Psychological Review
1992, Vol. 99, No. 3, 518-535

0033-295X/92/$3.00

Testing Global Memory Models Using ROC Curves

Roger Ratcliff
Northwestern University

Ching-Fan Sheu
Carnegie Mellon University

Scott D. Gronlund
University of Oklahoma

Global memory models are evaluated by using data from recognition memory experiments. For recognition, each of the models gives a value of familiarity as the output from matching a test item against memory. The experiments provide ROC (receiver operating characteristic) curves that give information about the standard deviations of familiarity values for old and new test items in the models. The experimental results are consistent with normal distributions of familiarity (a prediction of the models). However, the results also show that the new-item familiarity standard deviation is about 0.8 that of the old-item familiarity standard deviation and independent of the strength of the old items (under the assumption of normality). The models are inconsistent with these results because they predict either nearly equal old and new standard deviations or increasing values of old standard deviation with strength. Thus, the data provide the basis for revision of current models or development of new models.

In the long tradition of modeling the structures and processes that underlie memory, there have been two main considerations in theory building. The first has been the ability of a theory to cover a wide range of the phenomena under examination. This consideration has become especially important in response to the plethora of simple models that were developed for extremely limited domains 15 or more years ago. The second consideration is the standard criterion in modeling in all disciplines, that is, modeling detailed aspects of data. In memory research, many early models were flawed in one or the other of these respects and, in particular, most were criticized because they dealt with only a handful of experimental procedures. In contrast, the global memory models (e.g., Eich, 1985; Gillund & Shiffrin, 1984; Hintzman, 1986, 1988; Murdock, 1982, 1983, 1989; Pike, 1984; Ratcliff & McKoon, 1988) that have been developed recently have gone beyond the earlier models in their scope of application. They are capable of explaining a variety of phenomena across a range of experimental paradigms, and they can account for functional relationships with coherent variation of model parameters. The approach so far has been one of making the models as wide-ranging as possible by applying them in as many different domains as possible. This effort is of tremendous importance because of the serious criticism concerning lack of scope of earlier models. It is unfortunate that, concurrent with the ability of the models to fit a range of phenomena comes the suspicion that the models are too flexible to allow tests of their basic assumptions. This suspicion arises because models with quite different structures and processes seem to account for a wide range of data equally well and because new phenomena can be accommodated in the models by adding post hoc assumptions. However, the suspicion is unfounded; the models are testable and falsifiable. Some results require only minor changes in the structure of the model, whereas other results necessitate radical revisions of the models. For example, Ratcliff, Clark, and Shiffrin (1990) provided data with which, at that time, all of the global memory models were inconsistent. Our article provides additional tests of the current global memory models—tests that deal with basic assumptions of the models.

The global memory models are designed to account for performance across a range of tasks, including recall, recognition, frequency judgment, categorization, and serial-order recall. Each model deals with only a subset of the tasks, but all of the models attempt to deal with recognition. Thus, with recognition the models can be directly tested, both individually and against each other. The recognition paradigm is elementary: A list of words is presented to the subject for study and then a test list is presented that is composed of some old words from the study list and some new words that were not on the list. The subject is required to indicate whether each word is *old* or *new* by pressing one of two response keys, and the accuracy and reaction time of responses are recorded. The models account for recognition performance by supposing that a test item interacts with all items in memory to produce a value of match or familiarity or strength (*familiarity* is used interchangeably with the terms *strength, degree of match,* and *relatedness,* and all denote the output of the match between the probe and memory). The match value is used in a signal detection analysis to determine a response: If the value is higher than a criterion value, respond *old;* if lower, respond *new.* Although the familiarity response dimension is the same for the various models, the structures of the models are radically different. TODAM (Mur-

This research was supported by National Institute of Mental Health Grants MH 44640 and MHK00871 to Roger Ratcliff and National Science Foundation Grant 85-16350 to Gail McKoon.

We thank Gail McKoon for programming Experiments 1 and 3 and for extensive comments on this article. We also thank Richard Shiffrin, Douglas Hintzman, and the reviewers for comments on the article.

Correspondence concerning the article should be addressed to Roger Ratcliff, Psychology Department, Northwestern University, Evanston, Illinois 60208.

dock, 1982) assumes that an item is a vector of attributes and that each studied item is added to a common memory vector. For recognition, the dot product between the test vector and the memory vector gives the match value. In MINERVA 2 (Hintzman, 1986), items are also vectors of features but are stored separately in memory. At retrieval, a (modified) dot product is calculated between the test item and each item in memory, and then the value of each dot product is cubed and all of the dot products are summed to give a match value. In the SAM model (Gillund & Shiffrin, 1984), strengths between each item as a cue and every item in memory are built up during encoding, and at test, the sum of the strengths between a test item and each item in memory serves as the match value. In all of these models, the test item interacts with all of the information in memory.

The critical tests of the global memory models presented in this article are concerned with the relative behaviors of the variances of the distributions of familiarity for old and new items. All of the predictions of all of the models are based on the way variance is introduced into the model and substantive alterations to the assumptions about variances would change many of the published predictions of the models. Thus, tests of the variance assumptions are critical to evaluating the models. Each of the models makes strong predictions about the behavior of old- and new-item familiarity variances as a function of strength of the old items. For example, the models of Gillund and Shiffrin (1984) and Hintzman (1986, 1988) predict that the variance in old-item familiarity is greater than the variance in new-item familiarity and that this difference becomes greater as the average value of familiarity or strength of the old items becomes greater. Murdock's (1982) model, in contrast, predicts that new-item familiarity has about the same variance as old-item familiarity for all values of strength of old items.

Predictions of the Memory Models

For the three memory models that we consider in detail, predictions are derived using similar methods. In each of the models, memory comprises a representation of the studied items. In SAM and MINERVA 2 (instance-based models), the items are kept separate, whereas in TODAM (a composite model), the items are stored in a composite trace. At test, a test item is compared with all of memory, and the value of the overall degree of match or familiarity is computed. In all of the models, familiarity is computed by summing the match values between the test item and each stored item in memory. In the instance-based models, this is easy to understand, but in the composite model, it is more difficult because items are jumbled together in memory and not represented separately. However, to derive predictions, the vectors stored in the composite memory are assumed to be independent, which means that from a computational point of view, the contribution from each stored item can be treated separately. Thus, in the composite model (as in the instance-based models) the match between the test item and each stored item is assessed, and familiarity is the sum of these individual match values.

