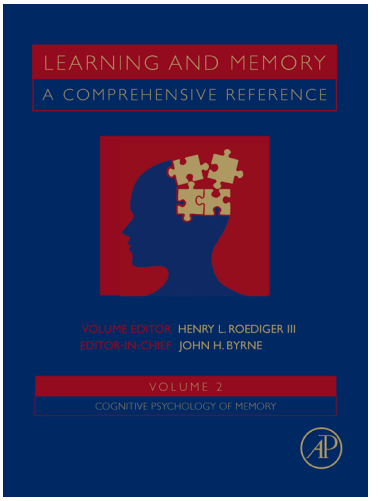

J. Raaijmakers. Mathematical Models of Human Memory. In H. L. Roediger, III (Ed.), Cognitive Psychology of Memory. Vol. [2] of Learning and Memory: A Comprehensive Reference, 4 vols. (J.Byrne Editor), pp. [445-466] Oxford: Elsevier.

2.25 Mathematical Models of Human Memory

J. G. W. Raaijmakers, Universiteit van Amsterdam, Amsterdam, The Netherlands

2.25.1	Introduction	445
2.25.2	The ACT Model	448
2.25.3	The SAM and REM Models	453
2.25.3.1	The SAM Model and Related Models	453
2.25.3.2	The REM Model	456
2.25.4	Neural Network Approaches	459
2.25.5	Models for Serial Order Memory	462
2.25.6	Concluding Remarks	463
References		464

2.25.1 Introduction

In this chapter, I will provide a brief introduction to formal models of memory. Although such approaches have become quite successful, it would be an overstatement to say that they enjoy a great popularity among mainstream experimental researchers interested in human memory processes. There are probably several reasons for this skepticism, but an important one seems to be that it is not always easy to see what a model adds compared to a verbal theory or explanation. In this chapter, I will discuss a number of the most important theoretical approaches, paying special attention to the issue of what these models can do that could not be done using only verbal theorizing.

Formal or mathematical models of memory can be broadly classified in terms of their scope and generality. At the simplest end, we have descriptive models that try to characterize lawful empirical regularities. Memory researchers, for example, have tried to characterize the form of the forgetting function, the function that relates memory performance (percent recalled or some other measure) to the retention interval, the time since the item was studied. Although several promising candidate functions have been proposed (most notably power and logarithmic functions; see Wixted and Ebbesen, 1991), the issue of which function best describes the forgetting curve has not been resolved. One reason is that many candidate functions capture the basic aspects of the forgetting curve, i.e., a curve that is characterized by a decreasing rate of decline (the older the trace, the less likely it is that it will be forgotten in the next unit of time). Another reason is that the comparison between different functions is complicated by the fact that some models are more versatile than others (can handle more different shapes, can mimic data generated by other models), which means that it is easier for such a model to fit any given set of data, although at the expense of its generalizability to new data (for a discussion of these issues, see Lee (2004) and Myung and Pitt (2002)). Hence, although such descriptive models may be useful for predictive purposes, a shortcoming of these models is that they are limited in scope, predicting only one type of relation. What is lacking in such models is an account of what causes the forgetting, making it difficult to devise experimental tests that would pit one model against the other. Similar issues arise in attempts to model the learning curve, the function that describes the increase in performance as a function of the number of learning or training trials.

At the next level, we have models that try to account for the basic learning and forgetting data in terms of what happens to individual memory traces. One issue that the descriptive models usually do not discuss is whether the proposed forgetting (or learning) function describes each and every memory trace or just the average of a large number of separate curves. This question was the main focus of a large number of studies conducted in the 1950s and 1960s. In a series of studies using the so-called RTT paradigm, in which one study or reinforcement trial (R) was followed by two test trials (T) without any additional study in between, it was shown that the probability of a correct response (success) on the second test trial given no success at the first test trial was nearly zero and much lower than the average probability of a success. This seemed to be

indicative of one-trial or all-or-none learning: The item was either completely learned on the study trial or not at all. This contradicted the standard assumption that learning was gradual. Such gradual learning functions were predicted by so-called linear operator models that assumed that the probability of a success on a given trial n was a simple linear function of the probability of success on the previous trial. Thus,

$$p_{n+1} = Q(p_n) = \alpha p_n + \beta \qquad [1]$$

where α and β are parameters that depend on the nature of the reinforcement given on trial n. The crucial assumption here was that this function described the behavior of each and every item independent of whether the response to that item had been correct on trial n.

To account for the results of the RTT paradigm, an alternative model was proposed in which the learning of an item was all-or-none: The item was either learned, always leading to a correct response, or not learned, in which case the probability of a success was at chance level. This model still predicts a gradual learning curve because such a curve represents the average of a number of items and subjects, each with a different moment at which learning takes place. The learning process in the all-or-none model may be represented by a simple Markov chain with two states, the conditioned or learned state (L) in which the probability correct is equal to 1, and the unconditioned state (U) in which the probability correct is at chance level (denoted by g). The following matrix gives the transition probabilities, the probabilities of going from state X (L or U) on trial n to state Y on trial $n+1$.

$$\begin{array}{l} \text{state on trial } n+1 \quad \text{P(Correct)} \\ \qquad\quad L \quad\quad U \\ L \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} \begin{bmatrix} 1 \\ g \end{bmatrix} \\ U \\ \text{state on trial } n \end{array} \qquad [2]$$

Strong support for the all-or-none model was obtained in an experiment by Bower (1961) in which subjects were presented lists of ten paired associate items consisting of a consonant pair and either the digit 1 or 2. This experiment was a breakthrough in the mathematical modeling of learning and memory because it did not just fit the learning curve but also a large number of other statistics (such as the distribution of the number of errors and of the trial of last error). The model fitted Bower's data remarkably well and this set a new standard for mathematical modelers.

One of the key predictions of the model was what became known as presolution stationarity: If the all-or-none assumption holds, the probability of responding correctly prior to learning (or prior to the last error) had to be constant:

$$P(e_{n+1}|e_n) = \text{constant for all } n \qquad [3]$$

Figure 1 shows the data from Bower's (1961) experiment and the predictions from the all-or-none and linear models. The data are in almost perfect agreement with the predictions of the all-or-none model and clearly inconsistent with those of the linear model. It may be shown that this presolution stationarity property is crucial for the all-or-none model in that the combination of this property together with the distribution of the trial of last error is a sufficient condition for the all-or-none model. That is, if both of these properties hold, the all-or-none model has to be the correct model. Since this property is strong evidence for the all-or-none model, it is understandable that proponents of gradual learning models tried to reconcile the finding with a model in which learning was more gradual. The argument that was used was based on the idea that the result might be explained if individual differences in the speed of learning were assumed. If items and/or subjects differ in their learning rate, errors on later trials might be

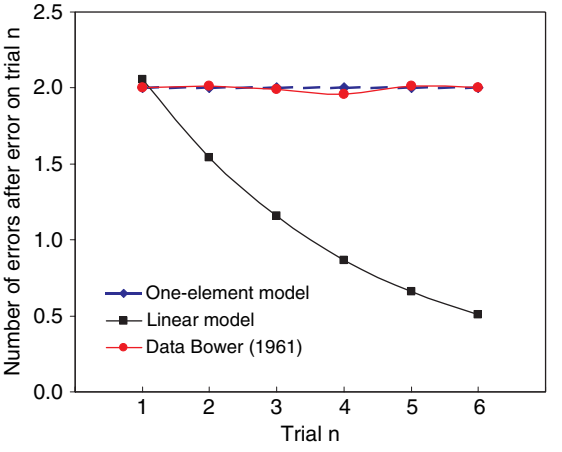

Figure 1 Number of errors following an error on trial n as a function of n. Data from Bower GH (1961) Application of a model to paired-associate learning. *Psychometrika* 26: 255–280; predictions for the all-or-none model and the linear model.

coming mainly from the more difficult items and from subjects with slower learning rates. However, in an ingenious analysis, Batchelder (1975) showed that this could not work. Batchelder analyzed the predictions of the linear operator model (eqn [1]) using a completely arbitrary distribution for the learning rate parameter and proved that it was impossible for the linear operator model to fit these results.

The success of the all-or-none model soon led to a series of related models that were based on the notion of discrete changes in the state of studied items. One issue that was investigated was whether this notion could account for transfer effects based on underlying conceptual categories. For example, suppose that several lists of paired associates are learned in succession where the stimulus items that belong to a particular conceptual category all have the same response. If learning is all-or-none, we might assume that a particular item will be learned in an all-or-none fashion as long as the conceptual relation is not yet discovered, but that once the relation has been discovered (which itself involves an all-or-none process) any new item belonging to the same category will start in the learned state rather than the unlearned state (i.e., no errors will be made on this item). Greeno and Scandura (1966), Batchelder (1970), and Polson (1972) showed that a relatively simple generalization of the all-or-none model gave a good account for the results of such experiments.

Although the all-or-none model was quite successful, the experiments that it was applied to were extremely simplified (simple stimuli, coupled with one of two possible responses). From the outset it was clear that the model would not hold for more complex experiments. However, perhaps the basic idea of the all-or-none model could be generalized in such a way that more complex learning tasks might be described as involving a series of stages, each stage being completed in an all-or-none manner. The most successful attempt at this type of generalization of the all-or-none model can be seen in the work of Greeno and associates (Greeno, 1968, 1974; James and Greeno, 1970; Humphreys and Greeno, 1970). Greeno did an extensive theoretical and empirical analysis of a two-stage learning model. As there are now two learning rate parameters, one for each stage, it becomes possible to look at the factors that affect each of these parameters and hence provide an interpretation for what the separate stages stand for. Contrary to the traditional two-stage theory of paired-associate learning (Underwood and Schulz, 1960), which maintained that the first stage involved a process of response learning and the second stage stimulus–response association, the results from the two-stage model proposed by Greeno were largely consistent with the idea that the first stage involved storage of the pair and the second stage learning to retrieve the pair.

Perhaps the most significant extension of the all-or-none model was proposed by Atkinson and Crothers (1964), who included the notion of a short-term memory state. The assumption here was that an item could move to a short-term state when it was studied but that it could move back to the unlearned state on subsequent trials when other items were being studied. Thus, such an item would show short-term forgetting: When tested immediately after having been studied, the response would be correct; however, when retested after several intervening trials, the probability of a correct response would be back at the baseline level (unless the item had moved to the learned state). The learning process in such models can be described using two transition matrices, one that applies when the target item is presented ($\mathbf{T}_1$) and one that applies when another item is presented ($\mathbf{T}_2$):

$$\mathbf{T}_1 = \begin{array}{c} \\ L \\ S \\ U \end{array} \begin{array}{c} \begin{array}{ccc} L & S & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ d & 1-d & 0 \\ wc & w(1-c) & (1-w) \end{bmatrix} \end{array} \qquad [4a]$$

$$\mathbf{T}_2 = \begin{array}{c} \\ L \\ S \\ U \end{array} \begin{array}{c} \begin{array}{ccc} L & S & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ (1-f)r & (1-f)(1-r) & f \\ 0 & 0 & 1 \end{bmatrix} \end{array} \qquad [4b]$$

where L is the state in which the item has been learned, S is the short-term memory state, and U is the state in which the item is not learned.

