15 Using ROC Data and Priming Results to Test Global Memory Models

Roger Ratcliff and Gail McKoon
Northwestern University

INTRODUCTION

In this chapter, we summarize two lines of work that relate closely to Ben Murdock's interests and to his approach to research. The first line of work uses ROC curves to test fundamental properties of the recently proposed global memory models (Gillund & Shiffrin, 1984; Hintzman, 1986, 1988; Murdock, 1982, 1983). The second uses empirical results to support a compound cue interpretation of priming phenomena (implemented in global memory models, Ratcliff & McKoon, 1988). While these two lines are quite independent, they form a nice contrast between two of the different kinds of research issues that arise from theoretical modeling, one concerning the search for critical tests with which to evaluate complicated and non-intuitive models (cf. Hintzman, this volume), and the other concerning the role of models in guiding and developing new explanations of known phenomena.

The global memory models have been designed to provide a comprehensive account of a number of memory phenomena. Their range of coverage is impressive, and they attain this coverage with relatively few degrees of freedom. The models were developed in part as a response to criticisms of the limited applications of earlier models. What has taken a back seat during this development effort has been the testing of fundamental assumptions. One reason for this is that coming up with tests of basic assumptions is by no means a simple task. A large element of luck is involved in discovering a potential test, and in finding that the test is truly constraining and not subject to variations in parameter values. The specific test we present in this chapter is a test of the models' accounts of recognition memory, and uses ROC curves to determine the relative variances of the signal and noise distributions for old and new test items.

The second line of research described in this chapter is an empirical investigation of priming phenomena. The issue is whether priming effects can be obtained only between concepts that are strong direct associates, as would be claimed by some of the global memory models, or whether priming can also be obtained between concepts that are not so strongly connected. Initial data (de Groot, 1983; Balota & Lorch, 1986) showed priming in lexical decision only between high associates and not between more weakly associated words. However, later data (McNamara & Altarriba, 1988), with a slightly different procedure, did show priming between weakly associated words. McNamara and Altarriba (1988) argued that this weak associative priming effect came about because of a "mediating" concept between the prime and target. For example, *beach* primes *box* because of the mediating concept *sand*. Such mediated priming is inconsistent with the global memory models' explanations of priming, and so seems to indicate that the models are wrong. We describe new data that shows that what has been labeled "mediated priming" is actually the result of weak *direct* associations, and not mediating concepts.

ROC CURVES

It has long been known that in signal detection analysis, the relative standard deviations of the signal and noise distributions can be obtained easily if the two distributions are normal. When hit rate and false alarm rate are transformed to z-scores and plotted against each other, then the resulting ROC curves are straight lines with a slope that is equal to the ratio of the noise standard deviation to the signal standard deviation, σ_n/σ_s, and an intercept that is equal to the ratio of the mean of the signal distribution to the standard deviation of the signal distribution, μ_s/σ_s (setting the mean of the noise distribution to zero).

The global memory models make strong predictions about the behavior of the signal and noise distributions for old and new items. First, each of the models makes assumptions that entail normally distributed old and new item distributions, which in turn leads to a prediction of linearity for z-transformed ROC curves. Second, the models make strong, but different, predictions about the ratio of the standard deviations (SD) of the signal and noise distributions. Murdock's TODAM model predicts that the SD in the noise distribution is about the same as the SD for the signal distribution, so that the ratio should be about 1.0. In contrast, Hintzman's MINERVA 2 model and Gillund and Shiffrin's SAM model predict that the SD of the noise distribution is smaller than the SD of the signal distribution, so that the ratio should be less than 1.0, and also that the ratio should decrease as the signal strength increases.

Because the models make such strong predictions about the ratio of the standard deviations of signal and noise distributions, the slopes of empirical ROC curves provide tests of the models. ROC curves can be obtained from experiments which manipulate the response criterion used by subjects to make their recognition decision. The criterion is manipulated either by varying old/new test item probabilities or by using confidence judgments. If the resulting z-transformed ROC curves are linear, then the data are consistent with the assumption of normal

distributions, and so the slope of the curves can be used to test whether the ratio of the signal and noise standard deviations has a value that would be predicted by one or another of the models. We describe experiments that provide this test later.

For these experiments, there is another set of issues that concerns the "list strength effect." "List strength" refers to the effect on the strength of an item in memory due to the other items in its list. In a typical experiment, three kinds of lists are presented to subjects: A "pure weak" list in which all items are intended to have weak strength either because they are presented for only a short time or for only a few repetitions; a "pure strong" list in which all items are intended to have high strength because their presentation time is long or their number of repetitions is large; and a "mixed" list in which half the items are weak and half strong. Strong items should, of course, be recognized better than weak items. The prediction of the global memory models concerns how much better. The models predict that the difference between weak and strong items in a mixed list will be larger than the difference between weak and strong items in the pure lists. This is because in the mixed list, there is only one new item distribution (with one standard deviation) for both weak and strong items, whereas in the pure lists, the standard deviation for new items increases as a function of strength of the old items. Because of the increase in standard deviation for new items in pure lists, the increase in strength for the strong items in these lists should be attenuated relative to the increase in the mixed list. However, in contradiction to this prediction, data from a number of experiments show that the pure and mixed list conditions provide about the same difference in strength (as measured by d') between weak and strong items (Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990). One interpretation of this data is that increasing the strength of old items in a pure list does not increase the standard deviation of new item familiarity. Hence, the mixed-pure list experiments, like ROC curves, give a way of examining the behavior of old and new item standard deviations as a function of strength of the items. In total, the mixed-pure experiments and the manipulations of response probabilities and confidence judgments can provide a relatively comprehensive picture of how variance in item strength behaves as a function of strength.