Variance in familiarity values across either different items or the same item presented on different occasions is introduced differently in each of the models. In the vector models (TODAM and MINERVA 2), variability is a consequence of the assumption that the features or attributes that make up items take random values. In SAM, variability is derived from the process by which items are encoded into memory. To test the models against data, estimates of the mean familiarity value and the variance in the familiarity value are required for both old and new test items. The estimate of the mean value of familiarity is derived by computing the expected value (mean) of the match between a test item and each item in memory and summing these values over all items in memory. Similarly, for the estimate of variance in familiarity, the variance in the match value between a test item and each item in memory is computed, and these are summed over all the items in memory.

The predictions of the models for the means and variances in the familiarity values of old and new test items are critical. For MINERVA 2 and SAM, the difference in variance between old and new test items depends on the strength of the old items. An old test item matches one item in memory and mismatches all of the rest, whereas a new test item mismatches all items in memory. For a test item that matches an item in memory, the value of the match will depend on how strongly the item was encoded. When the match value becomes large, the variance in the value for the one match becomes as large or larger than the sum of the variances for all of the nonmatches; thus, variance for an old item dominates the sum of all of the nonmatching variances for new items. In other words, when old-item strength is high, old-item variance is larger than new-item variance. In contrast, for TODAM, the difference in variance of familiarity values for old and new test items is small for all values of strength of old items. The contribution to variance when an old item matches an encoded item is only about two times that for a nonmatch. Thus, SAM and MINERVA 2 predict increasingly large differences in variances for old and new items as a function of strength of the old items, whereas TODAM predicts about equal variances for all strength values.

The second variance prediction of the models that is tested is the behavior of new-item familiarity variance as a function of strength of the old, studied items. To illustrate the prediction, consider a stereotypic vector model such as that of Anderson (1973). In this model, items are vectors with features having values randomly selected from a normal distribution, with a mean of zero and a variance P/N (where P is usually set to one). Memory is another vector that is the sum of all of the item vectors. If the numerical values of all elements in each studied vector were doubled (representing twice as much strength), then the mean familiarity and the standard deviation in the familiarity of a test item would be twice as large for both new items and old items. This prediction of increased variance holds for TODAM, SAM, and MINERVA 2, though for different reasons in each model. (See Ratcliff et al., 1990, and Shiffrin, Ratcliff, & Clark, 1990, for details, including a SAM variant that does not make this prediction.) Full discussion is given in the sections that follow presentation of the various experiments.

To test the predictions about variance, we used a combination of two methods: receiver operating characteristic (ROC) curves to assess the relative variances of old and new familiarity values, and a "mixed/pure" experimental design in which some study lists have only strongly encoded items, some only weakly encoded items, and some in which the two kinds of items are

mixed. The mixed/pure design allows the variance of new items to be measured in a situation in which it would be presumed to change as a function of strength of the old items, that is, when strength increases from a pure weak list to a pure strong list; or in a situation in which there is only one variance, that is, when the weak and strong items are mixed. These two methods for testing variance are described in detail in the next sections.

Variance Estimates From ROC Curves

The method used to determine the ratio of old-item variance to new-item variance is based on the ROC curve. The method has been used for over 30 years (see Egan, 1958). However, its application to the recent global memory models has been overlooked.

In the first two experiments presented here, empirical ROC curves were obtained by varying the proportion of old and new test items in a test list. This manipulation leads subjects to vary their old/new criteria in responding so that when old items predominate, subjects are more likely to respond *old,* whereas when new items predominate, a *new* response is more likely. Figure 1 shows two normal distributions of familiarity with five criterion settings (vertical lines). An ROC curve is produced by plotting hit rate against false alarm rate as a function of criterion setting. By transforming the hit and false alarm rates to z scores (assuming normal distributions of familiarity), two simple equations can be written in terms of the criterion position c and the means and standard deviations of the signal (old) and noise (new) distributions, as follows:

$$z_{\text{hit}} = -(c - \mu_s)/\sigma_s$$

and

$$z_{\text{fa}} = (c - \mu_n)/\sigma_n,$$

where μ_s and σ_s, and μ_n, and σ_n are the mean and standard deviation of the signal and noise distributions, respectively.

If the criterion position is eliminated from these equations, then

$$z_{\text{hit}} = (\sigma_n/\sigma_s)z_{\text{fa}} + (\mu_s - \mu_n)/\sigma_s,$$

as shown in Figure 1. The bottom panel of Figure 1 depicts a z-transformed ROC curve (under the assumption of normality, as before).

There are three aspects of analyses based on this latter equation that are important for testing the global memory models. First, if the underlying signal and noise distributions are normal (as the models either assume or predict), then the result of plotting z_{hit} versus z_{fa} is a straight line, as in the bottom panel of Figure 1 (Green & Swets, 1966; McNicol, 1972). If the distributions are not normal, then under many conditions, the resulting curves will be indistinguishable from straight lines (Lockhart & Murdock, 1970). Second, the slope of the z_{hit} versus z_{fa} line (the z-ROC curve) provides an important test of the models because, for normal distributions, the slope gives the ratio of the noise standard deviation to the signal standard deviation. Third, the intercept of the linear z-ROC line (for normal distributions) is the difference between the two distribution means divided by the signal distribution's standard deviation. This provides an estimate of d', the discriminability between old and new items. Varying strength of the old items will produce a set of z-ROC curves, and the d' values obtained from this set of curves can be used to measure the discriminability difference between strong and weak studied items and thus test (for some models) predictions for the effects of strength on the variances of old- and new-item familiarity values.

Note that there are several measures of d' that could be used when the signal and noise distribution variances are different from each other (McNicol, 1972, pp. 87–90), but as long as one is used consistently, others can be calculated as needed from the z-ROC plot. We will mostly use μ_s/σ_s, setting μ_n to zero.

Mixed/Pure List Designs

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 because they are either 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 of the items are weak and half are strong.

A mixed/pure design allows examination of the effect of the strong items in a mixed list on the weak items in the list, and vice versa. Performance on weak items in a mixed list can be compared with performance on the weak items in a pure list to determine the effect on them of the strong items in the mixed list. Also, performance on strong items can be compared in the mixed and pure lists to determine the effect on them of weak items. If performance on weak items is poorer in a mixed list than in a pure list, and strong items are better in a mixed list than in a pure list, then this is termed a *list-strength effect* (Ratcliff et al., 1990).

