter model for single-item recognition, based on the probability of error, c, and the parameter of the geometric, g. Shiffrin and Steyvers (1997) produced qualitative predictions for a variety of standard phenomena in recognition memory, as shown in Fig. 4.1. The predictions were based on: $c = 0.7$; $g = 0.4$ (used in Equation 3 to

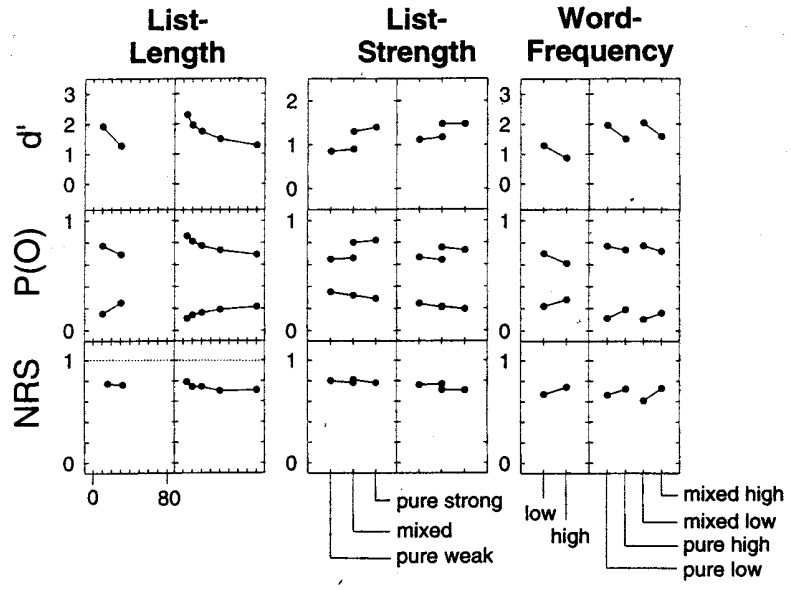

Fig. 4.1 Selected data from the literature (in the left-hand column of each set of two columns; citations are given in Shiffrin and Steyvers, 1997), and predictions of the REM model described in the text (right-hand column of each set of two columns). Left columns: variations in list length. Middle columns: variations in strength of words (top right-hand two points are stronger) and strength of other list items (right-hand point in each connected group of two is the case of stronger other words). Right columns: variations in word frequency, in lists of pure and mixed frequency, as labeled. Top panels: performance measured as d'; middle panels: hit rates ($P(old|old)$) and false alarm rates ($P(old|new)$); bottom panels: slope of the linear fit to the receiver operating characteristic (ROC) plotted on normal–normal axes.

calculate odds); $g_H = 0.45$ (used to generate high-frequency lexical/semantic images); $g_L = 0.325$ (used to generate low-frequency lexical/semantic images); $u* = 0.04$ per storage attempt; $t = 10$ storage attempts for strong words, and $t = 7$ storage attempts for weak words. Figure 4.1 demonstrates that this model correctly predicts the qualitative patterns for d', hit and false alarm rates (including the symmetric changes in these across conditions, termed the 'mirror effect'), and the slope of the normal ROC functions (labelled NRS in the figure, for the variables of list length, strength (e.g. study time or repetitions), list strength (the strength of other list words than the test word), and word frequency. (Details may be found in Shiffrin and Steyvers, 1997.) It is fairly remarkable that a normatively derived model based on two parameters captures the major trends and findings in recognition memory, including some results that have proved troublesome for almost all extant models.

Optimality: effectiveness of retrieval

These results are based on a normative calculation of probabilities, and represent optimal retrieval: given what was assumed about storage, all the information in memory is used to calculate the odds that the test word is old. This is important to note in light of the many theorists who propose recognition is carried out by a mixture of global familiarity processes (like the present REM model), and recall-like processes. Such mixture models have a long history and recently they have seen prominence in theories based on Jacoby's process-dissociation techniques (e.g. Jacoby, 1991); these mixture models have prompted procedures in which subjects are asked to classify their recognition judgements into two classes: know judgements' (presumably corresponding to generalized feelings of familiarity) and 'remember judgements' (presumably corresponding to recall of specific episodic events). It is an implicit assumption in most of this research that a judgement based on recall, when available, is superior to one based on familiarity, an assumption making it seem like a truism that the addition of recall to a global familiarity process will improve performance. Suppose for example that recall occurs in parallel with the familiarity calculation (e.g. familiarity in REM is the odds): the recall part consists of sampling a single image from those stored, and examining the contents; if the sampled trace is judged to contain the test word then an old decision is made without consulting the odds calculation. Although it seems plausible at first glance that such a procedure would improve performance beyond that achievable with a global familiarity process alone, this reasoning is incorrect in the case of the REM model. In REM the odds calculation uses all the information in memory—it is as if the decision maker is given a card with all the stored vectors, and given as much time as needed to utilize these images to make the best decision. Another way to say this is that the odds calculation in effect is based on sampling and recalling every trace, and using correctly all the information found.

This observation may help explain why a single-process global familiarity model like REM (and others) have fared so well in predicting recognition data, despite subjective impressions and a variety of other results and analyses that suggest recall occurs on some trials. In the example of the previous paragraph, one image is

sampled and recalled in parallel with the odds calculation in REM, and when judged to contain the test word, is used to give an 'old' judgement without reference to the odds. We know from the above reasoning that this procedure can only harm performance, in comparison with using the odds only. However, as such recalls will tend to occur in cases where the image in question is quite strong (cases in which many accurately stored features are in the episodic image), the odds calculation on that same trial would also have led to an 'old' decision, with a high probability. Thus the predictions for the single process and dual process models would be correlated to a very high degree, and the single process model would give accurate predictions even if the dual process model is correct.