DESCRIPTION OF THE EXPERIMENTS

Ratcliff, Sheu, and Gronlund (1991) report five experiments that examined ROC curves in mixed list/pure list designs. Four of these experiments used presentation time per item to vary the strength of items and the fifth used number of repetitions per item. ROC curves were produced in two ways: by manipulating the probabilities of old and new test items in the test list in order to alter the criterion setting, and by using confidence judgments to provide hit and false alarm rates at different levels of confidence.

Experiments 1 and 2 varied criterion by means of manipulating the probabilities of old and new test items in the test list. Experiment 1 used 16 pairs of words for study and 48 test words, varying from probabilities of 4 old to 1 new, to 1 old to 4 new. Strength was varied by presentation time per item: 1 s per pair for weak items and 5 s per pair for strong items. Experiment 2 varied strength of items by repetition. Weak items were presented once and strong items five times, so that

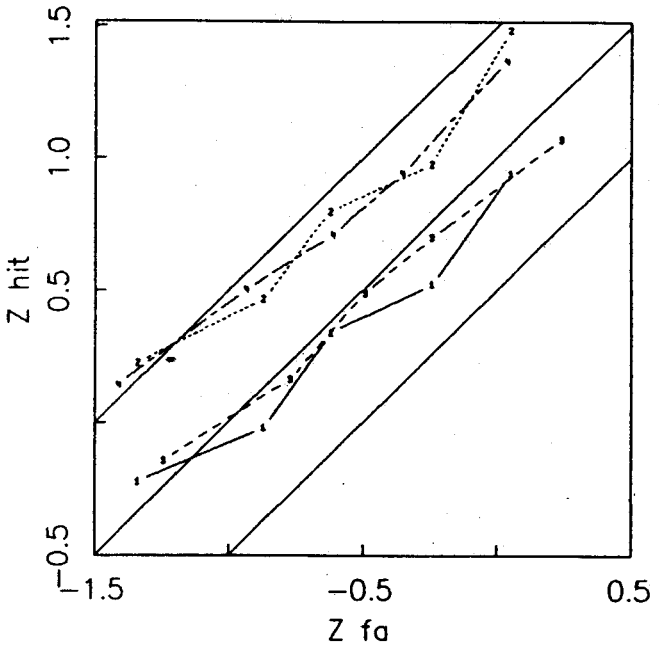

FIGURE 15.1 z-transformed ROC curves for Experiment 1 (presentation time manipulations and probability of old and new test items varied). The curves represent mixed strong, pure strong, pure weak, and mixed weak, reading top to bottom on the left hand data points.

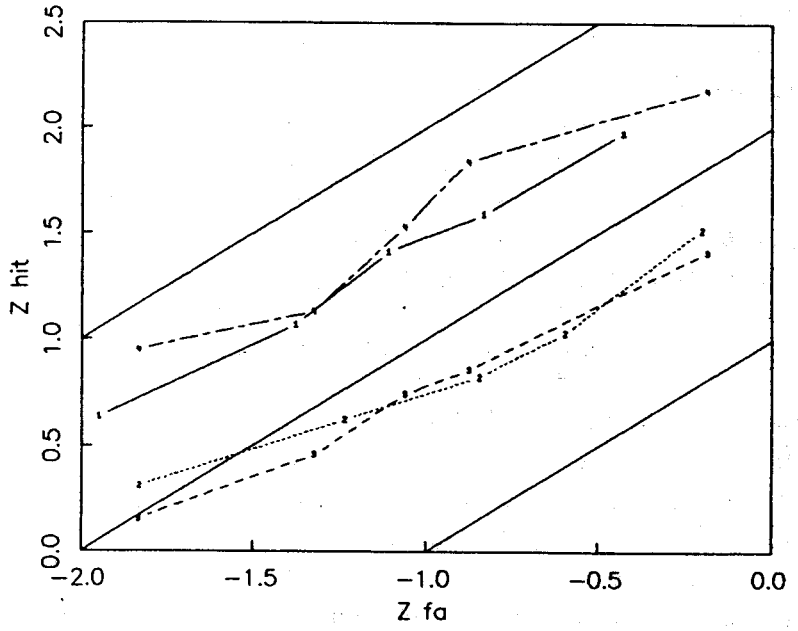

FIGURE 15.2 z-transformed ROC curves for Experiment 2 (repetition manipulation and probability of old and new test items varied). The curves represent mixed strong, pure strong, pure weak, and mixed weak, reading top to bottom on the left hand data points. Conditionalized data means tests were performed on items from the same range of study positions for the long (5 repetitions) and shorter lists.

study lists varied in length, with weak lists of 20 items, mixed lists of 60 items, and strong lists of 120 items. Test lists were 24 items long, and varied in probability from 5 to 1 old to new, to 1 to 5 old to new. Experiments 3, 4, and 5 were variants on Experiment 1 but used confidence judgments to construct the ROC curves. Responses were made on a six point scale using the *x* through *m* keys on a CRT keyboard. There has been some discussion of possible differences between estimates of ROC curves based on the two methods (confidence judgments and varying old/new probabilities in the test list). While the suggested differences (Markowitz & Swets, 1967) probably are not relevant to recognition (they had to do with the relative familiarities of different auditory tests), it is still worth using the two different but converging methods.