The models predict this list-strength effect, and the prediction rests on the behavior of the variance of the noise distribution. In the example presented earlier using a stereotypic vector model, the noise variance will be larger in a pure strong list than in a mixed list and larger in a mixed list than in a pure weak list. Given that the mean strengths of items are predicted to be the same in mixed and pure lists (for all the models considered here, see Shiffrin et al., 1990), then the noise variance differences entail the list-strength effect: d' for strong items in a mixed list will be greater than d' for strong items in a pure list; in a similar manner, d' for weak items in a mixed list will be smaller than d' for weak items in a pure list. The statistic used to assess this difference between mixed and pure list d's is the ratio of ratios of d' values: the ratio of d' for mixed strong to d' for mixed weak divided by the ratio of d' for pure strong to d' for pure weak, $\text{Rr} = (d'_{\text{ms}}/d'_{\text{mw}})/(d'_{\text{ps}}/d'_{\text{pw}})$. This measure is used because in the models, mean strength values divide out, leaving a simple expression in terms of standard deviations, $\text{Rr} = SD(\text{ps})/SD(\text{pw})$, where SD is the standard deviation in the noise distribution.

The ratio of ratios allows a critical test: The models predict that standard deviation of the noise distribution for a pure strong list must be greater than the standard deviation of the

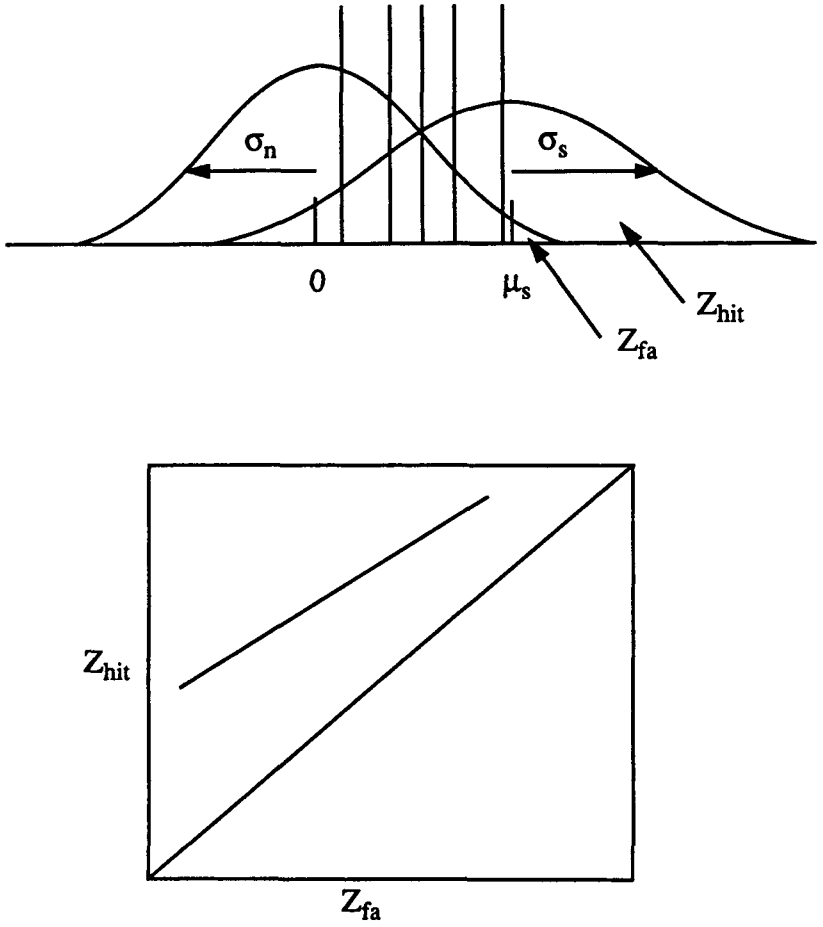

Figure 1. Illustration of normal distributions with different criterion settings and the resulting *z*-transformed receiver operating characteristic curves. (fa = false alarm.)

noise distribution for a pure weak list; therefore, the ratio must be greater than one, and this ratio must increase as the strength of the strong items is increased relative to the weak items. When the mixed/pure list design was used by Ratcliff et al. (1990) to test the global memory models, the models all failed. In results from a number of experiments, the pure and mixed list conditions did not show strength differences between weak items and between strong items (Ratcliff et al., 1990; Shiffrin et al., 1990); therefore, the ratio of ratios was about one, not greater than one. Thus, the results show that noise variance did not change as a function of pure versus mixed list types.

In this article, we combine the tests by using ROC curves and mixed/pure list designs to examine both the relative variances of old and new items and the variance of new items in mixed and pure lists as a function of the strength of the old items in the lists. This combination will provide stringent tests of the behavior of variances for old- and new-item distributions in the models, which, as we show later, constrain the models severely. The next sections present experiments designed to vary the strength of the items in a mixed/pure design using ROC curves. Following the experiments, the various models and their detailed predictions are presented.

Experiment 1

The aim of Experiment 1 was to vary jointly the probability of old and new items in the test list and the strength of items in the study list by using a mixed/pure list design. The study list was composed of pairs of words, and old/new recognition test lists were interspersed with occasional cued-recall test lists. Strength of items was manipulated by varying the amount of study time per pair from 1 s to 5 s. Probability was manipulated by varying the ratio of old and new test items in the test list from 1:4 to 4:1.

Method

Subjects. There were 4 paid subjects recruited from the student population at Northwestern University. Each subject took part in either 19 or 20 sessions, each lasting about 50 min.

Procedure. There were two kinds of study lists presented to a subject: pure lists and mixed lists. In a pure list, each of 16 pairs of words was presented for the same amount of time, 1 s for weak or 5 s for strong. In a mixed list, sequential blocks of pairs in a study list had different study times: the first 2 pairs at 1 s, the next 6 pairs at 5 s, the next 6 pairs at 1 s, and the last 2 pairs at 5 s (the first and last pairs were buffer pairs),

or the reverse ordering of presentation times. Subjects were instructed to learn the pairs for later cued-recall tests. The cued-recall tests were included to encourage the subjects to rehearse each pair of words together and not rehearse across pairs (see Ratcliff et al., 1990, for further discussion). A recognition test followed each study list. The probability of old and new test items was varied in the test list: There were 8 old and 32 new, 16 and 32, 32 and 32, 32 and 16, or 32 and 8. Subjects were informed of these probabilities after the study list and before the test list. There were 20 recognition lists per session, leading to one list of each type per session (4 mixed/pure × 5 test-item probability). After the study phase of a list, subjects pressed a key to begin the test phase. In the test phase, old and new items were presented in random order, and subjects were instructed to be fast without sacrificing accuracy. After each response, there was a 250 ms blank interval followed by the next test item. For two of the study lists, the recognition test list was followed by a cued-recall test (the left member of the study pair was presented and the subject was required to recall the right member). Instructions recommended that pairs be learned for cued recall and the test lists were preceded by one practice list in which recognition was followed by cued recall.