A somewhat different question of optimality concerns storage error versus retrieval error. In REM, there are two kinds of storage 'error': failure to store a feature (incompleteness) and incorrect storage of a feature. In REM it seems appropriate to localize incompleteness in the storage process because the amount and type of coding and rehearsal are the primary determinants of performance, a result that would not occur were storage complete. In models differing from REM, however, feature storage might occur for all features, but with different degrees of strength; then retrieval of a feature value might occur on only some trials, depending on that strength. In such models, the 'incompleteness' seems to be shared between storage and retrieval. Turning to error in feature assignments next, we note that REM assumes these errors occur during storage (governed by the 'c' parameter). However, the REM theory would be mathematically identical if the error occurred during retrieval instead. It is only if errors and incompleteness are assigned to storage that it can be said that retrieval is optimal. It is hard to find any convincing reason to prefer the assumption that the errors occur in storage, but there is one rather weak line of reasoning that led us to make the assumptions we did: if the errors occur during retrieval, then it is hard to find a reason why these would not be at least partially random over successful retrieval attempts. If errors are randomized over successive retrieval attempts, then the law of large numbers will insure that performance can be made to improve (to whatever are the limits imposed by storage) by accumulating evidence over multiple retrieval attempts on a trial. This reasoning led us to place the error in storage, but either model could probably be defended.

Another question of optimality concerns the set of simplifying assumptions that had to be made to allow the REM model to be derived. For example, almost any deviation from the simple assumptions we made for the simplest REM model (those listed earlier), and applications to almost any task more complicated than simple recognition, greatly complicate the form of the Bayesian solution: the optimal odds calculation is no longer based on the likelihood ratios for individual images, as in Equations 4.2 and 4.3, but generally turns out to be a division of two different sums, each containing an astronomically large number of differing terms. The number of terms is so large that it is not feasible to simulate the predictions even with the fastest available computers. For the case of single word recognition, we have explored a variety of ways to relax the assumptions needed to derive the simple form of the Bayesian solution, and allowed more realistic task assumptions to be made (e.g. allowing occasional separate storage of repetitions, images from words not on the list to be in memory, and context features to be part of the representations). We were

able to show that the use of the derived optimal calculations (Equations 4.2 and 4.3) produces predictions virtually indistinguishable from those for the simplest case (Shiffrin and Steyvers, 1997). These findings lend some robustness to the model.

At the end of this chapter, we shall discuss a final example demonstrating the effectiveness of an approximation to an optimal solution for recognition decisions. The approximation bases the recognition decision on the maximum of the likelihood ratios across the $2n$ images (as suggested by McClelland and Chappell, in press), rather than the sum of the likelihood ratios that is required by an optimal solution. Discussion of this case is deferred because one important implication of the result is a potential application to cued recall.

Relative effectiveness of retrieval across tasks

What we have been discussing so far concerns what might be termed absolute optimality, something a real system might not be expected to achieve. What is in many respects more enlightening, and perhaps having more important implications, is relative optimality: If we move from one experimental condition or task to another, can we say something about whether the retrieval in one case uses the information in memory as effectively as in another? Of course, to look at this question with empirical data, it is essential that the study conditions and instructions are identical across the conditions of interest, so that the information in memory prior to retrieval is identical across conditions being compared. We will discuss conditions where this empirical proviso holds true. We start by considering recognition situations in which more than one word is tested, and ask how these compare with each other, to single-item recognition, and to cued recall.

Nobel's (1996) study of multiple word recognition

There are a number of studies in which groups of words are studied without foreknowledge of the upcoming test, and then followed by a variety of single- and multiple-item recognition tests. For example, Clark and Shiffrin (1987) carried out such a study with word triples, followed by all combinations of single, double and triple word tests, under three different instructional conditions. For present purposes we will discuss instead recent studies by Peter Nobel (Nobel, 1996), using a signal-to-respond procedure. Twenty word pairs (AB, CD, EF, etc.) were presented for study, without the subject knowing what sort of test would follow. Four kinds of test blocks were used.

1. Single-word recognition, denoted (A versus X). One old word or one new word is presented and the subject judges old versus new.
2. Paired recognition, denoted (AB versus XY). Two words are presented, both old or both new, and the subject judges old or new.
3. Associative recognition, denoted (AB versus CF). Two words are presented for test. Either both words are old and had been studied together, or both words are old and had been studied in different pairs. The subject judges which is the case.
4. Cued recall, denoted (A–?). One word is presented and the subject tried to generate the other member of the studied pair.

In each of these four conditions, subjects withheld a response until a (variably delayed) signal occurred and then had to respond within a very short period of time (several hundred milliseconds).

The curves giving the growth of accuracy with time are given in Fig. 4.2: note that these rise to an asymptotic level reflecting the maximum attainable level of performance in each condition. It may be noted that the approach to asymptote is slower for associative recognition and for cued recall than for paired and single word

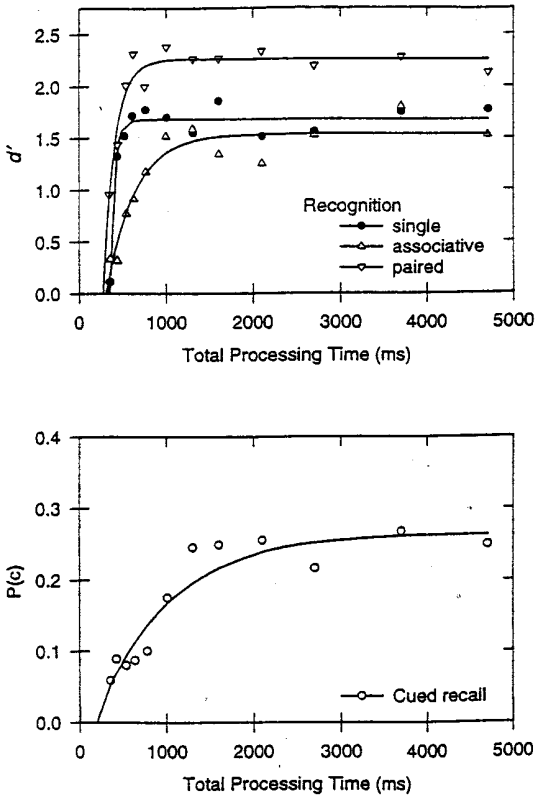

Fig. 4.2 Results from signal-to-respond conditions from Nobel (1996). Top panel: performance (d') for three recognition conditions as a function of the sum of the signal delay plus response time. Bottom panel: performance (probability correct) for cued recall. Solid lines are three parameter exponential functions fit to the data, governed by an intercept (I), a growth rate (G), and an asymptote. For recognition the asymptotic d' values are given in the first three rows of Table 4.1. Joint confidence intervals for I and G are given in Fig. 4.3.

recognition (also see Fig. 4.3). The asymptotic levels of d' for the three recognition conditions are given in the first three rows of Table 4.1: note that paired recognition is better than single-item recognition, but not hugely so, and that associative recognition is almost at the same level as single-item recognition.