The variable that distinguished Experiments 3, 4, and 5 was presentation time per item. For Experiment 5, pairs were presented for study at 1 s and at 5 s per pair, in order to replicate Experiment 1 with the confidence judgment procedure. Experiment 4 used single words during the study phase instead of pairs, and presentation time was 50 ms per item for weak items and 200 ms per item for strong items. Experiment 5 was the same as Experiment 4 but with study times of 100 ms and 400 ms per item.

The motivation for the rapid presentation rate was a result reported by Yonelinas, Hockley, and Murdock (1990). Using rapid presentation rates, they found a larger difference between d′ values in mixed lists than in pure lists, which counters the generality of the results of Ratcliff, Clark, and Shiffrin (1990) who had found no such effect at slower rates. One difference between the studies of Yonelinas et al. (1990) and Ratcliff, Sheu, and Gronlund (1991) was that the latter used blocked study lists, whereas Yonelinas et al. used study lists in which items with the two presentation rates were randomly intermixed. For the Ratcliff, Sheu, and Gronlund (1991) studies, the blocked design was used to minimize rehearsal redistribution (i.e., rehearsal of weak items during presentation of strong items). However, Yonelinas et al. (1990) point out that at presentation times of a hundred ms or faster, rehearsal redistribution is unlikely if not impossible. On the other hand, there is another possible problem with intermixed presentation rates when the fast presentation rate is as fast as 50 ms per item: There may be inverse rehearsal redistribution. It might be that in a mixed list, long items are rehearsed during the presentation of subsequent short items, so that the short items are not processed at all. This would lead to a larger mixed list difference than pure list difference between strong and weak items. To examine an inverse rehearsal redistribution as a possible explanation of the mixed/pure difference in the Yonelinas et al. (1990) experiments, Experiments 3 and 4 used a blocked design.

RESULTS

The main results are shown in Figures 15.1 through 15.5 (details of the analyses are presented in Ratcliff, Sheu, & Gronlund, 1991). First, the figures show that the ROC curves are mainly linear. This is consistent with the global memory models' assumption of normality for the old and new item familiarity distributions. Second, Figure 15.6 shows the values of slope for the ROC curves for the five experiments as a function of the intercept of the ROC curve. This represents the ratio σ_n/σ_s (the

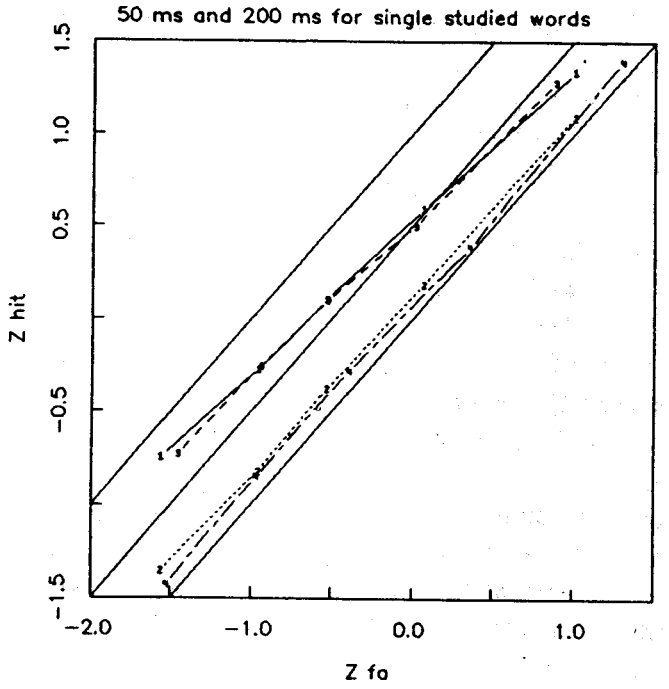

FIGURE 15.3 z-transformed ROC curves for Experiment 3 (presentation time manipulations and confidence judgments). The curves represent mixed strong, pure strong, mixed weak, pure weak, reading from top to bottom on the left hand data points.

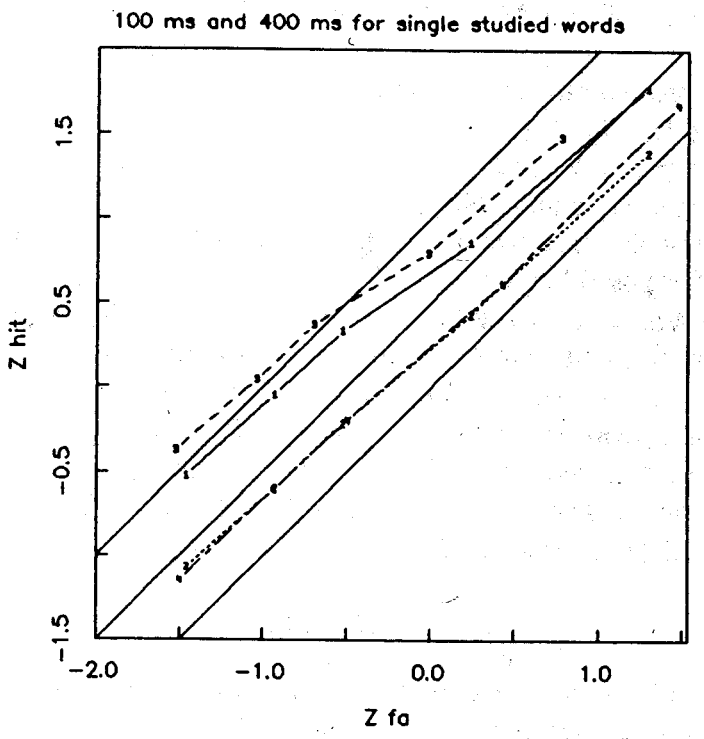

FIGURE 15.4 z-transformed ROC curves for Experiment 4 (presentation time manipulations and confidence judgments). The curves represent pure strong, mixed strong, mixed weak, pure weak, reading from top to bottom on the left hand data points.