Materials. The words for each session were chosen randomly without replacement from a pool of 1,650 two-syllable common English words not more than eight letters long.

Results

Responses with reaction times below 300 ms and above 2,500 ms and test items from the buffers (the first and last two studied pairs) were discarded from the analyses. For mixed lists, performance was averaged over the two mixed-list orders. Table 1 gives the hit and false alarm rates for each mixed/pure and weak/strong condition for each test-list probability averaged across subjects (later analyses examine performance of individual subjects).

The analyses of importance for the global memory models concern the ROC curves. The global memory models predict that both the old and new test-item familiarity distributions will be normal in shape. If the data are consistent with normal distributions, the z transformation of the hit plotted against the z transformation of the false alarm rate will be linear. Note that linearity does not necessarily mean that the distributions are normal; other nonnormal distributions can produce linear or approximately linear functions (Lockhart & Murdock, 1970). Figure 2 shows the z-transformed functions averaged across subjects for mixed and pure, and weak and strong items (using only the first 40 test items in each test list to equate test position as a function of old/new test-item frequency). The graphs show nearly linear curves with approximately equal slopes. Table 2 shows linear regression analyses on the group data for the slope, the intercept, and the standard deviation in each. The same linear regressions were carried out on the data from the individual subjects, and the means of the slopes and intercepts are also shown in Table 2.

The main results to note from the linear regression analyses are that the slope of the z-transformed ROC curve is about 0.84 on average and that the slope is independent of the strengths of the items. In all but one case, the slope is significantly less than one, and none of the slopes are significantly different from each other (using the standard errors given in Table 2). These results clearly contradict predictions of the models. First, the 0.84 slope shows that the variance for new-item familiarity is less than that for old-item familiarity, which contradicts TODAM's prediction. Second, this value of slope is constant as a function of the strength of the old items, which contradicts SAM and MINERVA 2.

Another prediction of the models is that variance of new-item familiarity should increase with strength of the old items, leading to a larger d' difference between strong and weak items in mixed lists than in pure lists. Contrary to this prediction, there is little, if any, difference between performance in mixed and pure lists, replicating results of Ratcliff et al. (1990). Ratcliff et al. obtained d' values from one criterion value only (i.e., only one value of old/new test-list probability); here, they are obtained from the whole ROC curve. As noted earlier (and in Ratcliff et al., 1990), the ratio of ratios of d', $\mathrm{Rr} = (d'_{\mathrm{ms}}/d'_{\mathrm{mw}})/(d'_{\mathrm{ps}}/d'_{\mathrm{pw}})$, which reduces to $SD_{\mathrm{ps}}/SD_{\mathrm{pw}}$ for the models, is used to test the models. For the data in Table 2, d' is calculated from the intercept divided by the slope; this gives the difference between old and new familiarity divided by the standard deviation in the new-item familiarity. The ratio of ratios, Rr, is 1.04, based

Table 1
Results From Experiment 1

Probability condition	Presentation time per item	Mixed list Hit rate	Mixed list F/A rate	Pure list Hit rate	Pure list F/A rate
1 new/4 old	1 s	0.817	0.510	0.849	0.586
	5 s	0.928	0.510	0.907	0.505
1 new/2 old	1 s	0.691	0.405	0.751	0.402
	5 s	0.828	0.405	0.820	0.363
1 new/1 old	1 s	0.636	0.270	0.685	0.313
	5 s	0.788	0.270	0.757	0.270
2 new/1 old	1 s	0.498	0.196	0.561	0.224
	5 s	0.680	0.196	0.694	0.179
4 new/1 old	1 s	0.419	0.096	0.446	0.111
	5 s	0.595	0.096	0.554	0.083

Note. F/A = false alarm.

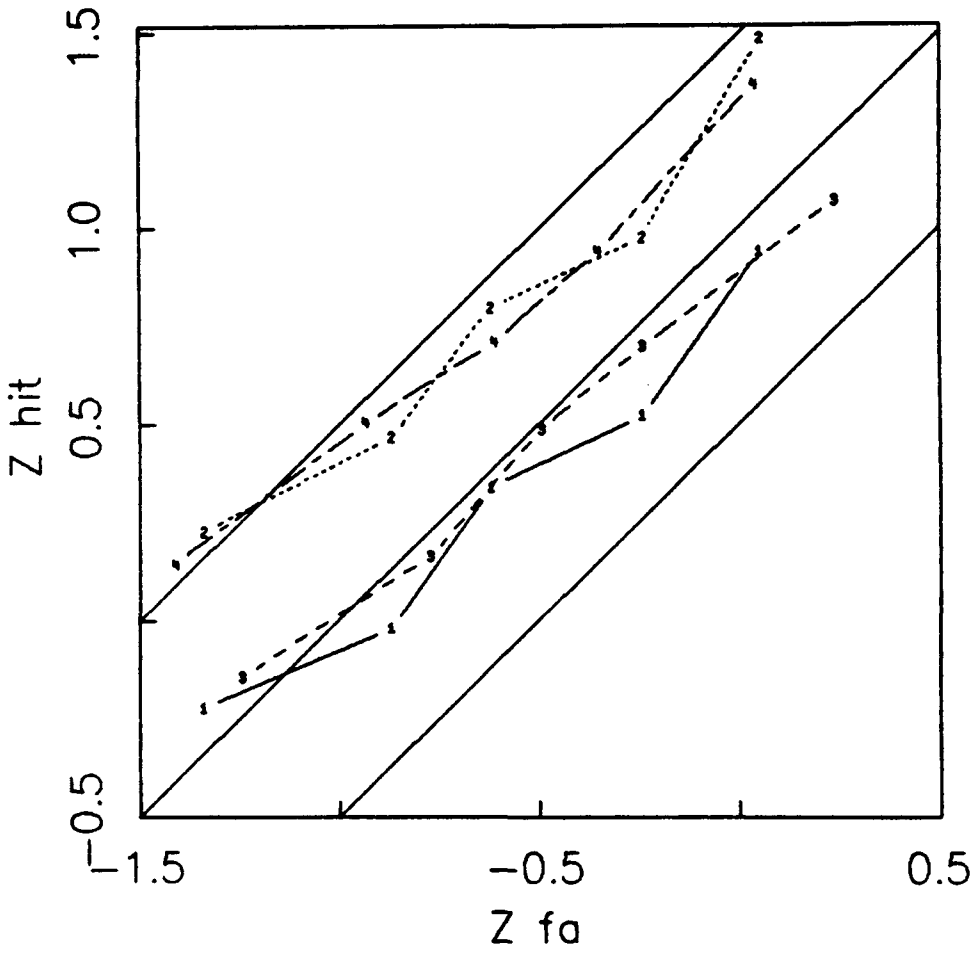

Figure 2. *Z*-transformed receiver operating characteristic curves for the group in Experiment 1; presentation time is varied to produce changes in strength. (The diagonal lines represent slopes of 1. The curves represent mixed strong, pure strong, pure weak, and mixed weak, reading top to bottom on the left-hand data points. fa = false alarm.)

on the data from all subjects combined, and 0.99, based on the average of the values for the individual subjects. These values replicate those presented in Ratcliff et al. (1990) and show that, counter to the predictions of the models, the new-item familiarity variance is not larger for pure strong lists than for pure weak lists. In other words, new-item familiarity variance does not change as a function of strength of the old items.