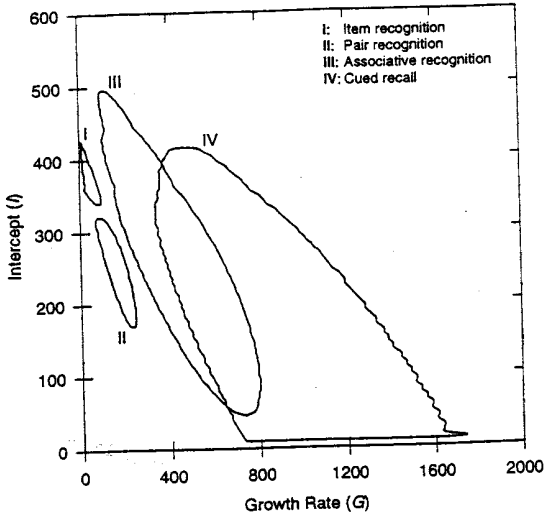

Fig. 4.3 Contours of the 95% joint confidence regions for the values of intercept (I) and growth rate (G), for the exponential functions fit to the signal-to-respond data of Nobel (1996). The data and exponential functions are shown in Fig. 4.2.

Table 4.1 Comparison of experimental data of Nobel and Shiffrin (1996) and the REM simulations

Condition	Hit rate	False alarm rate	d'
Data			
(1) Single	0.78	0.21	1.69
(2) Paired	0.75	0.06	2.2
(3) Associative	0.75	0.21	1.62
Prediction			
(4) Single	0.77	0.16	1.74
(5) Paired	0.91	0.07	2.81
(6) Paired ($\beta = 0.775$)	0.837	0.105	2.234
(7) Paired (product rule)	0.84	0.12	2.18
Associative Model; pseudo-optimal			
(8) Associative	0.51	0.008	2.44
(9) Associative ($C = \exp[-4]$)	0.86	0.12	2.2
(10) Associative ($\beta = 0.775$)	0.37	0.009	2.05
(11) Associative ($\beta = 0.775$; $C = \exp[-4]$)	0.8	0.2	1.67
Associative Model; paired methodology			
(12) Associative ($\beta = 0.775$)	0.84	0.5	0.99
(13) Associative ($\beta = 0.775$; $C = \exp[2]$)	0.65	0.23	1.14
(14) Associative	0.91	0.5	1.32
(15) Associative ($C = \exp[2]$)	0.78	0.27	1.4

Retrieval effectiveness in paired recognition

What do these results say about relative optimality? We can use REM to compare these findings at asymptote. Consider first the case of paired recognition. To carry out paired recognition in optimal fashion, one compares the joint probe vector, consisting of both test words, with each stored joint vector, in both possible orders. A likelihood ratio is calculated in the usual way for each of the $2n$ images. The $2n$ likelihood ratios are averaged to obtain the odds, and an old decision is made if the odds are greater than 1.0. We first choose the parameter which come close to predicting Nobel's recognition data for single word recognition; it turns out that the parameters used to generate the predictions of Fig. 4.1 provide a reasonable fit, so these were simply carried over. The predictions are given in row 4 of Table 4.1. For the same studied list, one can produce predictions for the paired recognition case using these same parameter values; the corresponding predictions are given in row 5.

What we see by comparing rows 1 and 2 with 4 and 5 is that predicted paired recognition performance is too good relative to single-item recognition. Apparently, the subjects are not retrieving as effectively in the paired case. Why not? It is conceivable that there is an error in the assumption that both orders of the test words are compared, but an analysis of the data showed no difference between test pairs that matched the study order and those that did not, making such an explanation unlikely. Most likely in our view is the hypothesis that retrieval capacity is limited in the capacity to utilize multiple cues in a single simultaneous probe of memory; an idea featured prominently in the SAM (search of associative memory) model (Raaijmakers and Shiffrin, 1980, 1981). The REM model that produces the predictions in row 5 of Table 4.1 assumes that all the features of both words are used in the probe of memory. Perhaps not all these features can fit in the probe. We therefore tried varying the proportion of features that make up a probe—each feature in the vector representing the test pair was allowed to join the probe with an independent probability β. Row 6 of Table 4.1 shows that when $\beta = 0.775$, predicted d' for paired recognition drops approximately to the observed level.

Other types of suboptimal retrieval are of course possible. One of these models would have the subject calculate the familiarity of each word separately, based on all the features of each word, and then combine these. For example, if the subject makes a separate odds calculation about the 'oldness' of each word, a reasonable strategy is to multiply these, and respond old if the product is greater than 1.0. As shown in row 7 of Table 4.1, this model also produces predictions for paired recognition that are roughly aligned with the data.

Choosing between these models is a delicate matter. We note that the growth rate for paired recognition is if anything more rapid than single-item recognition, as illustrated in Figs. 4.2 and 4.3, suggesting that if familiarity is calculated separately for both words, these calculations are carried out in parallel. Some researchers might therefore prefer a model with a joint probe but with a limit on the features in the probe. On the other hand, if features join the probe with probability β, then it may be plausible that different features join the probe on different retrieval attempts. If so, multiple retrieval attempts will enable an accumulation of evidence that would eventually produce performance equaling that obtainable with a complete set of features in a joint probe, a result contrary to the data.

Thus there do not seem compelling reasons as yet to prefer one of these models over the other. None the less, the conclusion that retrieval is less effective in paired recognition than in single-item recognition is probably well founded, and we suspect would hold true in many model frameworks. This conclusion is bolstered by the following observation: it is easy to imagine reasons why paired recognition would be even better than that predicted (e.g. testing a pair might allow configural/relational features to join the probe, features not available for single word tests), but then an even greater limitation of retrieval than we have assumed would be required to fit the data.