Several variants of such LS-models (Long-Short) were introduced, including ones that assumed that there could be additional storage (as well as forgetting) on intervening trials (note the parameter r in $\mathbf{T}_2$). This notion is of course related to the more general concepts of rehearsal and consolidation. The idea of storage on trials intervening between presentations might provide an explanation for the spacing effect, the finding that (in general) spaced study presentations are more beneficial for

later recall than massed presentations. Bjork (1966), Rumelhart (1967), and Young (1971) developed (increasingly complex) models to account for such spacing effects in paired-associate recall, leading to a model that became known as the General Forgetting Theory. However, these models never gained much popularity, perhaps because they were introduced at a time when the emphasis in the formal modeling of memory processes shifted to the next level following the 1968 publication of the Atkinson-Shiffrin model.

The theoretical framework that was proposed by Atkinson and Shiffrin (1968) made a distinction between structural properties of the memory system that were fixed and permanent, and control processes that operated on those structures. Control processes included such processes as rehearsal, coding, and retrieval strategies. The Atkinson-Shiffrin model assumed that information first enters a Short-Term Store (STS) and that the processing within STS determines storage in a permanent memory system, Long-Term Store (LTS). Information that is still present in STS at the time of testing will be readily available, but information that is no longer in STS will have to be retrieved from LTS. The probability of successful retrieval from LTS was a function of the strength of the LTS trace, which was itself determined by the nature of the processing in STS.

An important advancement of the Atkinson-Shiffrin theory was the model that was proposed for rehearsal processes in STS. It was assumed that at any time only a few items could be simultaneously in STS and that once STS was filled, any new item would have to replace one of the other items in STS. This idea led to the introduction of the concept of a rehearsal buffer as a simple model for rehearsal in STS, or rather a family of models since various alternatives were considered that differed in whether older items were more or less likely to be replaced by a new item. In these models, it was assumed that storage in LTS is directly related to the length of time that a particular item stays in the buffer. This storage assumption has often been misinterpreted as implying that the Atkinson-Shiffrin theory would assume that only time in STS determines how much information gets stored in LTS. However, Atkinson and Shiffrin proposed that rehearsal in STS is a control process and that the nature of the processing in STS will vary depending on the requirements of the task. In some tasks, the emphasis will be on simply maintaining the information in STS, but in other tasks the emphasis is on coding the information in LTS. This distinction between coding and rehearsal (or elaborative and maintenance rehearsal as it was later called) made it possible to accommodate levels-of-processing effects, i.e., the notion that the nature of the processing in STS determines the probability of later recall. Thus the standard textbook story that supposes that there is a fundamental difference between the Two-Store model and the levels-of-processing framework proposed by Craik and Lockhart (1972) is incorrect (see also Raaijmakers, 1993).

The major significance of the Atkinson-Shiffrin model was that it was not simply a model for one specific experimental task but a general framework within which models could be formulated for specific tasks. Thus, in addition to the short-term memory tasks investigated in the 1968 paper, the same general framework was applied to search and retrieval processes in long-term memory (see Shiffrin, 1968; Shiffrin and Atkinson, 1969), free recall (Shiffrin, 1970), and recognition memory (Atkinson and Juola, 1974). This was a major step forward compared to, for example, the General Forgetting Theory that did not allow a simple generalization to free recall or recognition paradigms. This type of approach in which a general framework is presented within which specific models are developed for specific tasks is a common characteristic of most current models of memory. In the next sections, I will discuss a number of such approaches with special attention given to the question how these models are able to provide novel explanations for experimental findings.

2.25.2 The ACT Model

The first model that we will discuss in more detail is the ACT theory developed by John Anderson. The ACT theory (Adaptive Control of Thought) has its roots in early theories of spreading activation (Collins and Loftus, 1975) and the work of Newell and Simon on cognitive architectures. ACT is not just a model for memory processes but aims to provide a general framework or architecture for all cognitive tasks (sometimes termed a Unified Theory for Cognition). Although ACT has undergone many changes since it was first presented in 1976, there are a number of aspects of the theory that have remained more or less the same over the years. First, ACT does not make a fundamental distinction between semantic and episodic memory. All knowledge facts and all experiences are stored in a single declarative memory system. Second, ACT makes a distinction between a working memory system, a declarative memory system, and a procedural memory

system. Declarative memory is modeled as a large set of interconnected nodes or chunks, while the procedural system has the form of a large set of production rules (rules of the form IF conditions A, B, and C are satisfied, THEN action Y is performed) that fire whenever their conditions are satisfied. Although one sometimes gets the impression that ACT is more a programming language in which various cognitive tasks may be modeled, the ACT framework has been used to develop detailed quantitative models for various memory tasks that do make specific and testable predictions.

In the original ACT model (Anderson, 1976), retrieval of a target item B from a cue item A was based on a notion of spreading activation in which a particular node was either active or inactive. The spreading of activation was controlled by the relative strength of the links from the cue to the nodes that were connected to the cue node. Once a node was activated, it would in turn start to activate other nodes associated with it (a threshold was assumed to prevent activation of all nodes). Since activation is all-or-none, response latency was determined by the time it took for activation to spread to the target node. However, using a primed lexical decision task, Ratcliff and McKoon (1981) showed that the semantic distance between the prime and the target does not affect the time at which the facilitation due to priming begins to have its effect, although it does affect the magnitude of the facilitation. Anderson (1983a,b) proposed a revised version of ACT, named ACT*, in which nodes were no longer activated in an all-or-none fashion. In ACT*, each node had a continuously varying activation value. The larger the activation value, the faster and the more likely it was that the trace would be retrieved.

Anderson (1981, 1983b) showed how this model could be used to explain a number of memory phenomena. In ACT*, performance is determined by the strength of the target trace relative to that of other traces associated with the retrieval cues used. On each presentation of an item, there is a probability that a trace will be formed and once formed, further presentations provide additional strength to the trace. The strength added to a trace was assumed to decay according to a power law. More specifically, the trace strength (S) for a trace that has been strengthened n times is equal to:

$$S = \sum_{i=1}^{n} t_i^{-b} \qquad [5]$$

where t_i is the time since the i-th strengthening and b is a decay parameter (between 0 and 1).

Anderson (1981, 1983b) showed that the ACT* model predicts a large number of standard findings from the memory literature. One intriguing result that came out of this analysis was that performance in recall tasks is a function of both the absolute and the relative strength of the target trace. In ACT*, the probability of recall is a function of both relative and absolute strength, but the latency is a function of the relative strength only. Anderson (1981) demonstrated that this implies that in a standard interference task there will be an interference effect on latency, even when the conditions are equated on percent correct. This result implies that it will not be possible to completely equate interference and control conditions at the end of second-list learning, as was implicitly assumed in many experiments on interference and forgetting (e.g., when both conditions learn to the same criterion). Basically, this prediction is due to the fact that if probability of recall is a function of both relative and absolute strength, it must be the case that in the condition in which it takes longer to reach a particular recall criterion, the absolute strength will be larger at the point where the criterion is reached. Hence, to get equal percent recall, this must be compensated for by a lower relative strength, hence a longer latency.

In a similar way, it can be shown that if the second list is again learned to a fixed criterion, performance on the second list may show proactive facilitation instead of interference, when it is tested after a delay in such a way that differences in relative strength are less important and performance is mostly determined by the absolute strength of the target trace. The latter may be experimentally accomplished by giving an unpaced test in which subjects are given ample time to produce the response. In such a test, differences in relative strength become less important since eventually the trace will be retrieved, although it may take a long time. Anderson (1983b) reports results that confirm this counterintuitive prediction. Mensink and Raaijmakers (1988) showed that these predictions hold not only for the ACT* model, but for all models in which performance is a function of both relative and absolute strength.

The latest version of ACT, called ACT-R (ACT-Rational), is based on a number of assumptions that are quite different from ACT*, yet the model shares enough features with the older models to justify using the same acronym. There are two important differences with ACT*. First, ACT-R no longer assumes a spreading activation conception of memory retrieval. Rather, it is assumed that activation of a memory trace or chunk is a

direct function of the association between the source elements (the retrieval cues) to that chunk and there is no spread of activation to other chunks from a chunk that is not itself a source of activation. Second, ACT-R is based on the assumption that the cognitive system is a rational system, i.e., the rules that govern the activation of information from memory are such that they optimize the fit to the environmental demands. This rational approach to cognition has been very influential (see also more recent models such as the REM model (Shiffrin and Steyvers, 1997) that will be discussed later in this chapter).

To appreciate this rational approach, it is helpful to consider some of the results discussed by Anderson and Schooler (1991). Anderson and Schooler showed that many of the functional relationships that we know from standard memory experiments (e.g., the typical learning and forgetting functions) can also be seen in the environment with material that has little to do with memory *per se*. For example, the probability that a particular word will appear in the headline of *The New York Times* or the probability that one will get an e-mail from a specific person obey the same functional relations as we know from memory research. If a particular word has appeared in the headline the probability that it will appear again after X days follows the same power law that we are familiar with when looking at standard retention functions. Thus, the basic idea of ACT-R is that the cognitive system has developed in such a way as to provide an optimal or rational response to the information demands of the environment: The probability that a particular item will be remembered at a particular time reflects the probability that it will be needed at that time.

This rational approach is reflected in the equations that ACT-R uses to describe the activation of a particular trace given that specific cues are present. In the ACT-R approach to memory (see Anderson et al., 1998) it is assumed that the activation of a chunk i depends both on its base-level activation (B_i, a function of its previous use) and on the activation that it receives from the elements currently in the focus of attention:

$$A_i = B_i + \sum_j W_j S_{ji} \qquad [6]$$

where S_{ji} is the strength of the association from element j to chunk i and W_j is the source activation (salience) of element j. If we interpret the base-level activation as similar to the prior odds of the chunk being needed and the second term as similar to the (log) likelihood of the trace given the available evidence (the cues), then the similarity of eqn [6] to Bayes' rule becomes evident. (According to this rule, the logarithm of the posterior odds is equal to the log prior odds plus the log likelihood ratio.) According to ACT-R,

$$S_{ji} = S + \ln(P(i|j)) \qquad [7]$$

where $P(i, j)$ is the probability that chunk i will be needed when element j is present or active. Note that since $P(i, j) \leq 1$ the logarithm of $P(i, j)$ will be ≤ 0 and hence S represents the maximum value that S_{ji} can obtain. For all practical purposes, these S_{ji} may be viewed as reflecting the associations between the cues j and the target trace. In ACT-R (see Anderson et al., 1998: 344), it is typically assumed that if there are m elements associated to the cue j, each will have a probability of $1/m$, hence:

$$S_{ji} = S + \ln(1/m) = S - \ln(m) \qquad [8]$$

Note that this equation assumes that for the associative activation S_{ji} it does not matter that a particular association may have become stronger in the course of the experiment: all that matters is the number of associative links from the cue to other elements or its fan. This seems a rather strong assumption, yet it does play an important role in ACT-R's handling of data from recognition experiments.