Experiment 2

Experiment 2 was designed to vary strength by varying the number of repetitions of studied items instead of varying study time. This method was used by Ratcliff et al. (1990) and leads to the same predictions by the models as does the study time manipulation (Shiffrin et al., 1990).

Table 2
Linear Regression Fits to the Z-Transformed ROC Curves for Experiment 1

Condition	Slope	Intercept	Slope *SD*	Intercept *SD*	S × S slope	S × S intercept
Mixed weak	0.824	0.809	0.054	0.044	0.792	0.774
Mixed strong	0.872	1.316	0.115	0.089	0.809	1.328
Pure weak	0.841	0.876	0.041	0.029	0.842	0.903
Pure strong	0.827	1.271	0.054	0.044	0.801	1.277

Note. S × S slopes and intercepts were derived from the fits to individual subjects' data, which were averaged over subjects. ROC = receiver operating characteristic.

Method

Subjects. There were 7 paid subjects recruited from the student population at Northwestern University. Each subject took part in 20 to 25 sessions, lasting about 45 min each. One of these subjects had participated in Experiment 1.

Procedure. In the study list, single words were presented in either pure or mixed lists. In a pure strong list, each of 20 words was repeated five times in a random order (total list length was 100 words). In a pure weak list, each of 20 words was presented once (total list length was 20 words). In a mixed list, 10 words were repeated five times, and 10 words were presented once (randomly intermixed; total list length was 60 words). The only constraint on the presentation order was that repetitions of words could not occur within a lag of four. Subjects were informed before each study phase which type of list would be presented (all fives, all ones, or half and half). Each word in a study list was displayed for 750 ms.

Each study list was immediately followed by a recognition test list. The probability of old and new test items was varied so that there were 4 old and 20 new, 8 and 16, 12 and 12, 16 and 8, or 20 and 4. Before the beginning of the test phase, subjects were told what the composition of the test list would be (17% old, 33% old, 50% old, 67% old, or 83% old). Old and new items were presented in random order.

In the recognition test phase, subjects were instructed to be fast without sacrificing accuracy (responses longer than 1,500 ms were followed by a "too slow" message for 500 ms). Error feedback was given after every incorrect response, displayed for 500 ms. To encourage subjects to make use of the probability information, an error in the extreme probability conditions (83% old and 17% old) was given a 2,500-ms time penalty (the word "error" was spelled out letter-by-letter at 500 ms per letter). Otherwise, the next test item followed the response immediately. A summary of performance was also given to subjects after every three study lists of trials.

There were 30 study-test lists per session, two lists of each type: 3 mixed/pure study lists × 5 test-item probability conditions.

Materials. The words for each session were chosen randomly without replacement from the same word pool used in Experiment 1.

Results

Responses faster than 300 ms and slower than 2.5 standard deviations above the mean for each subject and condition were eliminated from the analyses. Study-list lengths varied as a function of list type. For pure weak lists (1 repetition), study-list length was 20 words; for the mixed lists, list length was 60 words; and for the pure strong lists, list length was 100 words. For the analyses of slope of the *z*-transformed ROC curves, two analyses can be performed: one in which responses are averaged over all serial positions and one in which responses from early study positions in longer study lists are discarded so that weak items in a pure list are compared with weak items in a mixed list equated over equally recent study positions (analyses of strong items are performed in a similar manner). This latter equated analysis was used to examine the mixed/pure list difference and to form the ratio of ratios for the mixed/pure list analysis.

Figure 3 shows the *z*-transformed hit and false alarm rates for the group average data over all serial positions, and Figure 4 shows the equivalent results for the data with recency of study position equated. The functions are linear functions with equal slopes, replicating Experiment 1. Table 3 shows the hit and false alarm rates, and Table 4 shows the linear regression analyses separately for weak and strong items equated for study positions. The slopes lie between 0.70 and 0.90 for the group data and between 0.68 and 0.80 by fitting data for individual subjects and averaging. As noted earlier, the prediction of SAM and MINERVA 2 is for the slope to decrease as old items get stronger. In fact, although the differences are not significant (but close), the slope increases as old items get stronger.

The ratio of ratios computed from the group data is 1.21, and the average of each individual subject's ratio of ratios is 1.13. Given that the strong items were presented five times and the weak items were presented only once, the models would predict a value much larger than one (e.g., Gillund & Shiffrin, 1984, would predict a value of 2.2). The obtained values are only a little larger than one and are again consistent with the results of Ratcliff et al. (1990). Therefore, the results again contradict the prediction that the variance of new-item familiarity will increase as the strength of old items increases.

Experiment 3

Experiment 3 was designed to use an alternative method for obtaining ROC curves, a confidence judgment procedure. In the confidence judgment procedure, subjects are required to make a recognition response on a 6-point scale with values ranging from *sure old* to *sure new.* The response probabilities in each confidence category can be used to construct an ROC curve, as in Experiments 1 and 2. The confidence judgment procedure was used to make sure that our results were not specific to the probability manipulation used in these two experiments. Markowitz and Swets (1967) compared the two methods in auditory detection and found differences between results from the two procedures that they attributed to practice on the signal stimuli when the signal probability was high. In recognition memory procedures, this possible confound is not a problem because practice at test on one old word will not transfer to another old word because each word is tested only once per list. However, similar results on the two procedures will improve our confidence in their generality.

Method

Experiment 2 was similar to Experiment 1, except that old and new test items were presented with equal probability, and responses were on a 6-point scale from *sure old, probably old, maybe old, maybe new, probably new,* to *sure new* response categories. The keys on the cathode-ray tube keyboard used for the confidence judgments were the *x* through *m* keys on the bottom row of the keyboard. As in Experiment 1, both pure and mixed lists of pairs of items were studied, and single items were tested for recognition. Strong items were studied for 5 s per pair and weak items for 1 s per pair (as in Experiment 1, instructions required pairs to be learned for cued recall). After the recognition phase of a practice list and after one other randomly chosen list, a cued-recall test was given to the subjects to encourage learning of pairs.