Retrieval effectiveness in associative recognition

Given that there may be a limitation of capacity in combining two words in a single probe, it seems best to compare associative recognition with paired recognition, as both might be expected to share in the same limitation of capacity when constructing a probe. In the case of associative recognition it is no longer possible to consider a strategy involving the calculation of separate odds for each test word, as both test words have been studied, and hence will be equally familiar. It would be desirable to start with the optimal Bayesian solution for associative recognition, but this cannot be simulated in real time.

Although a strictly optimal Bayesian solution for the associative case is not computationally feasible, an approximation suffices to make the points necessary for this section: under the assumption that an intact pair is being tested, one can find the assignment of one target image and $n-1$ distractor images for which the likelihood of the observed data (the matching and mismatching features in all the images) is maximized. Similarly, under the assumption of a rearranged test, one can find the assignment of two partially matching images, and $n-2$ mismatching images, for which the likelihood of the data is maximized. In practice, what is done to implement this idea follows: the single half-image that best matches either of the test words is termed I1, and the test word it matches is termed T1. The other half of the image containing I1 is matched to the other test word; that is, I2 is matched to T2. A likelihood ratio based on this match is then calculated according to Equation 4.3. Term this ratio λ_1. If λ_1 is high, there is evidence that the test pair is 'intact'. Next test word T2 is matched to all the remaining double word images, excluding only the one that produced the best match, and the best matching half image is termed J2. A likelihood ratio is calculated according to Equation 4.3 based on the match between J2 and T2. Term this ratio λ_2. If λ_2 is high there is evidence that the test pair is 'rearranged'. Thus the decision is based on the ratio of these two likelihood ratios (i.e. on λ_1/λ_2). It is important to keep in mind that this model is clearly suboptimal, and the true optimal model would produce predictions of even higher performance.

The predictions for this pseudo-optimal model are given in row 8 of Table 4.1. As can be seen, performance is predicted to be better than that observed. As the model predictions for hits and false alarms are not centred, the criterion for responding 'intact' was lowered from 1.0 to exp[−4]. The resulting predictions are given in row 9, and are still too high. Thus it seems advisable to look at models for associative recognition that include even less effective retrieval. One such model is based on the assumption that both words are used together in the probe, and that not all the features can fit in the probe. We therefore included features with probability β, using

the same value for β that fit the paired data, but otherwise used the pseudo-optimal model just described. The predictions are given in row 10, and are still too high. Finally, the criterion for this version of the model was lowered to exp[−4], giving rise to the predictions shown in row 11. Although performance for this version is now almost as low as the observed data, it must be remembered that, even ignoring the use of $\beta = 0.775$, the basic approximation we used is less than optimal. Thus, no matter how one looks at this matter, if we assume that subject's effectiveness of retrieval is adequately measured by the REM model applied to these tasks, subjects seem to be retrieving less effectively in associative recognition than in pair recognition (and less effectively in pair recognition than in single-item recognition).

One account of this finding would hold that the recognition system is only capable of calculating basic 'familiarity' for words or groups of words, and not capable of carrying out the subtle sorts of analyses required to approach optimality for associative recognition. With this idea in mind, we tried the following suboptimal model. The subject was assumed to use the same calculation used in paired recognition: a single probe is formed for a pair, with each feature included with probability β, and odds are calculated just as for paired recognition. Of course, as both words in rearranged words are familiar, the use of a criterion of 1.0 for a decision will tend to produce unacceptably high false alarm rates, as illustrated in row 12 of Table 4.1. We therefore adjusted the criterion to exp[2], approximately at the point where the distributions of odds for targets and distractors cross, and generated the predictions given in row 13. Both these sets of predictions are too low, so we tried generating predictions for such a model without the limitation of capacity (with $\beta = 1.0$), and achieved the predictions given in row 14; the predictions in row 15 result from raising the criterion to exp[2]. These are still too low relative to the observed performance. Thus it seems clear that a more effective retrieval strategy is employed than that copied from paired recognition: assessing familiarity of the pair of test items is not sufficient to carry out associative recognition.

Given that another model is clearly needed, we turned to an approach based on that utilized by Nobel (1996). This approach assumes that associative recognition utilizes an extended search process with recall-like components, an assumption motivated in part by results on the speed of retrieval. Figure 4.3 gives confidence regions for the intercept and rate of growth parameters for the various conditions of Nobel's (1996) study (the rate of growth of the curves in Fig. 4.2). The main thing to note is that the retrieval dynamics are very similar, and rapid, for single-item recognition and paired recognition; they are very much slower for associative recognition and cued recall (which are not statistically distinguishable).

Nobel's (1996) version of a recall-based model for associative recognition worked quite well, and, although embedded in the framework of the SAM model of Raaijmakers and Shiffrin (1980, 1981), ought to prove easy to implement in the REM framework. The idea is to rely on the sequential sampling of images, using for each attempt a probe cue consisting of one of the two words in the test pair. An image is sampled in proportion to its strength to the probe word. The search can stop with the sampling of a target image for which both halves seem to match the probe (in which case the search stops and an 'old' response is given), or the sampling of an image only one of whose halves seems to match the probe (in which the search stops

and a 'new' response is given). Because each probe uses all the features of just one word, a capacity limitation on the joint use of both words in a single probe is overcome, but at the cost of extra time used in retrieval (time used for the sequential search). Qualitatively, at least these anticipated predictions match the patterns of observed data, but a real test must await quantitative fits.

It is worth noting that the consideration of retrieval time considerably complicates analyses of optimality. In most studies subjects are either required to respond quickly (as in signal-to-respond studies) or given ambiguous instructions to respond 'as accurately and quickly as possible'. In both cases optimality is therefore a matter of joint minimization of retrieval time and maximization of accuracy. However, especially in light of the well known fact that subjects can trade improvements in one for decrements in the other, there are no generally accepted metrics for simultaneous optimization of both time and accuracy. Note that it may not help matters to instruct subjects to put all their 'weight' on accuracy. Subjects have a high resistance to the allocation of extra time and effort to retrieval, especially when the marginal gains are modest or worse. Thus, regardless of instructions subjects will face a conflict between situational demands for high accuracy and rapid responding.