The first part of eqn [6], the base-level activation, reflects the activation that remains from previous presentations of the target trace or chunk. The activation of a chunk is subject to decay so that the longer ago the chunk was strengthened, the less the contribution of that activation to the current base-level activation. The equation for the base-level activation is thus given by:

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) + B \qquad [9]$$

In this equation, n is the number of times the chunk has been retrieved from memory, t_j indicates the length of time since the j-th presentation or rehearsal, and d and B are constants. It is evident that eqn [9] is closely related to eqn [5] that describes the activation in ACT*.

Finally, as in ACT*, it is assumed that the latency of a response is an exponentially decreasing function

of the activation level of the corresponding chunk. However, unlike ACT*, ACT-R does not simply look at the activation of the target trace but takes into account other traces or chunks that might be activated. It is assumed that the system will always retrieve the chunk with the highest activation (provided it is above the threshold). Due to the presence of noise in the system, the activation values will not have a fixed value but rather a probability distribution (a logistic distribution is assumed). The probability that a chunk with a mean activation value of A_i (and variance σ^2) is above a threshold τ is then equal to:

$$\Pr(i) = \frac{1}{1 + \exp[(A_i - \tau)/s]} \quad \text{where} \quad s = (\sigma\sqrt{3})/\pi \qquad [10]$$

If there are more chunks above threshold, the system will choose the one with the largest activation. The probability that the target chunk has the largest activation is given by an equation similar to the Luce choice rule:

$$P(\textit{choose } i) = \frac{\exp(A_i/t)}{\sum_j \exp(A_j/t)} \quad \text{where} \quad t = (\sigma\sqrt{6})/\pi \qquad [11]$$

Although ACT-R is much more than a model for memory, it does explain quite a number of findings from the memory literature. We will briefly discuss two such applications, the analysis of recognition memory proposed by Anderson et al. (1998) and the model for spacing effects developed by Pavlik and Anderson (2005).

Any ACT-R model begins with the specification of a number of production rules. In the recognition model, the basic production rules are the rules for Yes and No responses, which simply state that if a trace is found that corresponds to seeing the item in the list context, a Yes response will be made and another rule that applies when the first one fails and that generates a No response. Hence, contrary to most other current models for recognition, ACT-R is not based on a signal-detection-like approach but rather on the retrieval of a trace representing the item in the list context. Note that in such a model negative responses (No responses) are not based on a low familiarity value but on the fact that the rule for generating a positive response passes a waiting time threshold. Although such an approach may work well for explaining data observed on positive responses, there are some problems when negative responses are to be explained. First, this type of model has no simple solution to generate fast negative responses. Second, the model predicts that negative responses are not affected by various experimental factors (e.g., list length) unless one assumes that the waiting time threshold itself is a function of those factors (a solution that is hard to defend).

According to ACT-R, performance in a standard recognition task is determined by the activation of the chunk representing the tested item. According to eqn [6], this is a function of the base-level activation and the associative activation that it receives from the cues (the presented word and the list context). Hence,

$$A = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) + B + W_W S_W + W_L S_L \qquad [12]$$

where W_W is the weighting given to the word, S_W is the strength of the association from the word to the trace, W_L is the weight of the list context, and S_L is the strength of the context association. According to Anderson et al. (1998: 348), the first term may be approximated by:

$$\ln\left(\sum_{j=1}^{n} t_j^{-d}\right) = \ln\left(\frac{anT^{-d}}{1-d}\right) = C + \ln(n) - d\ln(T) \qquad [13]$$

where C captures the constant terms. Since $W_W S_W$ is also a constant and S_L is equal to $S - ln(L)$ according to eqn [8], the activation of eqn [12] may be written as:

$$A = B' + \ln(n) - d\ln(T) - W_L \ln(L) \qquad [14]$$

where B' combines all the constant effects, n equals the number of presentations/rehearsals, T is the time since presentation, L is the list length, d is the decay rate, and W_L is the attentional weighting of the list context. In their analyses, Anderson et al. (1998) set d and W_L equal to 0.5.

One interesting finding that this model predicts (and that would have been difficult to foresee without actually running the simulations) is the differential effect of list length and list strength in recognition. The list length effect refers to the effect of the number of other items on the list, while the list-strength effect refers to the effect of the strength of those other items (where strength might be manipulated by such factors as presentation time or additional presentations). In recall paradigms, both of these effects are present but in recognition tasks there is no effect of list strength (or a slightly reversed effect), although

there is a list-length effect. Shiffrin et al. (1990) showed that it is very difficult for many models to predict both the presence of an effect of the number of other items, yet no effect of the strength of those other items. ACT-R's recognition model, however, does explain this intricate pattern of results. The basic reason is that in ACT-R, strength manipulations affect the base-level activations whereas the length of the list mainly affects the associative activation (i.e., the fan effect; see eqn [13]). There are a few other factors that play a role (such as small differences in retention interval when presentation time or the length of the list is varied) but the main effects are due to these two factors. Hence, increases in strength affect the base-level activation for the tested item but do not affect the interfering effect of the other items on the list. Of course, one might question the assumption that strength manipulations do not affect the associative activation (as was the case in ACT*), but even so, the ACT-R analysis points to a possible solution to the puzzle of length and strength effects, a pattern of results that has proved difficult to accommodate in other models for recognition.

Pavlik and Anderson (2005) presented an application of ACT-R to account for spacing effects in paired associate recall tasks. They showed that their model could account for all of the standard findings in the spacing literature including a new experiment that they performed in which spacing was varied over much longer intervals than is normally the case in these experiments. In their experiment, there were two sessions separated by 1 or 7 days. During the first session, the subjects learned the English translations for a number of Japanese words. The pairs were presented four or eight times during the first session with interpresentation spacings of two, 14, or 98 trials. During the second session, they were given a number of test trials on the pairs learned during the first session. The data showed a crossover interaction such that the shorter spacings led to better performance at the end of the first session but worse performance at the start of the second session (see **Figure 2**).

In the application of ACT-R to this experiment, the associative activation will be constant and hence the analysis focuses on the base-level activation. Without any modifications, the ACT-R model does not predict such spacing effects (Pavlik and Anderson, 2005: 570), so some changes are necessary. The most likely candidate is the decay rate parameter d (see eqn [9]). In order to account for spacing effects, the decay rate has to be made

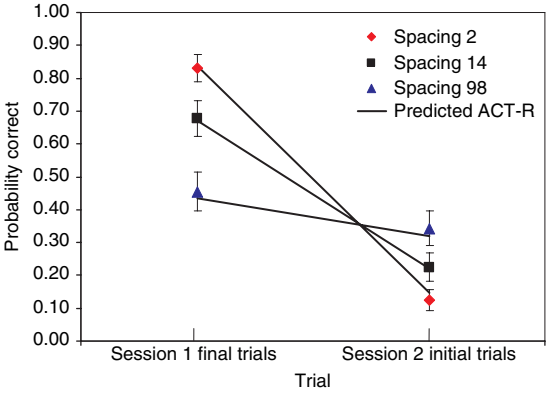

Figure 2 Probability of a correct response before and after the retention interval as a function of the spacing between the presentations during session 1. Observed data from Pavlik PI and Anderson JR (2005) Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cogn. Sci.* 29: 559–586; predictions from the ACT-R model. Error bars correspond to two standard errors.

sensitive to the intervals between successive presentations. The formulation that Pavlik and Anderson (2005) used is based on the assumption that the decay rate for the contribution from the j-th presentation is a function of the activation at the time of the j-th presentation. Thus, eqn [9] is replaced by the following equation for the activation after n presentations:

$$B_n = \ln\left(\sum_{j=1}^{n} t_j^{-d_j}\right) \text{ with } d_j = ce^{B_{j-1}} + a \qquad [15]$$

If at the start of the j-th presentation the activation was high (i.e., the activation after j–1 presentations, B_{j-1}), d_j will be larger and thus the contribution from that trial at later tests will be lower due to the more rapid decay. Hence, the effect of long spacing intervals (characterized by low activation at the end of the retention interval) will be longer lasting and this more than compensates for their longer retention intervals, thus leading to a spacing effect.

These two examples illustrate the way in which task-specific models are constructed within the ACT-R framework. As mentioned before, ACT-R is an ambitious attempt to provide a unified theory of cognition. As such, restricting the evaluation to just its contribution as a memory model clearly does not do justice to the theory as a whole. However, even though the ACT-R has not been evaluated as extensively as some of the other memory models, the theory has already made a large number of contributions (see Anderson et al., 1998; Pavlik and Anderson,

2005). There have also been extensions of the framework to implicit memory effects, but these need to be investigated more thoroughly to determine whether they are indeed viable explanations of priming effects. A recent extension of the ACT-R framework is the identification of specific modules within ACT-R with specific regions in the brain. Anderson and colleagues (Anderson et al., 2003, 2004, 2005) have shown that the duration of those components can be mapped onto the BOLD response obtained in the associated brain regions (using the assumption that the duration but not the intensity of a specific component is reflected in the BOLD response). This of course opens up a whole new approach to the validation of the general ACT-R theory and also provides a much-needed theoretical framework for the interpretation of neuroimaging data. All in all, then, ACT-R represents an excellent example of the trend toward more general theories that has characterized recent research on mathematical models for memory processes.

2.25.3 The SAM and REM Models

2.25.3.1 The SAM Model and Related Models

The next model that we will discuss is the SAM model (Raaijmakers and Shiffrin, 1980, 1981b) and a number of related models that have been proposed in recent years. The SAM model (Search of Associative Memory) started out as a model for free recall (Raaijmakers, 1979). It was soon realized that the model could be generalized to paired-associate recall (Raaijmakers and Shiffrin, 1981a) and recognition (Gillund and Shiffrin, 1984). The model was subsequently extended to handle interference and forgetting (Mensink and Raaijmakers, 1988, 1989) and, more recently, spacing effects (Raaijmakers, 2003). Related models in which a semantic memory component was added have been proposed by other researchers, e.g., PIER2 (Nelson et al., 1998) and eSAM (Sirotin et al., 2005). In addition, Shiffrin and coworkers have developed a new model, REM, that is in many ways similar to SAM, but provides a solution to some problems relating to recognition memory, and that has also been extended to semantic and implicit memory paradigms (Shiffrin and Steyvers, 1997; Schooler et al., 2001; Wagenmakers et al., 2004).