Subjects were 19 paid volunteers from the Northwestern undergraduate population. They each participated in one 45-min session with 16 study-test lists per session. The same experimental materials were used as in Experiment 1.

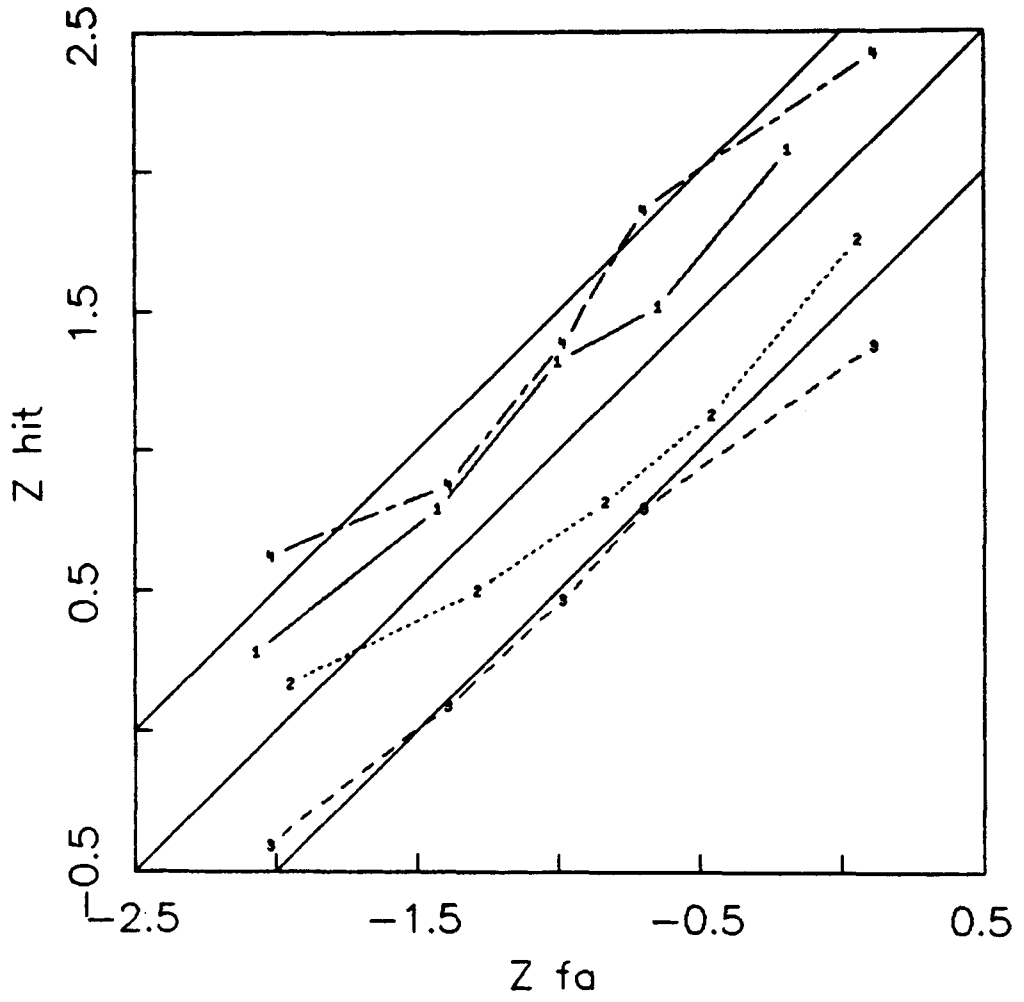

Figure 3. Z-transformed receiver operating characteristic curves for the group in Experiment 2 for all the data, that is, study position for the items not equated for the pure and the mixed lists. (The curves represent mixed strong, pure strong, pure weak, and mixed weak, reading top to bottom on the left-hand data points. fa = false alarm.)

Results

Data analyses used only study items from the middle of the blocks in the mixed lists (Items 2–5 in the blocks of six pairs) and items matched on serial position in the pure lists. Reaction times less than 250 ms and greater than 5,000 ms were also eliminated.

Figure 5 shows *z*-transformed ROC curves for the group average data. To derive these, cumulative ROCs are derived by cumulating the raw frequencies for each key, converting to probabilities, and then taking a *z* transformation (see McNicol, 1972). This contrasts with Experiments 1 and 2, in which the hit and false alarm rates were obtained directly from the probability conditions. As in the prior experiments, the functions are parallel and do not differ much from linearity (Experiment 3 shows the most systematic deviation of all the experiments, but the deviations are still small). Table 5 shows the regression analyses for the four conditions. As in Experiments 1 and 2, the slopes are in the 0.8 range, are significantly different from one, and do not vary as a function of strength of the old items.

The ratio of ratios was calculated by collapsing the old and new confidence judgments into two categories (old and new) and using these to compute d' values for the four conditions. The ratio of ratios was 1.03, again little different from one. Calculating the ratio of ratios from the slopes and intercepts of the ROC curves, for μ_s/σ_s, the value is 0.99, and for μ_s/σ_n (the intercept divided by the slope), the value is 0.94.

Serial and test position effects. It is possible that the 0.8 slope of the *z*-ROC is an artifact of averaging over study or test positions. If the true state of affairs was that, for any study or test position, the signal and noise familiarity distributions had equal variances, then averaging across study and test positions could produce a signal distribution with greater variance than the noise distribution. The overall average signal distribution would be a mixture of the individual signal distributions from different study and test positions, and if these had different