In summary, our optimality analyses suggest retrieval in paired recognition suffers relative to single-item recognition due to limited retrieval capacity for multiple word probes. Our optimality analyses also suggest retrieval in associative recognition is bounded between the optimal level and the level available from the use of a paired-recognition retrieval strategy. We suggest subjects carrying out associative recognition may use single word probes in an extended process of recall involving sampling and recovery. Although we have not yet implemented this model in the REM framework, it has the potential of predicting both the observed levels of accuracy and the slow time course of retrieval.

Retrieval effectiveness in cued recall

In cued recall, one member of a studied pair is provided, and the other must be generated. Presumably, the process must begin with retrieval of stored episodic information and then generation of a response requires in addition access to the information in the word lexicon. In some models of the first stage of this process (e.g. SAM and REM), access to the specific episodic information in a given image is needed, rather than access to a global composite of the episodic information in all images (as in most models of recognition). The extra complications associated with cued recall tasks make defining and assessing optimality quite difficult. Given full information concerning the stored images, and full information concerning a lexicon of possible responses, one could perhaps work out an optimal decision, but it is clear that subjects do not access and use this information in optimal fashion. Consider for example just one of the processes needed in cued recall, access to the lexicon. We know that subjects given a word fragment completion task with a unique completion (e.g. what word has the form: _a_v_n_?) do not always find the correct completion, although optimal retrieval would produce perfect performance. Further, we know that access is far from optimal in the sense of retrieval time: Nobel's research, described earlier, makes it clear that cued recall is carried out in quite different fashion than single-item or paired recognition recall: it has a distribution that is far more extended over time.

Faced with non-optimal retrieval, we have proposed a search process borrowed from Raaijmakers and Shiffrin (1980, 1981) in which images are selected and examined successively, the subject balancing demands of accuracy and response time in order to decide when to stop the search. The process we have in mind involves choosing a pair image, examining the contents of the selected vector, deciding whether the vector is the one encoding the test item, trying to determine what word is encoded in the other part of the vector (which requires retrieval from the lexicon), and deciding whether to emit a response, or continue searching. If the search continues, another selection of an image is made, and so on. The general determination of optimality in such a situation is beyond our capability, but some interesting and useful results can be obtained, especially concerning sequential selection.

For our simple version of REM, it can be shown that the probability that image j is the one containing the test word, termed P_j, is just the sum of the two likelihood ratios for the two parts of the image divided by the sum of all $2n$ likelihood ratios, as illustrated in Equation 4.4. Thus, it is clearly optimal to consider first the image with the highest sum of likelihood ratios. It also seems plausible that it is optimal to consider additional images in descending order of their P_j values.

$$P_j = \frac{\lambda_{j1} + \lambda_{j2}}{\sum_k (\lambda_{k1} + \lambda_{k2})} \qquad 4.4$$

Is it plausible that images are examined in strict order of their likelihood ratios? Table 4.2 gives the distribution for the proportion of instances in which the image actually containing the target will have the highest likelihood ratio, the second highest, etc. (for the parameter values we have been using throughout). For a 10-item list, the probability that the target will have the highest λ is 0.89, and the probability of being in the top two λs is 0.96. For a 20-item list these probabilities are 0.83 and 0.91. Thus, if subjects did sample in order of likelihood ratios, there would be little point in searching past the first few samples. In addition, even if search continues until an image is encountered that is judged to contain the test word, and then stops, this critical event will tend to occur very early in the search. (This conclusion applies as well to any monotonic transformation of the likelihood ratios, such as a log or root, as the ordering of λs is not changed by such a transformation.)

Table 4.2 Distribution of the rank of the likelihood ratio for the target image

Ordinal position	Length=10	Length=20	Length=100
1	0.89	0.833	0.66
2	0.069	0.08	0.106
3	0.022	0.035	0.051
4	0.011	0.018	0.032
5	0.003	0.013	0.022

Consequently, predicted search time will peak early and fall off sharply, and it would probably be hard to predict the kind of extended search indicated by Figs 4.2 and 4.3. (It would also be hard to predict the free response times skewed toward slow responses that are also observed in Nobel, 1996.) These observations can be added to those arising from a consideration of inter-response times in free recall (e.g. Raaijmakers and Shiffrin, 1980, 1981), suggesting that successive samples occur with replacement; that is, images already examined can be sampled again during the same search process. Thus, in cued recall successive samples probably do not occur in strict order of likelihood ratios, and images once sampled can probably be sampled again during the same search. Both factors will tend to flatten the distribution of time until the search reaches the target image.

Therefore, let us assume a retrieval system in which an image is selected in proportion to its likelihood ratio; i.e. the probability of sampling a given image is given by the term on the right side of Equation 4.4. Furthermore, assume that successive samples use the same sampling rule, so that re-sampling of the same image sometimes occurs. Will such a system produce reasonable predictions? Table 4.3 gives the distribution of the number of samples it takes to first sample the target image for this system (for a list of length 20, and the usual parameter values). There remains a marked tendency to sample the correct image very early in the search. Most of this tendency is due to the extreme skewing of the likelihood ratios toward large values: there is a high probability that a target will have an extremely high likelihood ratio, and hence will be sampled immediately. The skewing is so pronounced that the distributions of the raw likelihood ratios do not lend themselves to graphing. Thus we give in Fig. 4.4 the distributions of log λ_j, for s-images (those that match) and d-images (those that don't match). Table 4.4 illustrates the large likelihood ratios that exist in these distributions by giving the mean value of the n-th largest likelihood ratios, when the n-th largest is a target or a distractor.

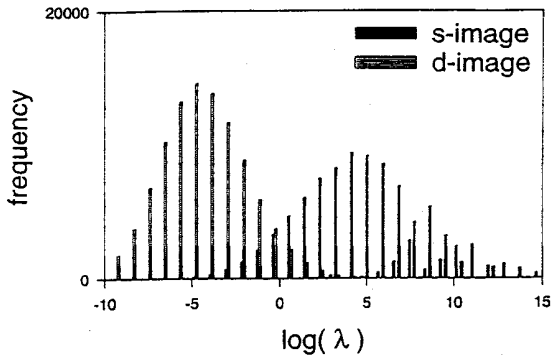

Fig. 4.4 Distribution of the natural logarithm of the likelihood ratio (λ) for an s-image (the image stored when the test word had previously been studied, illustrated with filled bars) and for d-images (images of other studied words; illustrated with open bars).