The original SAM model was based on a search model proposed by Shiffrin (1970). It shared a number of characteristics with the Atkinson-Shiffrin theory such as the notion of a STS buffer as a model for rehearsal processes and the assumption that storage in LTS is a function of the nature and duration of rehearsal in STS. SAM assumes that when a specific event occurs (this could be anything but in most analyses it is simply the presentation of an item on a study list) various types of information are stored in the memory trace representing that event. Any type of information might be stored in the trace (the memory image, as it is usually called in SAM), but the model uses a classification in item, associative (interitem), and contextual information. Retrieval of information from LTS is a cue-dependent process, i.e., what is retrieved from LTS depends on the information that is present in STS at the time of the retrieval. In applications of SAM to typical memory paradigms such as free recall or recognition, the cues may be words from the studied list, category cues, and contextual cues.

Whether or not a specific memory trace is retrieved depends on the relations between the cues and the information stored in the trace. These relations are defined in a retrieval structure, a matrix that gives the associative strengths between possible cues and the stored memory image. A crucial assumption in SAM is that when several cues are used simultaneously (e.g., context and a retrieved item), the overall strength of the set of cues (Q_1, Q_2, etc.) to a specific trace is given by the product of the individual associative strengths:

$$A(i) = \prod_{j=1}^{m} S\left(Q_j, I_i\right) \qquad [16]$$

where $A(i)$ is the combined strength or activation of image I_i, and $S(Q_j, I_i)$ is the strength of association between cue Q_j and image I_i. The most important aspect of this eqn [16] is the assumption that individual cue strengths are combined multiplicatively into a single activation measure. This multiplicative feature focuses the search process on those memory traces that are strongly associated with all cues, the intersection of the sets of traces activated by each cue separately. An important aspect of SAM is that retrieval strategies are implemented in the choice of retrieval cues but once a specific set of retrieval cues is used, the retrieval process is automatic and completely determined by the relations between the retrieval cues and the information stored in memory.

The activations $A(i)$ determine both the probability of retrieval of a memory trace in recall tasks as well as the probability that an item will be recognized

as having been presented on the study list. It is assumed that in recall tasks the probability of being able to generate the answer depends on selecting or sampling the correct target trace and on the probability that enough relevant features from the stored trace are activated to enable the reconstruction or recovery of the answer. It is assumed that the system may sample several times before giving up, but if recovery fails once sampled, it will fail again if the same trace is sampled a second time using the same cues.

More specifically, the probability of sampling a trace is assumed to be proportional to the activation strength of the trace:

$$P_S(I_i) = \frac{A(i)}{\sum A(k)} \qquad [17]$$

The probability of recovery is assumed to be an exponential function of the summed strengths of the retrieval cues to the sampled image:

$$P_R(I_i) = 1 - \exp\left[-\sum_{j=1}^{m} S\left(Q_j, I_i\right)\right] \qquad [18]$$

Combining these assumptions, an equation can be derived that gives the probability of recall for a simple cued recall test in which the same set of cues is used for a maximum of *Lmax* retrieval attempts:

$$P_{recall}(I_i) = \left[1 - (1 - P_s(I_i))^{L_{\max}}\right] P_R(I_i) \qquad [19]$$

The above equations apply to cued recall. SAM was, however, initially developed as a model for free recall, which is more complicated since during the search process other list items may be retrieved and these may then be used as new retrieval cues. In SAM it was assumed that during the presentation of the list items, a few items may be simultaneously rehearsed and that storage of context, item, and interitem information was a function of this rehearsal process. That is, the amount of information that is stored for an item was assumed to be a function of the time that that item was rehearsed or the time that a specific pair was simultaneously rehearsed (in case of the interitem associations). For this part of the model, a buffer model similar to that of Atkinson and Shiffrin (1968) was used. At the time of testing, any items still in the buffer are first recalled (unless of course there are no items available anymore in the buffer) and then the process of retrieval from LTS itself starts. Initially, the search process is based solely on the context cues that are available but as soon as a list item is retrieved, that item is used as an additional cue. If this item+context search is not successful (i.e., if there are *Lmax* consecutive retrieval attempts that do not lead to new items being recalled) the system will revert back to using only the context cue. This process continues until no more new items can be recalled (within a reasonable time). For this latter aspect, a stopping criterion was used based on the total number of failed retrieval attempts (*Kmax*), but other stopping rules are also possible (although we have not seen a case where the nature of the stopping rule seems to matter). SAM also assumes that new information may be stored during the retrieval process. That is, if a new item is successfully retrieved, the associative connections between the probe cues and the sampled image are strengthened. Although conceptually simple, it turns out to be virtually impossible to derive analytical predictions for the model for free recall, hence all analyses have been done using Monte Carlo simulations.

Raaijmakers and Shiffrin (1980) reported a large number of such simulation results and showed that the SAM model gave an excellent account of many standard findings from the free recall literature. These included serial position curves, the effects of list length and presentation time, cumulative recall data, the phenomenon of hypermnesia, and many others. As an example, **Figure 3** gives the predictions from SAM and the observed data for the experiment of Roberts (1972) in which presentation time and list length were varied over a wide range.

Of particular interest was the prediction by SAM of the part-list cuing effect (extensively discussed in Raaijmakers and Shiffrin, 1981b). This effect refers to the finding that presenting a random sample from the list items as additional cues did not have the expected positive effect on the recall of the remaining list items as one would have expected based on the notion that subjects use interitem associations during recall. SAM's ability to generate the part-list cuing effect was rather surprising since it ran counter to the then standard interpretation of that effect in terms of inhibitory factors. Subsequent experiments (reported in Raaijmakers and Phaf, 1999) demonstrated the viability of SAM's account of the part-list cuing effect.

SAM assumes that recall and recognition involve the same basic process of activating information. However, when a specific item X is tested for recognition, the response is not based on the retrieval of information from just the trace corresponding to X (although there is no principled reason why it could not be) but on the overall activation of the memory system induced by the retrieval cues. The overall activation is used as the familiarity measure in the

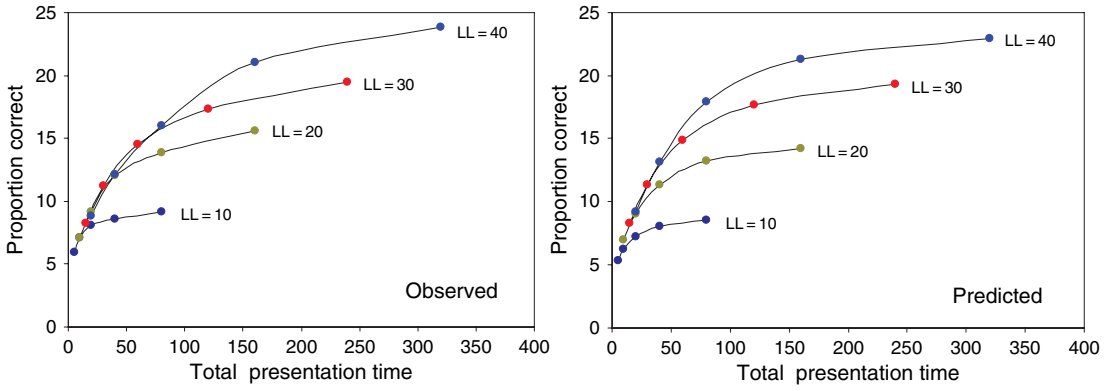

Figure 3 Observed (left panel, Roberts WA (1972) Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *J. Exp. Psychol.* 92: 365–372) and predicted (right panel) mean number of correct recalls in free recall as a function of presentation time and list length (LL). Predictions are based on the SAM model with parameter values as given in Raaijmakers JGW and Shiffrin RM (1980) SAM: A theory of probabilistic search of associative memory. In: Bower GH (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 14, pp. 207–262. New York: Academic Press.

standard signal detection model for recognition. This approach to recognition is termed a global familiarity model, in contrast to local familiarity models that are based on the familiarity or activation of the target trace. The global familiarity model is currently the most popular approach to modeling recognition and is used in a variety of models other than SAM (e.g., TODAM, MINERVA2, REM). One obvious advantage of the global familiarity approach is that it provides a simple way to deal with false alarms, the recognition of nonlist items (the distractor items), without having to make any additional assumptions. In the SAM recognition model developed by Gillund and Shiffrin (1984), the global familiarity measure is simply the overall activation in response to the retrieval cues used, i.e., $\sum A(k)$, with $A(k)$ as in eqn [16].

In the SAM model, the role of context cues in episodic memory retrieval is emphasized. Many experiments have shown that testing in a context that is different from the context at the time of encoding leads to a decrease in performance (especially in free recall tasks) compared to testing in the same context. This holds both for changes in the environmental context (e.g., Godden and Baddeley, 1975; Smith, 1979; Grant et al., 1998) and changes in the internal state or context (Eich et al., 1975; Eich, 1980). Mensink and Raaijmakers (1988, 1989) extended this notion to within-session changes in context. They assumed that within an experimental session there are gradual changes in context and that the context that gets stored in a trace is a selection from the currently available context elements. The model that they developed was adapted from Stimulus Sampling Theory (Estes, 1955) and assumed that there was a random fluctuation between a set of available or current context elements and a set of (temporarily) unavailable context elements. Mensink and Raaijmakers (1988) showed how such a notion of context fluctuation in combination with the SAM model for cued recall could account for many of the traditional results in the area of interference and forgetting. Using the same basic model, Raaijmakers (2003) showed that it could also account for standard spacing effects. A related analysis of contextual fluctuation processes as well as an application to free recall was developed by Howard and Kahana (1999; see also Kahana, 1996). Whereas in the original Raaijmakers and Shiffrin (1981b) analysis of free recall, a constant context was assumed during presentation and testing of a single list, Howard and Kahana (1999) made the reasonable assumption that context varies even within a single list and that upon retrieval of a specific trace not just the item information would be retrieved, but also the stored context information. They showed how such a model could account for a number of detailed aspects of recall processes.

Nelson et al. (1998) developed a model (PIER2) related to SAM that they showed could successfully explain a large number of findings on the effects of extralist cues on recall. In these experiments, a list of items is studied; at test, the subjects are given a cue and they are told that the cue item is meaningfully related to one of the list items. The basic idea of PIER2 is that during encoding of a list of words, explicit as well as implicit representations (traces) are formed. The implicit representation is an

automatic by-product of the comprehension process. Extralist cued recall may result from retrieving either the explicit or the implicit representation (or both). The PIER2 model focuses on the contribution to recall resulting from the implicit representation. It assumes that during encoding the study or target item as well as its associates are activated and that the activation strengths of both the target item and the associates are a function of their interconnectedness. At the time of testing, when the extralist cue is presented, a sampling function similar to that of SAM is used in which the probability of sampling the target item is proportional to the relative cue-to-target activation strength, relative to the strengths of the connections of the cue to its other associates and the strengths of the connections of the target to its other associates. Thus, the more unique the cue-to-target association (both at the cue side and at the target side) the higher the probability of sampling the (implicit) target representation. Using this sampling model, Nelson et al. (1998) showed that it successfully accounted for many results from previous experiments on extralist cuing.