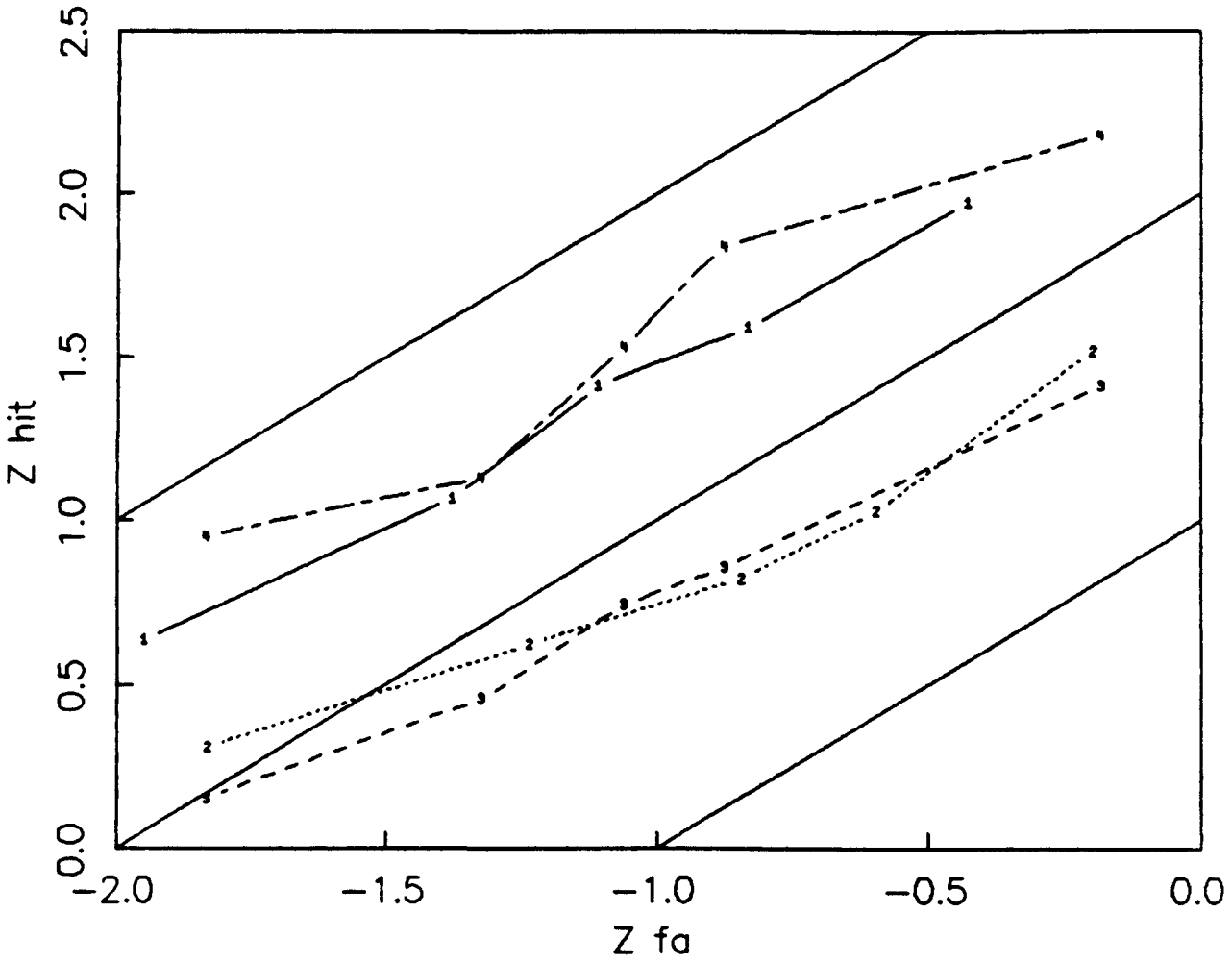

Figure 4. Z-transformed receiver operating characteristic curves for the group for Experiment 2 for the data equated for study position, that is, study position for the items equated for the pure and the mixed lists. (The curves represent mixed strong, pure strong, pure weak, and mixed weak, reading top to bottom on the left-hand data points. fa = false alarm.)

means, the probability mixture would have a greater variance than the individual distributions. The 0.8 slope would be a consequence of the mixture. Figure 6 illustrates this: The top panel shows a noise distribution with two signal distributions, and the bottom panel shows the same noise distribution with a signal distribution that is a mixture of the two signal distributions above it. The mixture has a larger variance than each of the signal distributions. There are two ways to address this issue: first, to examine the data for evidence favoring the mixture hypothesis, and second, to calculate how much separation of the distributions in the probability mixture would be needed to produce the observed slope of 0.8.

The mixture hypothesis can be tested with the data from Experiments 1 and 3. In these experiments, the study lists had primacy and recency buffers (two pairs each) that are not included in the analyses. For the nonbuffer pairs of items, we examined the effect of study position by partitioning the data *into first and second halves of the study list* and calculating slopes and intercepts of the *z*-ROC functions for each half. For the data to support the mixture hypothesis, the individual slopes for the first and second halves should be closer to one than the slope for all of the data combined. In Experiment 1, there is a small slope difference (0.06) as a function of first versus second half, but the individual slopes are not both closer to one; one is closer and the other is further away, with the average of these two about the same as the group average. For Experiment 3, the slope difference is reversed (−0.04), and as in Experiment 1, the average of the two slopes is about the same as the group average.

We also checked test position in Experiment 1. There was a difference of only 0.01 in the slopes for first versus second half of the test list (moreover, there was an intercept difference of 0.12, with early test items leading to higher accuracy). Thus, for neither study position nor test position do the data support the mixture hypothesis.

To examine the mixture hypothesis theoretically, we used simple calculations to determine how much separation of the distributions in the mixture would be needed to produce a

Table 3
Results From Experiment 2 for the Conditionalized Data

Probability condition	No. of repetitions per item	Mixed list Hit rate	Mixed list F/A rate	Pure list Hit rate	Pure list F/A rate
1 new/5 old	1	0.921	0.427	0.935	0.421
	5	0.985	0.427	0.978	0.334
1 new/2 old	1	0.803	0.190	0.848	0.275
	5	0.967	0.190	0.924	0.202
1 new/1 old	1	0.770	0.144	0.794	0.199
	5	0.937	0.144	0.930	0.134
2 new/1 old	1	0.675	0.093	0.732	0.109
	5	0.870	0.093	0.843	0.084
5 new/1 old	1	0.560	0.034	0.621	0.034
	5	0.829	0.034	0.734	0.026

Note. F/A = false alarm.

z-ROC slope of 0.8. To obtain a mixed distribution, two normal signal distributions, each with standard deviation of one and means of 1 and 2.5 were combined. With a normal noise distribution with standard deviation of one and mean set to zero, the slope is 0.79, about the same as for the empirical data. However, the *d'* difference between the two signal distributions is about 1.5, compared with *d'* differences in the data of only about 0.1 (for different study and test positions). Mixtures of several rather than just two signal distributions similarly required large separations between the signal distributions to obtain slopes around 0.8. Thus, the observed differences between different parts of the study and test lists were much too small to mix to produce the observed slope of 0.8.

The last point about this mixture hypothesis is that there might be large enough differences among individual study or test items so that mixing them would produce slopes around 0.8. However, such item differences are part of every model: In each of the models, item differences are introduced as the source of variance and so are exactly what the *z*-ROC tests evaluate.