If, as these results suggest, proportional sampling of the raw likelihood ratios fails to spread out sampling times sufficiently, it may prove helpful to assume that the likelihoods are compressed (by a fractional power or a log, say) before sampling.

Equation 4.5 illustrates this: it gives the probability of sampling an image when sampling is proportional to a function, f, of the likelihood ratios.

$$P_s(i) = \frac{f(\lambda_{j1}) + f(\lambda_{j2})}{\sum_j (f(\lambda_{j1}) + f(\lambda_{j2}))} \qquad 4.5$$

Table 4.3 Distribution of number of samples to reach the target, for proportional sampling with replacement

No. samples	Length=10	Length=20	Length=100
1	0.837	0.75	0.572
2	0.06	0.08	0.097
3	0.023	0.034	0.049
4	0.012	0.022	0.028
5	0.01	0.012	0.023

Table 4.4 Mean value of n-th largest likelihood ratios

	Target	Distractor
1	354553	16.2
2	9.3	2.03
3	1.6	0.75
4	0.6	0.39
5	0.34	0.23

The more compression is produced by f, the less will be the tendency to select first the images with the highest likelihood ratios. However, there is a conceptual problem with this approach. The entities on which sampling is based, the $f(\lambda)$, are different than the entities on which recognition likelihood is calculated, the λ. This state of affairs would be unappealing to many theorists. There is a solution to this problem that depends upon using an alternative to Equation 4.2 for recognition decisions, an alternative based on using the largest of the likelihood ratios rather than the sum. We therefore return briefly to the topic of recognition.

Recognition decisions based on largest likelihood ratios

The proposal that recognition decisions could be based on the largest of the likelihood ratios is due to McClelland and Chappell (in press) who have a model very similar to REM. They proposed further that sensitivity to list length be incorporated by dividing by n. For single word recognition the odds that an old item has been tested becomes:

$$\Phi^* = \frac{1}{n} MAX_j[\lambda_j] \qquad 4.6$$

Although this model is not optimal, there are reasons to think it may be a good approximation to optimality. To see whether this is the case, we generated predictions for the conditions of Fig. 4.1, using the same parameter values, simply replacing Equation 4.2 (the average model) by Equation 4.6 (the maximum model). Both sets of predictions are given in Table 4.5. An examination of Table 4.5 shows that the MAX model produces d' values uniformly lower than the SUM model, a result expected because the MAX model is suboptimal. On the other hand, the discrepancies are small in magnitude, and there is a surprising degree of similarity between the predictions of the two models. To a good degree of approximation, then, the use of the largest single likelihood (scaled by n so that the criterion remains at odds of 1.0) produces the same predictions as the optimal model.

Table 4.5. Results of REM simulations for list length, strength, list strength and word frequency

		Hit rate	False alarm rate	d'	NRS
List length					
REM (sum)					
	4	0.827	0.143	2.008	0.74
	10	0.769	0.162	1.722	0.716
	20	0.735	0.183	1.532	0.709
	40	0.696	0.212	1.314	0.696
REM (maximum)					
	4	0.871	0.213	1.926	0.745
	10	0.813	0.225	1.644	0.715
	20	0.767	0.23	1.468	0.695
	40	0.704	0.232	1.268	0.674
List strength					
REM (sum)					
	Pure strong	0.731	0.193	1.482	0.702
	Mixed strong	0.749	0.209	1.48	0.683
	Mixed weak	1.482	0.214	1.134	0.773
	Pure weak	0.702	0.241	1.148	0.759
REM (maximum)					
	Pure strong	0.764	0.242	1.416	0.697
	Mixed strong	0.769	0.248	1.418	0.666
	Mixed weak	0.662	0.244	1.11	0.771
	Pure weak	0.674	0.254	1.114	0.73
Word frequency					
REM (sum)					
	Pure high	0.731	0.186	1.508	0.714
	Pure low	0.771	0.108	1.984	0.63
	Mixed high	0.704	0.152	1.564	0.711
	Mixed low	0.767	0.107	1.972	0.615
REM (maximum)					
	Pure high	0.763	0.233	1.448	0.708
	Pure low	0.812	0.158	1.892	0.627
	Mixed high	0.743	0.21	1.46	0.709
	Mixed low	0.809	0.159	1.874	0.624

In retrospect, this apparently surprising result is not hard to understand. It is due to the extreme skewing of the likelihood ratios toward high values that is seen in Fig. 4.4 and Table 4.4. Note that Equation 4.2 is equivalent to saying 'old' if the SUM is greater than n, and Equation 4.5 is equivalent to saying 'old' if the MAX is greater than n. Thus, the equivalence of the model predictions implies that the sum of the likelihood ratios is approximately equal to the maximum likelihood ratio, especially with regard to the number of times that either is greater than n. This is in fact the case, due to the tendency for the distribution of likelihood ratios to be dominated by a single enormous quantity. Furthermore, the times when this approximation tends to break down occurs when there is not a single large likelihood ratio, but these tend to be cases when both the SUM and the MAX values are less than n, and hence both lead to a 'new' decision. It is a curious fact that this result makes the recognition models of McClelland and Chappell (in press) and Shiffrin and Steyvers (1997) extremely similar in both structure and parameterization, to an even greater degree than these sets of authors may have appreciated heretofore.

The near equivalence of the MAX and SUM rules could be important for several reasons. For one thing, it may be noted that any monotonic function of the likelihood ratios, in particular a compressive function like a log or fractional power, will leave unchanged the image with the largest likelihood ratio. Thus, the decision based on Equation 4.6, to say 'old' when Φ^* is greater than 1, is identical to that produced by a decision based on Equation 4.7, to say 'old' when Φ_f is greater than 1:

$$\Phi_f = \frac{1}{f(n)} MAX_j[f(\lambda_j)] \qquad 4.7$$

The reason is clear: as the image that is the greatest is unchanged, all that is necessary is that whenever λ_j is greater than n, $f(\lambda_j)$ is greater than $f(n)$, which is true when f is monotonic. In particular, this would be true for a log or power function.