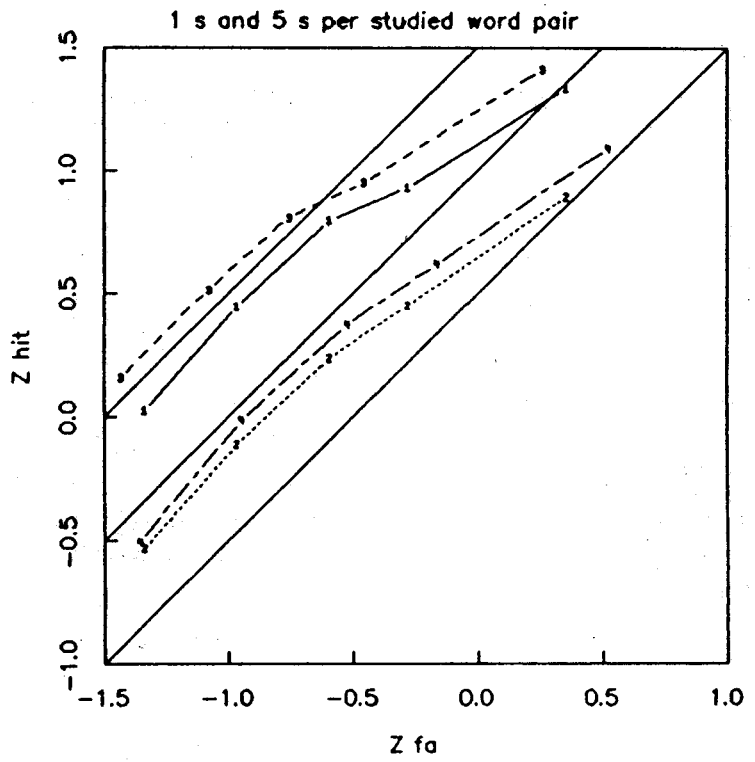

FIGURE 15.5 z-transformed ROC curves for Experiment 5 (presentation time manipulations and confidence judgments). The curves represent pure strong, mixed strong, pure weak, mixed weak, reading top to bottom on the left hand data points.

ratio of standard deviations) and μ_S/σ_S, closely related to strength or familiarity of the item. Given that σ_n/σ_S is constant at 0.8 except at the lowest values, the result would be the same for other definitions of d′ (in terms of the noise distribution or in terms of a pooled standard deviation).

The result that the ratio of standard deviations is constant at 0.8 as a function of old item strength causes problems for all the global memory models. The predictions of the current incarnations of the models are inconsistent with the data. The challenge then is to see if there is any way to modify the models to handle this result without altering the ability of the models to account for the range of other data to which they have been applied.

The third result of note is the mixed/pure list difference in d′. Strictly speaking, there is no invariant measure of d′ except when the signal and noise distributions have equal variance. When the variances are unequal, standard practice is to use either the difference in means divided by the noise, signal, or a pooled standard deviation (when these can be estimated from ROC curves). When the data provide only a single hit and false alarm rate, there is no way of estimating d′. (A discussion about the interpretation of d′ in this situation led to this research.) The estimate of d′ will be different at different criterion settings for the same sensitivity (see McNicol, 1972). Estimates of the ratio of the ratio of mixed strong/mixed weak to pure strong/pure weak are shown in Table 15.1. In all these ratios, there is no hint

Table 15.1

Ratio of Mixed Strong/Mixed Weak to Pure Strong/Pure Weak d' Values

	Experiment				
	1	2	3	4	5
Strength Ratio	5:1	5:1	5:1	4:1	4:1
Ratio of Ratios	1.04	1.21	.99	.69	.84

of a mixed/pure difference (the 1.21 value for Experiment 2 is reduced to 1.13 if the analyses are performed on individual subjects). This both replicates Ratcliff, Clark, and Shiffrin (1990) and fails to replicate the results of Yonelinas et al. (1990). Thus the arguments of Ratcliff, Clark, and Shiffrin (1990) carry through to analyses based on full ROC curves. (For further arguments against rehearsal redistribution, see Murnane & Shiffrin, 1990.)