Even though the SAM model has been quite successful in explaining a large variety of experimental results, the model in its original form fails to account for the list-strength effect (or rather the lack of it) and a number of other results in recognition (see Shiffrin et al., 1990). It soon became clear that in order to be able to explain these results, it would have to be assumed that the extent to which a trace is activated by an unrelated item cue should decrease as the number of features stored in that trace is increased (i.e., as the trace gets stronger). In SAM and most other models, it was assumed that the associative strength was a function of the number of overlapping elements, hence it should either stay the same or increase with the number of features stored.

A solution to this problem was found by adopting a so-called Bayesian or rational approach. In this type of approach (Shiffrin and Steyvers, 1997; McClelland and Chappell, 1998), it is assumed that the system, when confronted with an item that has to be accepted or rejected on a recognition test, makes an optimal decision based on the information that is stored in memory and knowledge of the rules that govern storage of information in memory. In the next section, we will discuss the REM model developed by Shiffrin and Steyvers (1997) as an example of this approach. A similar, independently developed model was presented by McClelland and Chappell (1998). Both of these models are based on the notion of differentiation, i.e., as an item gets stored better, it also becomes easier to differentiate from other items and will less likely be activated by cues representing other items. Although the models are quite similar in spirit (and would be considered equivalent on a purely verbal level), Criss and McClelland (2006) show that the two models are in fact not equivalent and will make different predictions for specific experiments (e.g., associative recognition). However, this analysis is beyond the scope of the present chapter.

2.25.3.2 The REM Model

As mentioned before, the REM model (Retrieving Effectively from Memory) is based on the assumption that the memory system behaves as an optimal decision-making system. On a simple recognition test, old and new items are presented and the subject has to decide whether the test item is old or new. REM assumes that the stored memory traces consist of samples of features from the studied items. Features may be stored correctly or incorrectly but as the study time increases, more features will be stored correctly. It is assumed that at test the system matches the features of the test item to each of the traces in memory. For a test item that was indeed on the list, there will of course be a relatively high number of matches and not many mismatches for the trace corresponding to that item. For all other traces (corresponding to the other items on the list) there will be more mismatches. For a distractor test item, all traces will have a relatively high number of mismatches and relatively few matches (since none of these traces corresponds to the test item). Hence, the number of matching and mismatching features gives information about whether the test item was on the list.

It is assumed that the system evaluates the evidence according to standard rules of probability theory and makes an optimal choice based on the available evidence. More specifically, the system chooses whichever response has the higher probability given the observed feature matches and mismatches in all the memory traces. Mathematically, the decision criterion is given by the posterior odds ratio, which according to Bayes' rule may be written as the product of the prior odds and the likelihood ratio:

$$\Phi = \frac{P(old|data)}{P(new|data)} = \frac{P(old)}{P(new)} \times \frac{P(data|old)}{P(data|new)} \qquad [20]$$

It can be shown that in REM, the likelihood ratio is given by the average likelihood ratio for the

individual list traces (assume L episodic images are compared to the test probe):

$$\Phi = \frac{1}{L}\sum_j \frac{P(D_j/old)}{P(D_j/new)} = \frac{1}{L}\sum_j \lambda_j \qquad [21]$$

Hence, an old response would be given if $\Phi > 1$. An interesting result from this analysis is that the decision rule turns out to be an example of the global familiarity approach to recognition memory. There are, however, two major differences between the REM and the SAM models for recognition. One is that in SAM the response criterion is basically arbitrary, whereas in REM there is a natural criterion corresponding to a likelihood of 1.0. The other difference is that in REM the activation value λ_j may be shown to be a function of both the number of matching and nonmatching features. For a simple version in which we simply count the number of matching and mismatching features, disregarding the exact value of the features (i.e., whether it is a very common or not so common value), it may be shown that the contribution to the overall likelihood for item j is given by:

$$\lambda_j = \left(\frac{\alpha}{\beta}\right)^{m_j}\left(\frac{1-\alpha}{1-\beta}\right)^{q_j} \qquad [22]$$

where α is the probability of a match given storage for the correct trace, β is the probability of a match given storage for an incorrect trace (α must obviously be larger than β), and m_j and q_j are the number of matches and mismatches, respectively, for trace j. Thus, the higher the number of matching features, the higher the likelihood, and the higher the number of mismatching features, the lower the likelihood. Earlier we mentioned the need to include information regarding the mismatching features in determining the activation of a trace in order to be able to account for list-strength effects. List-strength effects may be shown by comparing mixed lists composed of both strong and weak items, with pure lists consisting of only strong or only weak items. If there is a list-strength effect, the performance on the weak items in the pure weak list should be better than that on the weak items in the mixed list, and the performance on the strong items should be worse in the pure strong list compared to the mixed list. As shown in **Figure 4** (these results were obtained using a simulation program developed by David Huber), the REM model indeed predicts no decrease in recognition performance due to increasing strength of the other list items, although it does predict a decrease as a function of an increase in the number of other list items.

Equation [21] also suggests a similarity between REM and SAM in that the likelihood ratio for a particular trace in REM seems to play a similar role as the activation values in SAM. This suggests that it might be possible to generalize REM to recall paradigms by substituting the likelihood ratios for the activation values. This approach has the desirable feature that most, if not all, of the SAM recall predictions hold for REM as well. Diller et al. (2001) showed that this indeed produces a viable model for

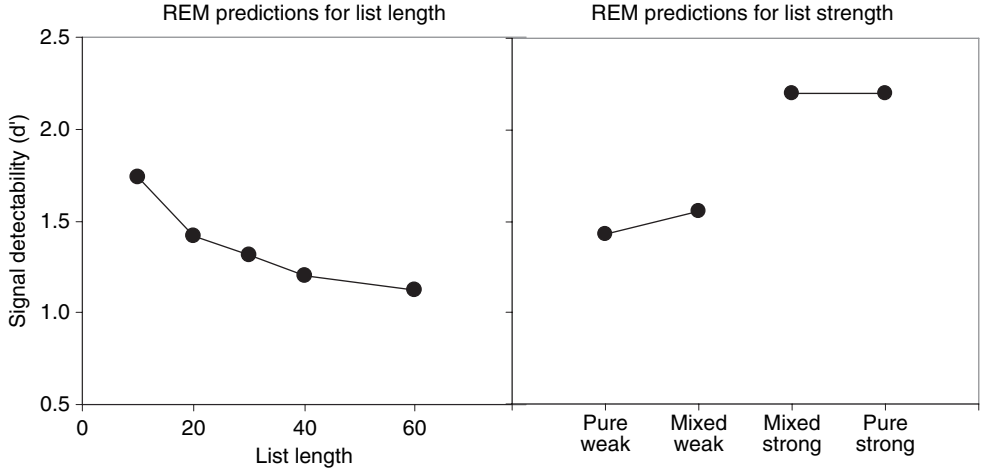

Figure 4 Predicted values for signal detectability (d′) as a function of list length (left panel) and list strength (right panel) according to the REM model (parameter values: g = 0.4, c = 0.7, u = 0.05; see Shiffrin RM and Steyvers M (1997) A model for recognition memory: REM: Retrieving effectively from memory. *Psychon. Bull. Rev.* 4: 145–166).

recall provided that one raises the likelihood ratios to a constant power. Thus, they defined the probability of sampling trace i as

$$P_S(I_i) = \frac{\lambda_i^{\gamma}}{\sum \lambda_k^{\gamma}} \quad [23]$$

Soon after the REM model for recognition was developed, it was realized that it might be fruitfully generalized to other domains, in particular semantic and implicit memory. In this more general version of REM, it is assumed that when an item is encountered, (a sample of) its features are stored in an episodic trace but also in a lexical/semantic system. Hence, the lexical/semantic trace accumulates information from all prior occurrences and is updated each time the item is presented (see Schooler et al., 2001).

Schooler et al. (2001) developed a REM-based model to account for priming effects in perceptual identification. The model gave a successful account of the results obtained by Ratcliff and McKoon (1997) in the forced-choice identification paradigm. In these experiments, a word (e.g., LIED) is briefly flashed and then masked. The subject is then presented with two alternatives (e.g., LIED and DIED) and has to choose which of these two was the word that was flashed. The critical result in this paradigm is that there is priming (i.e., an increase in the probability of choosing an item that was previously presented on a study list) but only when the two alternatives at the test are perceptually similar (LIED, DIED), but not when they are perceptually dissimilar (e.g., LIED, SOFA). Schooler et al. showed that this pattern of results can be explained in REM by the assumption that a small number of context features are added to the lexical/semantic trace of an item as a result of the prior presentation. These additional context features will obviously have a high probability of matching the later test context, hence will increase (although by a small amount) the number of matching features for the trace corresponding to the primed alternative. The crucial aspect in the REM explanation is that for similar alternatives the outcome of the feature match will often be the same, hence only a relatively small number of perceptual features will be relevant for the decision to choose one or the other alternative. As a result, the additional matches provided by the context features will have a larger effect when the alternatives are perceptually similar than when they are dissimilar.

To see this more clearly, **Figure 5** shows the distributions for the number of critical features for each of the choice alternatives that match the flashed item. Critical features are features that potentially can make a difference between the two alternatives. Since there are fewer critical features that differentiate similar alternatives, the number of matching critical features will also be lower. Assume that the foil item was presented on the prior study list and that this results in just one additional match due to context overlap between study and test. As shown in **Figure 5**, this additional match will have a clear effect for the similar alternatives: There is much more overlap between the distributions, and hence the probability that the target has more matches compared to the foil will decrease quite a bit. For the dissimilar alternatives, the added match due to context has only a small effect on the probability of choosing the target (the probability of a correct response). Hence, the effect of prior study will be much larger for the similar alternatives compared to the dissimilar ones.

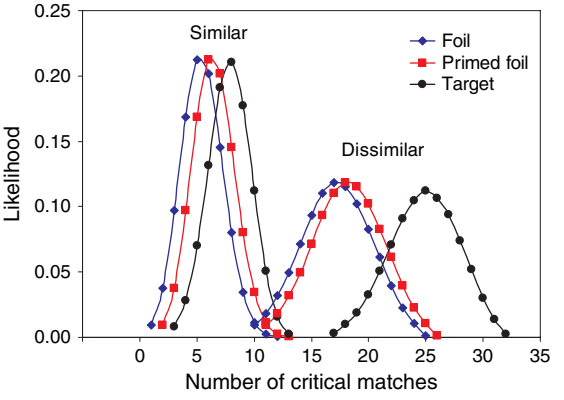

Figure 5 Predicted likelihood distribution for the number of critical matches for similar and dissimilar choice alternatives according to the REM model of Schooler LJ, Shiffrin RM, and Raaijmakers JGW (2001) A Bayesian model for implicit effects in perceptual identification. *Psychol. Rev.* 108: 257–272.