To summarize, the MAX model based on Equation 4.6 (that is shown in Table 4.5 to approximate the optimal REM model) is identical to a MAX model based on any monotonic function of the likelihood ratio (as long as a similar function is applied to the value n). Thus, the units upon which the system bases a decision can be compressed to a reasonably small number without altering the recognition predictions. This feature may be appealing to researchers who prefer systems that are potentially realizable in a neural architecture. Of more immediate significance, it becomes possible to use for recognition compressed values, that is, the $f(\lambda_j)$, without changing the approximation to the optimal model. These same compressed values can be used in the sampling equation for the selection of an image (e.g. Equation 4.5). When f is highly compressive, sampling is increasingly independent of the likelihood ratios; the more such independence exists, the greater is the tendency to spread out the time course of retrieval in cued recall. Without additional modelling we do not know the degree to which compressed sampling is needed to fit cued recall data, but it could well prove critical to have this option.

Optimizing cued recall strategies

Regardless of any possible compression, the existence of large likelihood ratios has some interesting implications for optimizing search strategies. For example, suppose that a cue is presented for recall, and that the likelihood ratio for one of the halves of pair image 1 is 10^8, and for one of the halves of pair image 2 is 10^2; thus both individually are quite likely to be matches. Suppose, however, that no response information at all was stored in the other half of pair image 1, but enough information was stored in the other half of pair image 2 to be certain of the response. If one decided that pair image 1 was the correct one, and decided to respond on this basis, a pure guess would have to be given. This seems like a poor strategy. Therefore, given that there is a reasonably high likelihood associated with image 2, it might seem that the best strategy would be to give the response in pair image 2. However, an optimal calculation reveals it would be better to make a random guess, excluding only the test word and the response found in pair image 2: despite the relatively high likelihood ratio of 10^2 for pair image 2, the odds against image 2 being correct are approximately $10^8/10^2 = 10^6$, larger than the subject's guessing vocabulary. This illustrates the kind of non-intuitive result that can occur when the likelihood ratios have extreme values.

Such observations just begin to touch on the complex issues concerning the most effective use of search strategies, many of which have to do with the competition between accuracy and response time; research on these matters must be left for the future. However, the scattering of results we have presented strongly suggest that cued recall departs significantly from an optimal process, and in addition perhaps points to some directions for future modelling.

Retrieval in generic and implicit tasks

We shall conclude with a few brief remarks concerning the effectiveness of retrieval during generic memory tasks (retrieval of our general knowledge), and implicit memory tasks (change in retrieval of knowledge caused by recent events). Our first observation is obvious but relevant: even when we have learned something quite well, successful retrieval is not guaranteed, especially in a short time frame. Nevertheless, in discussing explicit retrieval, we started with an optimal model for retrieval in recognition, and moved to a suboptimal model for retrieval in recall. It is conceivable that something similar occurs in retrieval from generic memory, with optimal retrieval for some tasks (perhaps including lexical decision, or word naming, say), and suboptimal retrieval for other tasks (perhaps including fact retrieval, say).

As a first pass at exploring this possibility, Schooler *et al.* (submitted) have extended the REM model to a set of word identification tasks explored by Ratcliff and McKoon (1997). We discuss first the application to forced choice identification: observers are given brief flashes of a word, followed by a mask. Then two words are presented in the clear, and the observer must choose which had been presented. Sometimes the two choices are visually similar, and sometimes dissimilar, although the same length. Sometimes one of the choices had been studied in an earlier list. The flash duration is manipulated in order to produce a rich set of parametric data. The results, illustrated in Fig. 4.5, show improved performance with longer flash time,

and improved performance with dissimilar choices, both unsurprising results that replicate much prior research. The effects of prior study were more interesting, with a substantial effect only for the case of similar alternatives. In this case the effect was relatively symmetric, demonstrating a tendency or bias to choose whatever choice had been studied, superimposed on any veridical perception; this bias was greatest at the lowest flash times.

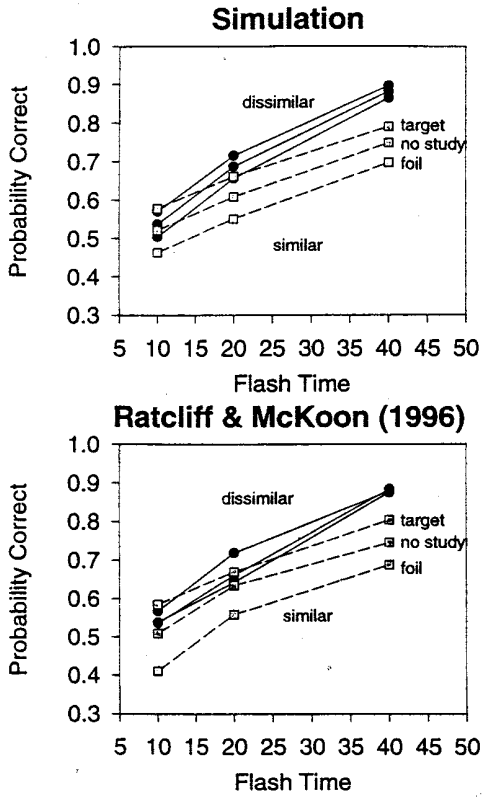

Fig. 4.5 Predictions (given in the top panel) of the REM model (Schooler *et al.*, submitted), fit to forced-choice word identification data (given in the bottom panel) collected by Ratcliff and McKoon (1997). Probability of correct forced-choice is given as a function of the display duration of the flashed word, for dissimilar forced-choice alternatives (solid lines) and similar forced-choice alternatives (dashed lines). In each set of three curves the top is for the case when the flashed word had been studied in an earlier list (labelled target), the middle is for the case when neither choice had been studied previously (labelled no study), and the bottom is for the case when the choice that had not been flashed had been studied in an earlier list (labelled foil).