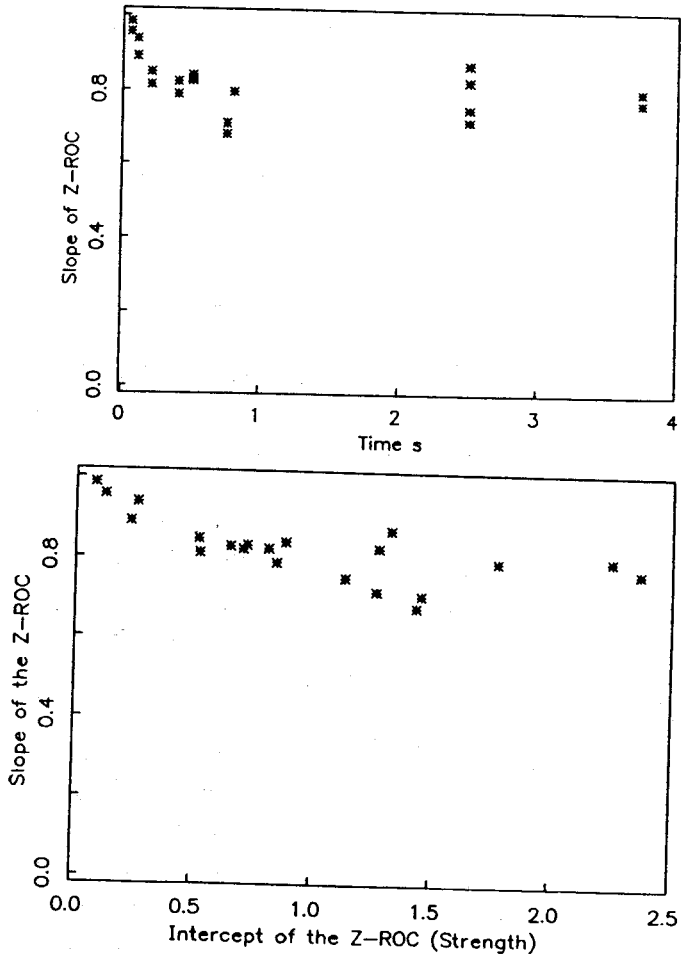

FIGURE 15.6 Ratio of noise standard deviation to signal standard deviation as a function of presentation time per item (bottom panel) and as a function of item strength (top panel). The data from Experiments 1-5 and results from Murdock & Dufty (1972) are included.

In the following sections, we review each model and its predictions for the ratio of standard deviations for old and new item familiarities. The details of the models can be found in the original papers, and details of the predictions in Ratcliff, Sheu, and Gronlund (1991).

TODAM

Murdock's TODAM model assumes that items are vectors of attributes, and that all studied items are stored in a common memory vector. For item recognition, a test item vector is matched against the memory vector by taking the dot product of the two vectors. Each attribute is assumed to be derived from a normal distribution with mean zero and standard deviation 1/N where the dimensionality of the vector is N. To illustrate why the model predicts that the signal and noise variances should be about equal, the parameters of the model are first set to simple values; that is, no forgetting, no probabilistic encoding, attention weight set to 1, and independent vectors. Then, each study item vector (in the common memory vector) that does not match the test item vector contributes 1/N to the variance of the test item, while each study item that does match contributes 2/N to the variance (Murdock, 1982; Weber, 1988). Hence, for study lists of length 32, the difference between the match and nonmatch variances will be 33/N vs. 32/N; the standard deviations are about equal.

MINERVA 2

Hintzman's MINERVA 2 model also uses a vector representation but instead of pooling the study item vectors into a common memory vector, each is kept separate. The results of matching a recognition test item against each study vector are pooled at retrieval. The model assumes that the elements of a vector are either -1, 0, or +1, and elements of a vector are encoded into memory with some probability, p. At test, the dot products between each encoded study vector in memory and the test item vector are obtained (with the dot product normalized by the vector length minus the number of times there is a zero in both memory and test vector in the same position). The values of these dot products are cubed and summed to give the "echo intensity" (familiarity or strength) on which the recognition decision is based. The prediction of the model that the variance of old test items is much larger than the variance of new test items arises because of the cubing operation. New items have a mean echo intensity of zero and the cubing operation shrinks the overall variance of the dot product. However, for old items, the matching comparison produces a positive dot product and the cubing operation stretches this scale (see also Sheu, 1990). This leads to the prediction that the standard deviation for old items will be considerably greater than that for new items.

SAM

Gillund and Shiffrin's SAM model assumes that each study item is represented separately in memory. For encoding, a simple buffer model builds a retrieval structure which takes the form of a matrix of connections between cues (test items) and images (representations of items in memory). The parameters of this encoding

process are expressed in terms of units per second, and measure the strength that accumulates per second. The parameter a refers to context strength, b to interitem strength (connections between all items in the buffer accumulate b units of strength per second), c to self strength, and d to residual strength; that is, strength existing prior to the experiment. Variability is added through a parameter v so that the value actually placed in the retrieval structure is either $(1\text{-}v)$, 1, or $(1+v)$ times the value computed from the encoding process, with probability 1/3 attached to each of these values. For recognition, the cue set (the item cue plus the context cue) probes memory to produce some value of familiarity. The value is calculated as the sum over items in memory of the product of the strength of the cue-to-item association. Recall involves probabilistic sampling and recovery in a search process. With these parameters, structures, and processes, the model accommodates a wide range of experimental data from recall and recognition experiments. Because the model assumes that the variance in strength values increases as the strength increases, the model predicts that the variance of familiarity for old items increases as the strength of the old items increases. This means that the standard deviation for old items will increase with presentation time or number of repetitions. The expression for the ratio of variances of noise to signal can be computed for a case analogous to Experiment 1: Assuming that a study list has 16 pairs of words, that interitem rehearsal occurs only between the two members of the pair (i.e., interitem strengths are only incremented between members of the pair), and that single items are tested, the expression for the ratio of noise to signal variances for weak and strong items in a pure list is: $32d / (32d + bt + ct)$, where t is the time the item is in the rehearsal buffer.

This expression indicates that the ratio of standard deviations cannot be 0.8 for both weak and strong items. Thus the model in its current form cannot fit the data reported earlier.