Wagenmakers et al. (2004) presented an application of REM to standard lexical decision tasks in which it was assumed that a lexical decision is based on the evaluation of the likelihood that the presented item corresponds to a word in the lexical system versus a nonword (just as a recognition decision is based on the evaluation that the test item corresponds to an item stored in the episodic system). There is a time-dependent encoding process such that as encoding time increases more and more probe features become available. The likelihood at time t is determined from the features available at

that time. The model was evaluated using signal-to-respond tasks and gave a good account for the effects of several major factors such as word frequency, repetition priming, and nonword lexicality.

Raaijmakers (2005) gives an outline of how the REM model may be extended to several other implicit and semantic memory paradigms such as associative priming, semantic categorization tasks, and associative repetition priming. A common feature of all of these applications is that the lexical/semantic system is assumed to be a much more flexible system than in many traditional accounts and that lexical/semantic traces do contain contextual features and hence are sensitive to recent episodes in which the item was presented.

2.25.4 Neural Network Approaches

All the approaches that I have discussed thus far do not make specific reference to how the processes that are postulated are actually implemented in the brain. The models in this section on the other hand take the analogy to neural processes in the brain as their starting point. It is assumed that information is distributed over sets of nodes in a neural network rather than as separate traces as in the models discussed thus far. Information is coded not in separate nodes or individual links but in the pattern of strengths over a large number of links or nodes. Hence, each individual node or link participates in the representation of many items or associations. Whenever a specific cue item is presented, the corresponding input nodes are activated and this activation is propagated through a network of links, leading to a specific pattern of activation at the output nodes and this pattern defines the output or the item retrieved from memory. The crucial property of these models (and the one that initially attracted the most attention) is that they provided a mechanistic account of the critical property that distinguishes human memory from other types of memory (such as hard disks), namely its associative character. That is, associative memory systems have the property that if the association *A-B* is stored, presentation of the cue *A* will automatically retrieve *B* without the need to know where *B* (or *A-B*) was stored. In models such as ACT and SAM, this property is assumed, but in neural network models, a computational account is given that generates the associative property, rather than assuming it.

To illustrate this, consider a very simple neural network model in which there is an input layer of neurons and an output layer of neurons and in which each input neuron is connected to each output neuron (e.g., Anderson et al., 1977). Items are represented by vectors, i.e., a series of activation values over the input or output neurons. In order to store the association *A-B*, the connections between the input *A* and the output *B* have to be modified in such a way that presenting *A* at the input side will produce *B* at the output side. This may be accomplished by modifying the connections between the *A* and *B* vectors in such a way that if the *i*-th value of *A* and the *j*-th value of *B* are both high, the connection is made stronger. More generally, if $\mathbf{f_i}$ is the feature vector for item *A* and $\mathbf{g_i}$ is the feature vector for item *B*, then the connections between the input nodes and the output nodes are increased by an amount equal to the product of the feature values. Using vector notation, this is equivalent to the assumption that the changes in the synaptic strengths are modified according to the matrix $\mathbf{M_i}$:

$$\mathbf{M}_i = \mathbf{f}_i \mathbf{g}_i' \quad [24]$$

Thus, if a list of such pairs is studied, the strengths are modified according to the matrix $\mathbf{M}$ with $\mathbf{M} = \sum \mathbf{M}_i$. Presenting an item as a cue to such a system amounts to postmultiplying the matrix $\mathbf{M}$ with the item vector. It is relatively easy to show that in the ideal case where all items vectors are uncorrelated and of unit length, such a model will show the associative property, i.e., on presentation of the item *A* ($\mathbf{f_i}$) the system will generate the associated item *B* ($\mathbf{g_i}$):

$$\mathbf{M}\mathbf{f}_i = \sum \mathbf{M}_j \mathbf{f}_i = \sum_{j \neq i} \left(\mathbf{g}_j \mathbf{f}_j'\right) \mathbf{f}_i + \left(\mathbf{g}_i \mathbf{f}_i'\right) \mathbf{f}_i = \mathbf{g}_i \quad [25]$$

The example given above is the simplest model of this kind and much more complicated models or networks have been proposed. All of these models, however, share the basic assumption that the associative information is encoded in the links or connections between the neurons. Item information is represented by the pattern or distribution of the activation values at the input and output layers. Note that the same nodes are used to represent all the items: The information is distributed over many nodes. Such models are therefore often called connectionist or distributed memory models. They may contain several layers of neurons with connections between successive layers (see Ackley et al., 1985; Rumelhart et al., 1986). Since the associative property that all of these models share may also be expressed as implying

that the model learns to predict the output vector given a specific input vector, it is not surprising that connectionist models have been developed not just to simulate human memory but also to compute any type of predictive relation between a specific input and specific output (i.e., associating a spoken output or phonemes based on the written input text, as in the NETtalk model; Sejnowski and Rosenberg, 1987). These more complex variants do not learn the associations in a single step (as in the simple model described earlier), but require several iterations in which the links between the nodes in the network are gradually changed. Basically what these models do is perform a kind of nonlinear regression using a least-squares fitting procedure to predict the output values given the input values.

Although these models have been quite successful in other domains, their success as a general framework for human memory is more limited. There are a number of features of these models that are problematic when they are used as models for episodic memory.

The most basic problem is known as catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990). This property is related to the fact that these models focus on extracting generalized rules from a series of exemplars rather than on storing individual items. The issue is that distributed memory models tend to forget all previously learned information on learning a new set of items. This is most clearly shown in the application of the back-propagation model to a retroactive interference experiment in which two lists are learned in succession (see McCloskey and Cohen, 1989). After learning the second list, humans will show some forgetting for the first list but the forgetting is far from complete. A typical back-propagation model, however, will show complete forgetting of the first list and in fact learning of the second list only starts after the first list has been completely unlearned. Such drastic forgetting is quite different from what is observed in experiments with humans, hence the name catastrophic forgetting. The basic reason for this incorrect prediction is that all the information is contained in the strengths or weights of the links in the network, and since these weights are freely adjusted during second-list learning to optimize second-list performance, there is nothing that prevents the complete forgetting of the first list information. Similar problems for recognition memory performance were demonstrated by Ratcliff (1990), who also showed that the model fails to predict a positive effect of amount of learning on the d′ measure for recognition.

It should be noted that these problems are not inherent to distributed memory models but seem to be limited to those connectionist models that assume that learning an item involves an optimization of the weights given to the links in order to tune the network to the information that it is currently being trained on. Murdock (1982, 1993), for example, developed a general framework for memory based on a distributed representation (TODAM, Theory of Distributed Associative Memory) in which item and associative information are added to a single memory vector (similar to the simple vector model described earlier) without any additional tuning. In TODAM, item information is simply added to the trace, while associative information (say the association A–B) is modeled by computing a vector that corresponds to the convolution of the vectors representing **A** and **B** (denoted as $\mathbf{A}^{*}\mathbf{B}$). Murdock showed that in such a model when A is presented as a cue, B (or at least a noisy version of the **B** vector) may be retrieved by computing the correlation of the **A** vector with the memory vector. TODAM does not suffer from the catastrophic forgetting problem presumably because a second list adds information (and hence noise) to the memory vector but does not destroy the information from the first list.

In order to prevent these problems in connectionist models, changes have to made to the basic structure of such models. One solution is to eliminate the strong version of the distributed memory assumption. For example, it might be assumed that there are a large number of nodes or connections and that learning a particular item or an association uses only a small proportion of these (e.g., so-called sparse distributed networks). Alternatively, it might be assumed that information concerning first-list learning continues to be stored in memory for a relatively long period after the learning of that list (a version of consolidation theory). In this way, the two lists become one list and a compromise is found between first- and second-list performance (see McClelland et al., 1995, for an ingenious version of this approach). Yet another approach is to relax the assumption that specific items are stored in a distributed manner, for example, competitive learning models using a winner-take-all principle in which retrieval results in a single unit are activated (retrieved) or a novelty-detection assumption that enables the system to allocate new items to units not already used to represent other items (e.g. Murre, 1992).

There are other problems that are not as easy to remediate in distributed memory models. For example, Shiffrin et al. (1990) showed that many network models have problems simultaneously predicting the presence of list-length effects and the absence of list-strength effects in recognition memory. Extra items harm performance by changing weights, but strengthening other items also changes the weights and should therefore cause similar harm. As yet, there is no clear solution for this problem within the framework of distributed memory models.

Despite these problems, neural network models continue to have a major influence on memory theories. These models have the advantage of a much closer link to neurobiological approaches and, more importantly, they still provide the only mechanistic explanation for the associative memory property. A nice example of a modern neural network model is the Complementary Learning Systems (CLS) approach proposed by McClelland et al. (1995) and further elaborated by O'Reilly and Rudy (2001) and Norman and O'Reilly (2003). The CLS approach is based on the realization that the memory system must combine two seemingly incompatible functions: Storage of episodic memories and integration of information to enable generalization. The first requires storage of specific, separated traces, whereas the second requires overlapping representations. The phenomenon of catastrophic forgetting shows that standard distributed representations are not a suitable model for episodic memory, although they do allow generalization. The solution in the CLS approach is to assume two separate but interactive systems: A rapidly changing system (assumed to be located in the hippocampal system) and a more slowly changing system (assumed to be cortical or neocortical). The hippocampal system is assumed to employ sparse compressed representations to minimize interference between traces, while the cortical system uses more standard distributed (overlapping) representations. It is assumed that there is a slow consolidation process that transfers information from the hippocampal to the cortical system. During recall, a cue will activate a corresponding pattern in the cortical system and if this pattern is sufficiently close to a stored hippocampal trace, the hippocampal system will settle on that trace, which then sends back activation to the cortical system, leading to the reinstatement of the original event pattern. Catastrophic interference in the neocortical system is avoided by a kind of consolidation process in which storage of new information is interleaved with renewed activation of older information.

McClelland et al. (1995) show how such a model may be used to explain a variety of findings from both human and animal experiments. For example, the fact that amnesic patients are unable to recall recent episodic experiences yet are able to recall older memories and do show implicit memory is attributed to a defect in the hippocampal system coupled with an intact cortical memory system. Norman and O'Reilly (2003) presented simulation results showing that the CLS model gives a good account of recognition memory. For example, the model may predict little or no list-strength effect in recognition if the recognition decision is mostly based on familiarity stemming from the cortical system (rather than on recall based on the hippocampal system). It is not clear, however, how the CLS model would handle both the absence of list-strength effects and the presence of list-length effects in recognition (see Norman and O'Reilly, 2003: 632).