Ratcliff and McKoon (1997) proposed an elegant counter model to predict these findings, but their model included a structural limitation in retrieval such that there was

a competition in the decision between the two choices that only existed for two similar word images in the lexicon. Schooler *et al.* (submitted) asked whether subjects might be operating in optimal fashion, given what had been stored in the lexicon, and what visual information was available from the flash. The model assumes that some visual features are seen veridically, and the rest are simply filled in by visual noise (perhaps caused by the mask). To these visual features are added a small number of current context features (visual or otherwise), and these features are held in visual short-term memory until the alternatives are presented. Each alternative is read and in accessing the visual/lexical images some of the context features stored in that trace are recovered as well. The decision rule is simple: each vector of features for the two choices (visual plus context) is compared with the stored vector of features from the flash (visual plus context): the alternative with the greater number of matching features is chosen. In this model, greater flash time produces more veridical features and hence better performance. Dissimilar alternatives makes the choice easier because the veridical features more clearly distinguish the choices in this case. Of greatest interest is the effect of prior study. In the model, prior study causes a few current context features to be added to the visual/lexical trace for the studied item. Some of these are recovered when the alternatives are read, and some of those will match whatever current context features have been added to the vector for the flash. The net result is that the number of matching features for a word studied earlier is slightly increased (by about one-half feature). For two similar alternatives, only about 13 features are diagnostic (differ between the alternatives) and the one-half extra feature produces a noticeable bias in favour of choosing the studied alternative. For two dissimilar alternatives the one-half extra matching feature tends to get lost in the roughly 44 total diagnostic features that differ between the two alternatives. The predictions of this model are illustrated in the lower half of Fig. 4.5. Clearly, the major patterns of the data are captured by this model, a model that assumes optimal retrieval and decision making. Of course the usual caveats must be stated that some of the error that this model places in registration of the visual flash (and in storage of the context features during initial study), could possibly be occurring in retrieval. The issues parallel those discussed for explicit recognition.

The next issue was possible extensions to other tasks. Ratcliff and McKoon (1997) also used yes–no matching (one alternative was presented, and the observer said whether it matched the flash), and naming (the observer tried to name the flashed word). Can any conclusions be reached about relative optimality? Consider yes–no first. In modelling this task, it seems clear that all the features registered from the flash become relevant for the matching decision (as opposed to the smaller number of features that differ between the two alternatives in forced choice). Thus, the model for yes–no has an extra parameter: the total number of features. This is a free parameter to be estimated, and makes it impossible to conclude anything concerning the relative degree of optimality of the two tasks (although the model fit quite well, even carrying over the common parameter values from forced choice). When the model is applied to naming, an optimal model must assume the system compares the vector of features extracted from the flash with all the visual/lexical images in memory, choosing the best match to produce, if anything is produced. As most of the errors were omissions, we employed a model variant in which a response was only given when the proportion of matching features for the best image exceeded a

criterion. The optimality of this variant depends on assumptions about the benefits of extra correct responses versus the costs of extra overt intrusions. Regardless of this issue, however, no conclusions about relative optimality in this task in comparison with forced choice and yes–no tasks could be reached, because a variety of assumptions had to be added to the naming model concerning the similarity structure of the entire lexicon in memory, issues that did not arise when decisions could be restricted to just two alternatives, or one. Thus, it can be concluded that an optimal retrieval model provides a plausible candidate for the processing and decision making in these various identification tasks, but a good deal of further research would be needed to come to more definitive conclusions.

Final remarks

It seems a bit anomalous to use as a starting assumption in this chapter the hypothesis that retrieval is 'optimal', given that the first author has for many years used as a starting assumption the view that decay of memories does not occur, but that forgetting is due instead to retrieval failure (e.g. Atkinson and Shiffrin, 1968; Shiffrin and Atkinson, 1969; Shiffrin, 1970*a*, *b*; Raaijmakers and Shiffrin, 1980, 1981; Shiffrin *et al.*, 1990; Murnane and Shiffrin, 1991). These two views are not as far apart as they seem at first glance. First, we have seen that recall is clearly suboptimal, and may well operate as a search based on proportional sampling. This is important because much of the discussion in the articles and chapters referenced above was concerned with free or cued recall. Second, retrieval failure in these articles included the use of poor cues with which to probe long-term memory; these poor cues could be the result of the unavailability of better cues, the change of context over time combined with a tendency to probe with the current context, or a difficulty in reconstructing an appropriate past context. What we have seen in this chapter is that retrieval under optimal conditions, with good probe cues in certain tasks, might well be close to optimal. On the other hand, we have also seen that optimality is a slippery concept, and some of the error we have assumed to lie in the storage process might instead be placed in retrieval without altering the predictions of the model.

Perhaps most important, analyses carried out under considerations of optimal retrieval allow relative effectiveness of retrieval to be assessed across tasks. We have seen, for example, that recognition of pairs of words is relatively less effective than recognition of single words, associative recognition is relatively less effective than recognition of pairs, and recall generally less optimal than recognition. Such analyses provide a relatively novel way of examining memory performance. The analyses in this chapter represent only a very small and tentative step down this road, but show promise for future development.

Acknowledgements

Support for the research reported in this chapter was provided by NIMH Grant 12717 to the first author. Requests for reprints should be sent to Richard M. Shiffrin, Psychology Department, Indiana University, Bloomington IN, 47405, or by email to shiffrin @indiana.edu.

References

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In *The psychology of learning and motivation: advances in research and theory*, (ed. K. W. Spence and J. T. Spence), Vol. 2, pp. 89–195. Academic Press; New York.

Clark, S. and Shiffrin, R. M. (1987). Recognition of multiple-item probes. *Memory and Cognition*, **15**, 367–78.

Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language*, **35**, 32–52.

McClelland, J. and Chappell, M. (in press). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*.

Murnane, K. and Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**(5), 855–74.

Nobel, P. A. (1996). Response times in recognition and recall. Ph.D. dissertation. Indiana University, Bloomington, IN.

Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In *The psychology of learning and motivation*, (ed. Bower, G. H.), Vol. 14, pp. 207–62. Academic Press, New York.

Raaijmakers, J. G. W. and Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, **88**, 93–134.

Ratcliff, R. and McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, **104**(2), 319–43.

Shiffrin, R. M. (1970*a*). Memory search. In *Models of memory*, (ed. D. A. Norman), pp. 375–447. Academic Press, New York.

Shiffrin, R. M. (1970*b*). Forgetting, trace erosion or retrieval failure? *Science*, **168**, 1601–3.

Shiffrin, R. M. and Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review*, **79**, 179–93.

Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**(2), 145–66.

Shiffrin, R. M., Ratcliff, R., and Clark, S. (1990). The list-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 179–95.