Differentiation Variant of SAM

In order to deal with the results from the mixed list/pure list experiments (the list strength effect), Shiffrin et al. (1990) introduced a new version of the SAM model. In this version, it is assumed that the better encoded an item, the more differentiated it is from other items in memory. Thus, instead of the residual strength of a test item to an image remaining constant as in the original SAM model, it decreases as a function of the strength with which the image is encoded into memory. This can be quantified with the assumption that the residual strength is an inverse function of context strength ($d = k / (at)$, where k is a constant). With this assumption, the variance in the new item distribution is independent of strength of old items, leading to the prediction that the difference between d′ values for weak and strong items will be the same in mixed and pure lists (i.e., no list strength effect; Shiffrin et al., 1990).

Using the expression $d = k / (at)$, the expression for the ratio of variances becomes: $32k / (32k + abt^2 + act^2)$.

This expression does not allow the ratio of standard deviations to be constant as a function of familiarity of old items. Thus the new version of SAM cannot deal adequately with the ratio of standard deviations derived from the ROC curves.

Other Models

There are other memory models such as the matrix model of Pike (1984) and the convolution model of Eich (1982). Both of these models have the same kind of structure as TODAM, with items stored in a common memory vector. To calculate the match values for old and new items, it is necessary to add contributions from matches and nonmatches as in TODAM, and for lists of, for example, 32 items, the nonmatch contribution dominates, so that the ratio of variances is near 1. Again, this is inconsistent with the data.

Ratcliff, Sheu, and Gronlund (1991) reviewed two connectionist models, Carpenter and Grossberg's (1987) ART1 model and a backpropagation-based encoder model (see Ratcliff, 1990). These models have various problems with the variance ratio. ART1 has an architecture that maps from a distributed representation at input to a local representation. Ratcliff, Sheu, and Gronlund (1991) examined several possible decision rules for recognition. For example, one decision rule was based on the activity of the most active local node. The problem was that ART1 predicts that the variance in the most active top level node becomes very small as a function of repetition of items. This leads to very small variances for old items after about four repetitions of an item, while the variance for new items is relatively large. Thus, the model's predictions conflict with the data.

The multilayer backpropagation encoder model discussed by Ratcliff(1990) predicts that the variance of old items is smaller than the variance of new items. In part, this is because of a scaling problem (the better the match, the nearer its value is to 1.0, and the smaller the variance by the binomial theorem). The second factor is that items that are trained are more uniformly learned than new items that might have a lot or relatively little in common with the test item. Modifications to the encoder model by Kortge (1990) that overcome some of the forgetting problems noted in Ratcliff (1990) still produce this inverse prediction about the old and new item variances.

COMPOUND CUE MODELS AND MEDIATED PRIMING

The second line of work we present in this chapter concerns application of the global memory models to priming. Ratcliff and McKoon (1988) proposed a compound cue model to account for priming phenomena (see also Dosher & Rosedale, 1989). In this model, a prime and target form a compound, and memory is probed with the compound. If the prime and target are associated with each other in memory, then the match of their compound against information in memory (the compound's "familiarity") is better than if the prime and target are not associated. A high degree of familiarity for the compound leads to facilitation ("priming") of responses to the target.

In MINERVA 2 and TODAM, the compound of a prime and target would be formed by placing the two items in a common vector and probing memory with that vector. In MINERVA 2, if the joint vector matches a memory vector, the match value is cubed which will add a much greater increment to echo intensity than if the prime and target match different, not joint, memory vectors. In TODAM, the

convolution of the two vectors would be formed, and if this compound is stored in memory, the degree of match between the compound and memory is increased. In the SAM model, compounding is already built into the model for decisions about pairs of items. The idea is that if a prime and target are connected to the same image (or images) in memory (i.e., they share an associate), then the prime and target match memory with a high degree of familiarity that is the result of multiplying two large values of strength together (see Ratcliff & McKoon, 1988). Thus, all the implementations of the compound cue model make strong predictions about the range of priming: if items are directly connected in memory, then MINERVA 2, TODAM, and SAM predict priming; if items are not directly connected but share a common associate, then SAM predicts priming (but less than if they are directly connected). All other possibilities will not produce priming, according to these models.

McNamara and Altarriba (1988) examined these predictions using a lexical decision task to look for indirect priming. Previous work by de Groot (1983) and by Balota and Lorch (1986) had found that direct associates gave priming in lexical decision, but mediated associates (e.g., *beach-box*, mediated by *sand*) did not. However, by changing the experimental procedure, McNamara and Altarriba were able to obtain mediated priming. In one experiment, they used a double word lexical decision task with only mediated pairs (no highly associated pairs) in the test list, and in a second experiment, they used a word by word lexical decision task with high associates included. In both experiments, they found mediated priming. They argued that this result ruled out the compound cue model for priming (at least in its TODAM and MINERVA 2 implementations). Furthermore, they argued that this result was consistent with spreading activation theories of priming, by which activation would spread from the prime through the mediator to the target.

Our claims are that there is a counter explanation for McNamara and Altarriba's results, and that the compound cue model is correct (so notions of activation spreading through semantic networks can be discarded). The counter explanation is that so-called mediated priming comes about instead from weak direct associations. The problem then is to operationally define the pairs of words that will give priming effects. By spreading activation theories, the pairs of words are those that are either directly connected in memory or connected through a mediator. By the compound cue account, the pairs are those that are directly connected so as to have sufficiently high familiarity as a compound.