However, even though these newer versions of connectionist modeling provide a solution for a number of the problems that plagued older connectionist models, there are several remaining issues. One is that it is not always clear which aspects of the model are responsible for a specific prediction. Although this is also a concern with other general modeling approaches, the issue is particularly relevant for these models. When a model successfully predicts a specific phenomenon, one also wants to know which aspects of the model are crucial for that prediction and which elements of the model (or the simulation) are incidental. For example, the model may employ a specific learning rule to optimize the weights or a specific equation for the decay of activation values. When one tries to understand why the model predicts the phenomenon, it is important to know whether it would still predict the phenomenon when a different learning rule or a different equation for decay (or perhaps no decay at all) is assumed. Thus, the ability to simulate a specific result does not yet mean that one has an explanation for that phenomenon (see also McCloskey, 1991; see O'Reilly and Farah, 1999, for a contrasting point of view). In many cases (for example, the prediction of the part-list cuing effect in SAM, see Raaijmakers and Phaf, 1999), a substantial amount of work is involved in figuring out why the model makes the prediction, but it is the additional work that ultimately leads to a model-independent explanation of the phenomenon. Such analyses are especially needed when it is difficult for other researchers to run the required model simulations.

2.25.5 Models for Serial Order Memory

In this section, I will discuss a number of models that have been proposed to account for memory for serial order information. Such models focus on explaining memory for item and order information in relatively short lists. For example, subjects might be presented with one or more lists of five items and then be given a test in which the items have to be recalled in the correct order, or they might be given the items at test (in a different order) and then asked to provide the correct order of presentation. The empirical evidence for (or against) these models is discussed by Healy and Bonk (*See* Chapter 2.05). We will restrict our discussion to the mathematical formulations that have been used.

A classic approach in this area is Estes' perturbation model (Estes, 1972). In this model, it was assumed that during study, items are associated or linked to their serial positions. However, during the retention interval, the item may shift (perturb) to a neighboring position. If one assumes that movements to an earlier or to a later position are equally likely, then the probability that an item occupies a particular position n at a given time t is given by the following difference equation:

$$P_{n,t} = (1-\theta)P_{n,t-1} + (\theta/2)P_{n-1,t-1} + (\theta/2)P_{n+1,t-1} \quad [26a]$$

For the endpoints we have a slightly different equation:

$$P_{1,t} = (1-\theta/2)P_{1,t-1} + (\theta/2)P_{2,t-1} \quad [26b]$$

for the first position and similarly for the final list position.

These relatively simple equations allow one to calculate the probability distribution for each item on the list. The model predicts better recall for items in the beginning and end positions than for items in the middle of the list since these items will have had less opportunity to perturb. Nairne (1992) obtained data for five-item lists at retention intervals of 30 s, 4 h, and 24 h and showed that the perturbation model gave a good quantitative account of the data. Note that in order to apply the model, one needs to estimate not just the perturbation parameter θ but also the number of cycles of perturbation (the number of times that eqn [26] is applied). It is easy to see that the model can also handle a number of other findings such as a higher accuracy if there are longer intervals between successive items (longer intervals will lead to less perturbation).

The perturbation model is an example of a bin model in which items are placed in or linked to serial positions rather than to one another. That is, a common view of serial order memory is that order memory is derived from item-to-item associations (the temporal order of a string such as ABCD is remembered through the pairwise associations A-B, B-C etc). What the perturbation model shows is that this type of view is not a necessary one and that an alternative view in which order information is not based on item-to-item associations but on memory for positional information can also give a good account of the data. However, a number of problems have been mentioned in the literature regarding such bin models, the most important one being that these models give no account of the recall of item information (cuing with a specific position automatically leads to recall of the linked item). In addition, it seems to be assumed that at test, the successive bins are always searched in the correct order (a rather strong assumption in the case of somewhat longer lists).

A prime example of a chaining model for serial order memory is the model proposed by Lewandowsky and Murdock (1989). Their model was based on the TODAM framework for memory, one of the distributed memory models discussed earlier. In this application of TODAM, it was assumed that recall starts by using a context cue to generate the first item, and then this item is used as a cue to generate the second item, and so on. A key problem for any type of chaining model is how to proceed if at a particular point no item is recalled. In TODAM, even though the retrieved vector may not enable the recall of a given item (the process of cleaning up the output vector via comparison to a lexicon may not succeed), the retrieved vector may still be used as a further cue.

Finally, Brown et al. (2000) developed a model for serial memory (termed OSCAR) that relies on contextual information to generate temporal information. In their model, context is represented as a series of oscillators that produce a dynamically changing state. The output from the oscillators forms a context vector. The model assumes that the overall context is made up of several such context vectors. During presentation of the list of items, each item vector is associated with the state of each context vector at the time of presentation. Thus, item 1 is associated to context vector 1 at time 1, context vector 2 at time 1, etc. Similarly, item 2 is associated to context vector 1 at time 2, context vector 2 at time 2, etc. All of the item–context associations for

each context vector are stored in an association matrix, similar to eqn [24]. At recall, the initial state of the context vectors is reinstated and these are then used to regenerate the context vectors at the following times. To recall the item that was presented at time *m*, context vector 1 at time *m* is multiplied with the memory matrix corresponding to context vector 1 (see eqn [25]), which produces an approximation to item *m*. Similarly, the context vector 2 is used in the same way, also leading to an approximation to item *m*, and so on for all context vectors. Finally, the item in a separately stored vocabulary of items that provides the best overall match to the various approximations of item *m* is then produced as the response. Thus, in this model, recall of a series of ordered items is based on the recall of gradually changing contexts that provide the temporal information for order memory. The OSCAR model is an example of a model for order recall that is based not on interitem associations but on the retrieval of temporal information that is specific to the time that a particular item was studied. The model provides a mechanism for how the system recalls the various contexts as well as the items that were presented. What is not clear, however, is how essential the specific formalization that Brown et al. (2000) used is for the predictions generated by OSCAR (e.g., which properties of the context vectors are essential, and are oscillators really required to enable the model to make these predictions).

2.25.6 Concluding Remarks

In the previous sections, I have presented an overview of several global frameworks for human memory. In this section, I return to the question raised in the introduction about what makes such models useful for understanding human memory processes.

Perhaps the most important advantage of having a formal model is that it makes it possible to prove that a specific argument or verbal explanation of a phenomenon is indeed valid (or the reverse: Show that it is not a valid argument). Many striking examples of such results may be found in the literature, for example:

- Batchelder's (1975) demonstration that the results from experiments on all-or-none learning could not be explained as being due to selection effects due to individual differences, as was thought by many proponents of theories in which learning was assumed to be more gradual.
- The demonstration by Hintzman and Ludlam (1980) that a purely exemplar-based classification model (MINERVA) could explain the finding that prototypical information seemed to be forgotten slower than the instances themselves. This finding had been generally interpreted as implying the existence of a prototype representation that was assumed to show a slower decay than the instance representations. The MINERVA model, however, did not contain any prototype representation and yet predicted the observed pattern of forgetting.
- The analysis of the part-list cuing paradigm using the SAM model (Raaijmakers and Shiffrin, 1981b) that showed that the lack of a positive cuing effect was entirely compatible with a model that was strongly based on the use of interitem associations. This analysis led to a new explanation for part-list cuing effects that we would not have thought of prior to running the analyses.

There are many such examples in the literature, and they do not necessarily have to be positive (in the sense of providing a new or alternative explanation). In some cases, computational analyses may show that a model fails to predict a finding that one would have intuitively thought that it should be able to predict. For example, the demonstration by McCloskey and Cohen (1989) of the catastrophic forgetting phenomenon shown by typical connectionist models had a big impact on the field. Similarly, Murdock and Lamon's (1988) demonstration that simple connectionist models failed to predict improved recognition performance with an increasing number of presentations was also initially met with disbelief.

What these examples show is that formal modeling may help to sharpen theoretical analyses by showing which results directly follow from a specific set of assumptions, which results cannot be predicted by the model, and which results may be predicted by the model but only under specific conditions (e.g., specific sets of parameter values). However, in order to be able to draw such conclusions, the modeler should not be content just to show that his or her model can predict the results of a particular set of experiments. This should be considered step one in the analyses and should be followed by additional analyses to determine the robustness of the prediction (does it vary in a qualitative sense when parameters are set to different values) as well as analyses to determine which aspects (assumptions) of the model are really crucial for the prediction. The latter aspect is often left out but is in my view the essence of the modeling approach: Models

should not be used as black boxes that in some mysterious way generate a specific pattern of data, but should preferably be used as analytical tools to assist the theoretical analysis of those data (what does it tell us about human memory processes?).

The latter point is related to the view that a model that is applied to a specific experimental paradigm is really a combination of (1) a set of core theoretical assumptions (the general theory), (2) a number of auxiliary assumptions related to the implementation of the model and specific computational aspects (e.g., an assumption that each trial adds the same amount of strength to a trace, or the specific learning rule used in a connectionist model), and (3) a set of task-specific assumptions (say a particular rehearsal strategy that is assumed or the rules that are used in generating an overt response based on the retrieved information). In this view, the ultimate goal of mathematical modeling is not simply fitting a set of data but to provide insight into the basic structure and processes in a particular domain. As such, there is no real difference with non-mathematical approaches. The basic advantage of the modeling approach is that it provides an analytical tool that can be used to experiment in a way that is not possible with verbally stated theories.

Viewed in this way, the progression of simple models that could only be applied to a single type of experiment to the more general approaches that we have discussed in this chapter is a major step toward a more coherent and comprehensive theory of learning and memory processes.

References

Ackley DH, Hinton GE, and Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. *Cogn. Sci.* 9: 147–169.

Anderson JA, Silverstein JW, Ritz SA, and Jones RS (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychol. Rev.* 84: 413–451.

Anderson JR (1976) *Language, Memory, and Thought*. Hillsdale, NJ: Erlbaum.

Anderson JR (1981) Interference: The relationship between response latency and response accuracy. *J. Exp. Psychol. Hum. Learn.* 7: 326–343.

Anderson JR (1983a) *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson JR (1983b) A spreading activation theory of memory. *J. Verb. Learn. Verb. Behav.* 22: 261–295.

Anderson JR and Schooler LJ (1991) Reflections of the environment in memory. *Psychol. Sci.* 2: 396–408.

Anderson JR, Bothell D, Lebiere C, and Matessa M (1998) An integrated theory of list memory. *J. Mem. Lang.* 38: 341–380.

Anderson JR, Qin Y, Sohn M-H, Stenger VA, and Carter CS (2003) An information-processing model of the BOLD response in symbol manipulation tasks. *Psychon. Bull. Rev.* 10: 241–261.

Anderson JR, Qin Y, Stenger VA, and Carter CS (2004) The relationship of three cortical regions to an information-processing model. *J. Cogn. Neurosci.* 16: 637–653.

Anderson JR, Albert MV, and Fincham JM (2005) Tracing problem solving in real time: fMRI analysis of the subject-paced Tower of Hanoi. *J. Cogn. Neurosci.* 17: 1261–1274.

Atkinson RC and Crothers EJ (1964) A comparison of paired associate learning models having different acquisition and retention axioms. *J. Math. Psychol.* 1: 285–315.

Atkinson RC and Juola JF (1974) Search and decision processes in recognition memory. In: Krantz DH, Atkinson RC, Luce RD, and Suppes P (eds.) *Contemporary Developments in Mathematical Psychology, Vol. 1: Learning, Memory, and Thinking*, pp. 242–293. San Francisco: Freeman.