The way that Balota and Lorch, de Groot, and McNamara and Altarriba defined mediation was through a free association production task. They reasoned that if concept A produces concept B with high probability, and B produces C with high probability, but A never produces C directly, then A and C are "mediated" and not directly related. The pairs of words used in the experiments demonstrating mediated priming were defined in this way. However, the pairs of words can also be described in another way: the A and C concepts are more related to each other than they are to randomly chosen other words (as measured by asking subjects to rate how "related" are the A and C words). This relatedness suggests that it might be their familiarity as a compound that is giving rise to a priming effect.

To test the compound cue hypothesis against spreading activation accounts, we used pairs of words with the same relatedness as the mediated pairs used in the earlier studies, but with no mediator between the A and C concepts; that is, no mediator between prime and target as measured by the same production task as was used by Balota and Lorch. Then, these pairs were tested in an experiment along with the previously used mediated pairs. The experiment, therefore, allowed degree of mediation to be varied from very high to very low, with relatedness kept constant. If priming is due to activation spreading through links from mediator to mediator, then there should be priming for the mediated pairs but not for the non-mediated pairs. However, if priming is due to the degree of familiarity that arises from direct associations between the words of the pairs, then the amount of priming should be the same for the two kinds of pairs.

The non-mediated word pairs were based on the mediated pairs used by Balota and Lorch (1986), and McNamara and Altarriba (1988). Each non-mediated pair had the same prime as one of the mediated pairs but a new target. The new target was picked to be intuitively about as related to the prime as the old target was, and ratings collected from subjects confirmed that the relatedness values of the mediated and non-mediated pairs were equivalent. The new targets were also chosen so that there was no mediator between prime and target that we could imagine. To verify that there was no mediator, we collected free associations in the same kind of production task that was originally used to produce the mediated pairs. One group of subjects was given the prime of each pair and instructed to give the first eight associates that came to mind. Another group of subjects was given the top four associates that were produced for each prime, and asked to generate four associates from these words. Finally, a third group of subjects was given the mediator for the original prime-target pair, and asked to generate four associates to this word. For all of the pairs, the new target was produced either as direct associate or as an associate of an associate with essentially zero probability. Thus, we defined the pairs as non mediated.

The norming tasks showed that the non-mediated pairs were actually non-mediated as defined by the production task, and that the mediated and non-mediated pairs had equivalent relatedness values. To measure priming, the two sets of pairs were tested in a lexical decision experiment (using the procedure in McNamara and Altarriba, Experiment 2). The results showed about equal amounts of priming for the two sets of pairs. For the mediated pairs, the priming effect was 14 ms (replicating McNamara & Altarriba, 1988), and for the non-mediated pairs, the priming effect was 13 ms. Thus, while mediation was varied from very high to very low (and relatedness was held constant), the size of the priming effect was constant. Thus, free association production, the measure used to calibrate distance between concepts, had no effect on priming. The data provide no evidence for activation spreading through a network.

In a second experiment, we examined another basis for evaluating whether two concepts are related in memory. The idea was to use the co-occurrence of properties of words to evaluate relatedness. Ken Church at AT&T Bell Labs has developed statistics to calculate the probability with which two words occur together in a large corpus of text (e.g. several million words from the AP Newswire),

and to calculate whether the probability is significantly greater than chance (Church, 1988; Church & Hanks, 1989). We chose a set of words such that each word had a high associate according to published norms (e.g., *child* has the associate *baby*). We sent these words to Church, who returned to us all words that co-occurred with our words with a probability greater than chance. Co-occurrence was defined as appearing together in a six word window in an AP newswire corpus of 6 million words. From the returned words, we picked one that had a relatively high significance value and one that had a relatively low value. Neither of these words was the high associate or a synonym of the high associate. From these words, we were able to form three pairs, each with the associate of the originally chosen word as target. For one pair, the target was the high associate (e.g., *child-baby*); for the second pair, the target co-occurred with high probability (e.g., *hospital-baby*), and for the third pair, the target co-occurred with lower probability (e.g., *room-baby*). Free association productions showed that the targets were produced in response to the high and low probability primes with essentially zero probability.

The three kinds of pairs, along with a control condition in which the target was preceded by an unrelated word, were tested in a lexical decision experiment. The amount of facilitation varied across the four conditions, with significant amounts of facilitation for the target when it was primed by its high associate from production norms and when it was primed by the word that co-occurred with it in the corpus with high probability.

This experiment suggests that co-occurrence leads to weak direct associations that give pairs of words enough familiarity to produce priming in an experimental situation. The implication of this result is that any word in the language probably has a number of direct weak associates that are capable of producing priming effects, in the absence of mediating associates.

GENERAL DISCUSSION

In the first section of this chapter, we presented tests of the current global memory models. The tests used standard recognition memory experiments to collect ROC curves for weak and strong test items (using confidence judgments or manipulating old/new test item probability in the test list). The slope of the z-transforms of the hit and false alarm rates for the various criterion placements were plotted, and the resulting curve was linear, which is consistent with the assumption of normal distributions for old and new item familiarity values. For normal distributions, the slope of this curve is the ratio of the standard deviations for the noise and signal distributions. The empirically obtained slope was approximately constant at 0.8 over a wide range of d′ values. None of the current models can accommodate this result. The composite models predict that the variances of the signal and noise distributions should be about the same (i.e., the slope should be 1.0), and the instance-based models (SAM and MINERVA 2) predict that the signal variance should be greater than the noise variance with the ratio increasing as the strength of old items increases.

The predictions of the models cannot be changed in any simple way to accommodate a ratio of the noise and signal standard deviations equal to 0.8. In TODAM and MINERVA 2, the predictions arise from the fabric of the model and

any attempt to modify the model would unravel fits to other data. SAM has proved more flexible, but the original version of the model and the differentiation variation both lead to incorrect predictions and there is no obvious fix. For all the models, the challenge will be to find a modification that will not upset the fits of the model to other data (cf. Gillund & Shiffrin, 1984).

These data, along with the data from the mixed/pure list designs (Ratcliff, Clark, & Shiffrin, 1990; Shiffrin et al., 1990), provide useful building blocks for testing and developing new models and variants on the old models. The issue for new models is how to decide what should be the fundamental data on which to build a model. Clearly, with regard to empirical effects, the list length effect, list strength effect, repetitions effects, and so on are basic. If a model does not produce better performance on an item the longer or more often the item is studied, then it is not adequate. But equally clearly, these empirical effects are not enough to constrain models, as shown by the tests in this chapter. If the empirical effects were completely constraining, then there would not be the variety of models reviewed earlier, or the models would be mimics of each other. To reiterate, these models have vastly different assumptions in terms of structure, representation, and processing (e.g., local vs. distributed, separate vs. common memory, etc.), they make many of the same predictions, but they produce different predictions in many other cases. Given that this variety of models can handle the empirical effects with varying but overall success, we believe that some of the fundamental assumptions (such as the variance effects) need to be tested in conjunction with the empirical effects. If a model fails on some of these fundamental properties, but can be modified to give success, then the modified model still has to account for a range of empirical effects (which may require a new and comprehensive set of fits to the data for which the older version was developed). We believe that the data specifying the behavior of the variance of the familiarity distributions are exactly the kind of data that will prove to be fundamental in evaluating old models and developing new models.

The second section of the chapter deals with the application of the global memory models to priming phenomena. Ratcliff and McKoon (1988, see also Dosher & Rosedale, 1989) presented a theory of priming phenomena based on the assumption that memory is probed with a cue made up of a compound of the prime and target items. Ratcliff and McKoon accounted for a range of empirical data with this compound cue notion and showed that it made a few strong predictions that the main competitor (spreading activation) did not make. In this chapter, we reviewed a test of one of these predictions, that weak priming effects should be obtained between weakly associated words for which there is no mediating concept through which activation could spread. We found that such effects did occur, consistent with the compound cue view but not with spreading activation theories. The effects can be predicted from the probability that the prime and target words co-occur significantly often in a large corpus of text, suggesting that there are large numbers of weakly related pairs of words in memory that can give small priming effects.

Spreading activation was first proposed as a general retrieval mechanism that would provide paths of connected concepts in semantic memory, so that the connections among the concepts could be evaluated. It is this function, long range spread of activation, that is perhaps the primary function of spreading activation. Thus, if the available data show no evidence of long range spreading, then the utility of the concept of spreading activation is diminished. In contrast, the compound cue theory predicts that the range of priming is short and that only direct associates (or, in the SAM version, pairs that share associates) will show priming. To the extent that there is no evidence for long range spreading activation, the compound cue model is supported. The data that McNamara and Altaribba presented as showing the compound cue model wrong ("mediated" priming) can be reinterpreted as showing weak (but direct) associative priming.

To conclude, we must echo the theme present throughout Ben Murdock's research: Our goal is to develop theories of representation and process in memory. Without theory (be it explicitly quantitative or more qualitative), it is impossible to decide what is important and what is trivial. In this chapter, we have presented two kinds of research. One kind is motivated by theory (ROC curves), and without theory would be of little interest. The other (priming) shows how theory can guide interpretation and understanding of "interesting phenomena." When is a phenomenon interesting? When it speaks to theory, whether formally or intuitively.

AUTHOR NOTE

The research described in this chapter was supported by NSF grant 85-16350 and AFOSR grant 90-0246 (jointly funded by NSF) to Gail McKoon and by NIMH grants MH44640 and MH00871 to Roger Ratcliff.

REFERENCES

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 336-345.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing, 37*, 54-115.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on applied natural language processing*. Austin: ACL.

Church, K., & Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 23rd annual meeting of the association for computational linguistics*, Vancouver: Association for Computational Linguistics.

de Groot, A. M. B. (1983). The range of automatic spreading activation in word priming. *Journal of Verbal Learning and Verbal Behavior, 22*, 417-436.

Dosher, B. A., & Rosedale, G. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. *Journal of Experimental Psychology: General, 2*, 191-211.

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

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

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

Hintzman, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

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

Kortge, C. A. (1990). Episodic memory in connectionist networks. In *The proceedings of the 12th annual conference of the cognitive science society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Markowitz, J. & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception & Psychophysics, 2*, 91-100.

McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language, 27*, 545-559.

McNicol, D. (1972). *A primer of signal detection theory*. London: Allen and Unwin.

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

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review, 90*, 316-338.

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

Murnane, K., & Shiffrin, R. M. (1990). *Interference and the representation of events in memory*. Manuscript submitted for publication.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review, 91*, 281-294.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285-308.

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

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*, 385-408.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1991). *Testing global memory models using ROC curves*. Manuscript submitted for publication.

Sheu, C.-F. (1990). *A note on the multiple-trace memory model without simulation.* Manuscript submitted for publication.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). The list strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179-195.

Weber, E. U. (1988). Expectation and variance of item resemblance distributions in a convolution-correlation model of distributed memory. *Journal of Mathematical Psychology, 32,* 1-43.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1990). *A demonstration of the list-strength effect in recognition memory.* Manuscript submitted for publication.