Atkinson RC and Shiffrin RM (1968) Human memory: A proposed system and its control processes. In: Spence KW and Spence JT (eds.) *The Psychology of Learning and Motivation: Advances in Research and Theory,* vol. 2, pp. 89–105. San Diego: Academic Press.

Batchelder WH (1970) An all-or-none theory for learning on both the paired-associate and concept levels. *J. Math. Psychol.* 7: 97–117.

Batchelder WH (1975) Individual differences and the all-or-none vs incremental learning controversy. *J. Math. Psychol.* 12: 53–74.

Bjork RA (1966) Learning and short-term retention of paired associates in relation to specific sequences of interpresentation intervals. Technical Report no. 106. Institute for Mathematical Studies in Social Sciences, Stanford University, Palo Alto, CA.

Bower GH (1961) Application of a model to paired-associate learning. *Psychometrika* 26: 255–280.

Brown GDA, Preece T, and Hulme C (2000) Oscillator-based memory for serial order. *Psychol. Rev.* 107: 127–181.

Collins AM and Loftus EF (1975) A spreading-activation theory of semantic processing. *Psychol. Rev.* 82: 407–428.

Craik FIM and Lockhart RS (1972) Levels of processing: A framework for memory research. *J. Verb. Learn. Verb. Behav.* 11: 671–684.

Criss AH and McClelland JL (2006) Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *J. Mem. Lang.* 55: 447–460.

Diller DE, Nobel PA, and Shiffrin RM (2001) An ARC-REM model for accuracy and response time in recognition and cued recall. *J. Exp. Psychol. Learn.* 27: 414–435.

Eich JE (1980) The cue-dependent nature of state-dependent retrieval. *Mem. Cognit.* 8: 157–173.

Eich JE, Weingartner H., Stillman RC, and Gillin JC (1975) State-dependent accessibility of retrieval cues in the retention of a categorized list. *J. Verb. Learn. Verb. Behav.* 14: 408–417.

Estes WK (1955) Statistical theory of spontaneous recovery and regression. *Psychol. Rev.* 62: 145–154.

Estes WK (1972) An associative basis for coding and organization in memory. In: Melton AW and Martin E (eds.) *Coding Processes in Human Memory*, pp. 161–190. Washington, DC: Wiley.

Gillund G and Shiffrin RM (1984) A retrieval model for both recognition and recall. *Psychol. Rev.* 91: 1–67.

Godden DR and Baddeley AD (1975) Context-dependent memory in two natural environments: On land and underwater. *Br. J. Psychol.* 66: 325–331.

Grant HM, Bredahl LC, Clay J, et al. (1998) Context-dependent memory for meaningful material: Information for students. *Appl. Cogn. Psychol.* 12: 617–623.

Greeno JG (1968) Identifiability and statistical properties of two-stage learning with no successes in the initial stage. *Psychometrika* 33: 173–215.

Greeno JG (1974) Representation of learning as discrete transition in a finite state space. In: Krantz DH, Atkinson RC,

Luce RD, and Suppes P (eds.) *Contemporary Developments in Mathematical Psychology, Vol. 1: Learning, Memory, and Thinking*, pp. 1–24. San Francisco: Freeman.

Greeno JG and Scandura JM (1966) All-or-none transfer based on verbally mediated concepts. *J. Math. Psychol.* 3: 388–411.

Hintzman DL and Ludlam G (1980) Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Mem. Cognit.* 8: 378–382.

Howard MW and Kahana MJ (1999) Contextual variability and serial position effects in free recall. *J. Exp. Psychol. Learn.* 25: 923–941.

Humphreys M and Greeno JG (1970) Interpretation of the two-stage analysis of paired-associate memorizing. *J. Math. Psychol.* 7: 275–292.

James CT and Greeno JG (1970) Effect of A-B overtraining in A-BR. *J. Exp. Psychol.* 83: 107–111.

Kahana MJ (1996) Associative retrieval processes in free recall. *Mem. Cognit.* 24: 103–109.

Lee MD (2004) A Bayesian analysis of retention functions. *J. Math. Psychol.* 48: 310–321.

Lewandowsky S and Murdock BB (1989) Memory for serial order. *Psychol. Rev.* 96: 25–57.

McClelland JL and Chappell M (1998) Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychol. Rev.* 105: 724–760.

McClelland JL, McNaughton BL, and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102: 419–457.

McCloskey M (1990) Networks and theories: The place of connectionism in cognitive science. *Psychol. Sci.* 2: 387–395.

McCloskey M and Cohen NJ (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In: Bower GH (ed.) *The Psychology of Learning and Motivation,* vol. 24, pp. 109–164. New York: Academic Press.

Mensink GJ and Raaijmakers JGW (1988) A model for interference and forgetting. *Psychol. Rev.* 95: 434–455.

Mensink GJM and Raaijmakers JGW (1989) A model of contextual fluctuation. *J. Math. Psychol.* 33: 172–186.

Murdock BB (1982) A theory for the storage and retrieval of item and associative information. *Psychol. Rev.* 89: 609–626.

Murdock BB (1993) TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychol. Rev.* 100: 183–203.

Murdock B and Lamon M (1988) The replacement effect: Repeating some items while replacing others. *Mem. Cognit.* 16: 91–101.

Murre J (1992) *Learning and Categorization in Modular Neural Networks*. Hillsdale, NJ: Erlbaum.

Myung IJ and Pitt MA (2002) Model evaluation testing and selection. In: Lambert K and goldstone R (eds.) *The Handbook of Cognition*, pp. 422–436. London: Sage.

Nairne JS (1992) The loss of positional certainty in long-term memory. *Psychol. Sci.* 3: 199–202.

Nelson DL, McKinney VM, Gee NR, and Janczura GA (1998) Interpreting the influence of implicitly activated memories on recall and recognition. *Psychol. Rev.* 105: 299–324.

Norman KA and O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.* 110: 611–646.

O'Reilly RC and Farah MJ (1999) Simulation and explanation in neuropsychology and beyond. *Cogn. Neuropsychol.* 16: 49–72.

O'Reilly RC and Rudy JW (2001) Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychol. Rev.* 108: 311–345.

Pavlik PI and Anderson JR (2005) Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cogn. Sci.* 29: 559–586.

Polson PG (1972) A quantitative theory of the concept identification processes in the Hull paradigm. *J. Math. Psychol.* 9: 141–167.

Raaijmakers JGW (1979) *Retrieval from Long-Term Store: A General Theory and Mathematical Models.* PhD Thesis, University of Nijmegen, The Netherlands.

Raaijmakers JGW (1993) The story of the two-store model: Past criticisms, current status, and future directions. In: Meyer DE and Kornblum S (eds.) *Attention and Performance XIV: Synergies in Experimental Psychology, Artifical Intelligence, and Cognitive Neuroscience*, pp. 467–488. Cambridge, MA: MIT Press.

Raaijmakers JGW (2003) Spacing and repetition effects in human memory: Application of the SAM model. *Cogn. Sci.* 27: 431–452.

Raaijmakers JGW (2005) Modeling implicit and explicit memory. In: Izawa C and Ohta N (eds.) *Human Learning and Memory: Advances in Theory and Application*, pp. 85–105. Mahwah, NJ: Erlbaum.

Raaijmakers JGW and Phaf RH (1999) Part-list cuing revisited: A test of the SAM explanation. In: Izawa C (ed.) *On Memory: Evolution Progress and Reflection on the 30th Anniversary of the Atkinson-Shiffrin Model*, pp. 87–104. Mahwah, NJ: Erlbaum.

Raaijmakers JGW and Shiffrin RM (1980) SAM: A theory of probabilistic search of associative memory. In: Bower GH (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory,* vol. 14, pp. 207–262. New York: Academic Press.

Raaijmakers JGW and Shiffrin RM (1981a) Order effects in recall. In: Long JB and Baddeley AD (eds.) *Attention and Performance IX*, pp. 403–415. Hillsdale, NJ: Erlbaum.

Raaijmakers JGW and Shiffrin RM (1981b) Search of associative memory. *Psychol. Rev.* 88: 93–134.

Ratcliff R (1990) Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychol. Rev.* 97: 285–308.

Ratcliff R and McKoon G (1981) Does activation really spread? *Psychol. Rev.* 88: 454–457.

Ratcliff R and McKoon G (1997) A counter model for implicit priming in perceptual word identification. *Psychol. Rev.* 104: 319–343.

Roberts WA (1972) Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *J. Exp. Psychol.* 92: 365–372.

Rumelhart DE (1967) The effects of interpresentation intervals on performance in a continuous paired-associate task. Technical Report 16. Institute for Mathematical Studies in Social Sciences, Palo Alto, CA, Stanford University.

Rumelhart DE, Hinton GE, and Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE and McClelland JL (eds.) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1: Foundations*, pp. 318–362. Cambridge, MA: MIT Press.

Schooler LJ, Shiffrin RM, and Raaijmakers JGW (2001) A Bayesian model for implicit effects in perceptual identification. *Psychol. Rev.* 108: 257–272.

Sejnowski TJ and Rosenberg CR (1987) Parallel networks that learn to pronounce English text. *Complex Systems* 1: 145–168.

Shiffrin RM (1968) Search and retrieval processes in long-term memory. Technical Report 137. California: Institute for Mathematical Studies in the Social Sciences, Palo Alto, CA, Stanford University.

Shiffrin RM (1970) Memory search. In: Norman DA (ed.) *Models of Human Memory*, pp. 375–447. New York: Academic Press.

Shiffrin RM and Atkinson RC (1969) Storage and retrieval processes in long-term memory. *Psychol. Rev.* 76: 179–193.

Shiffrin RM and Steyvers M (1997) A model for recognition memory: REM: Retrieving effectively from memory. *Psychon. Bull. Rev.* 4: 145–166.

Shiffrin RM, Ratcliff R, and Clark S (1990) The list-strength effect: II. Theoretical mechanisms. *J. Exp. Psychol. Learn.* 16: 179–195.

Sirotin YB, Kimball DR, and Kahana MJ (2005) Going beyond a single list: Modeling the effects of prior experience on episodic free recall. *Psychon. Bull. Rev.* 12: 787–805.

Smith SM (1979) Remembering in and out of context. *J. Exp. Psychol. Hum. Learn.* 5: 460–471.

Underwood BJ and Schulz RW (1960) *Meaningfulness and Verbal Learning*. Chicago: Lippincott.

Wagenmakers EJM, Steyvers M, Raaijmakers JGW, Shiffrin RM, Van Rijn H, and Zeelenberg R (2004) A model for evidence accumulation in the lexical decision task. *Cogn. Psychol.* 48: 332–367.

Wixted JT and Ebbesen EB (1991) On the form of forgetting. *Psychol. Sci.* 2: 409–415.

Young JL (1971) Reinforcement-test intervals in paired-associate learning. *J. Math. Psychol.* 8: 58–81.