Analyses of Other Data

Earlier research provides data that give ROC curves for recognition. Although these data generally confirm our findings, the strength of old items was not manipulated (except by Egan, 1958), nor was a mixed/pure design used.

Murdock and Dufty (1972) examined recognition memory by using confidence judgments. Subjects were required to respond on the same 6-point confidence scale as was used in Experiment 3. Slopes and intercepts of ROC curves for *z* transformation for individual subjects are shown in Table 6, and the plots of the *z* transformations are quite linear. The *z* transformed functions are linear and the slopes are less than one (they range from 0.53 to 1.05, with a mean of 0.80).

The results shown in Table 6 for individual subjects have important implications for models. One subject had a reliable *z*-ROC slope of 0.5, which was significantly different from the average slope of the group. Although the individual subjects in Experiments 1 and 2 did not show such large differences from each other (their data were somewhat noisier), other experiments in our lab have found subjects with slopes that are significantly different from each other. Thus, not only do the models have to account for the average values of slopes but also they must be capable of accounting for individual differences in these values.

Mandler and Boek (1974) also used confidence judgments in a recognition memory experiment. As the study phase, they had subjects engage in a word-sorting task in which 100 words were sorted into two to seven categories. After each sort, the deck was reshuffled and another sort was performed, and this sequence proceeded until two consecutive sorts produced 95% overlap in assignment. One week later, subjects were brought back to perform a yes/no recognition test (each recognition

Table 4
Linear Regression Fits to the Z-Transformed ROC Curves for the Conditionalized Data in Experiment 2

Condition	Slope	Intercept	Slope *SD*	Intercept *SD*	S × S slope	S × S intercept
Mixed weak	0.776	1.542	0.035	0.042	0.710	1.434
Mixed strong	0.803	2.372	0.123	0.146	0.772	2.246
Pure weak	0.709	1.526	0.094	0.103	0.680	1.450
Pure strong	0.895	2.332	0.119	0.149	0.802	2.368

Note. S × S slopes and intercepts were derived from the fits to individual subjects' data, which were averaged over subjects. ROC = receiver operating characteristic.

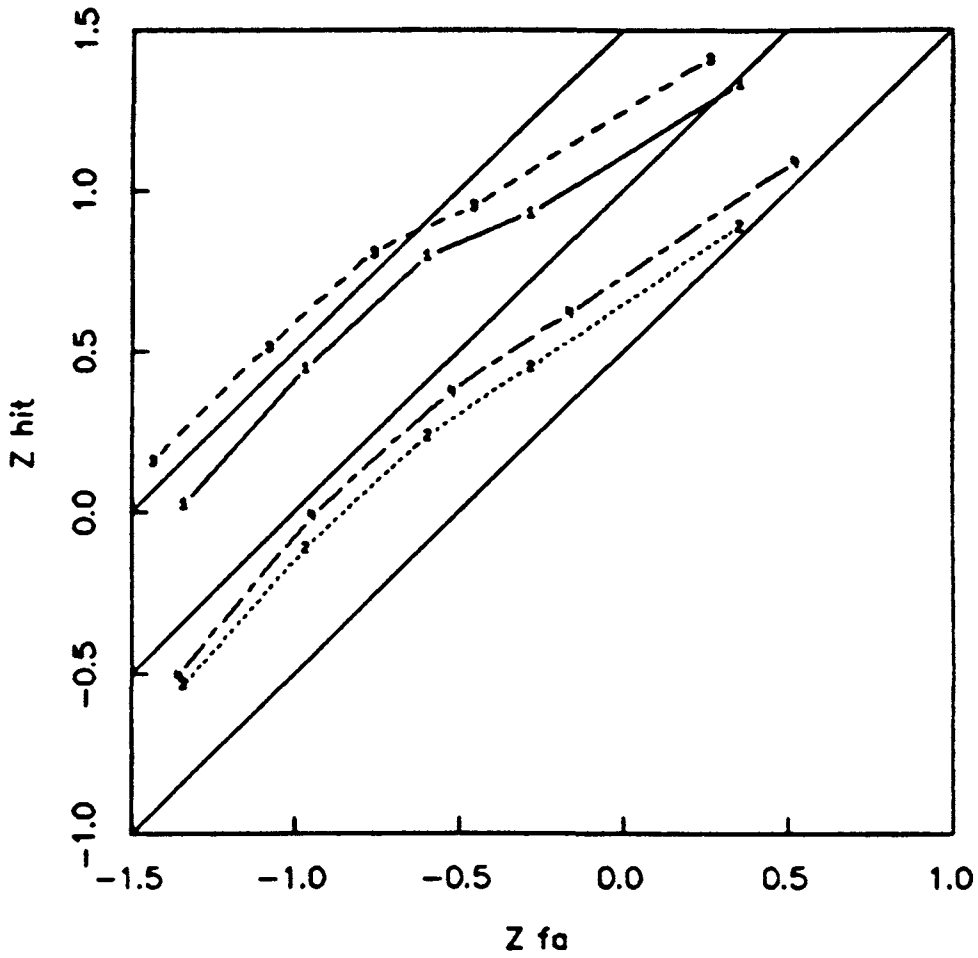

Figure 5. Z-transformed receiver operating characteristic curves for the group for Experiment 3. (The curves represent pure strong, mixed strong, pure weak, and mixed weak, reading top to bottom on the left-hand data points. fa = false alarm.)

decision was followed by a confidence judgment) on the words from the prior session. We analyzed the confidence judgment data and found that the slope and intercept of the *z*-transformed ROC curves were 0.787± 0.042 and 1.51± 0.04, respectively, in agreement with the results presented earlier.

Egan (1958) performed a recognition memory experiment in which 100 words were studied (either once or twice) and 200 words (100 old and 100 new) were tested. Responses were collected on a 7-point confidence scale. Results showed that for both once- and twice-presented items, the slopes of the *z*-transformed ROC curves were around 0.7 (0.675 for twice-presented items, intercept 2.7; and 0.713 for once-presented items, intercept 1.4) and thus roughly independent of strength of the old items, again agreeing with the results from the experiments presented earlier.

Glanzer and Adams (1990) used confidence judgments to examine the mirror effect in recognition memory. Out of 36 *z*-transformed ROC slopes in their experiments, one slope was 1.03, one was 1.00, and the rest were below 1.00, with the smallest being 0.56. The mean was 0.762. The phenomenon of interest in that article was the mirror effect (Glanzer & Adams, 1985). A large number of results show that, in general, in experiments in which one set of stimuli is recognized more accurately as old when old (e.g., low-frequency words compared with high-frequency words), they will be more accurately recognized as new when new. The slopes of the ROCs showed systematic differences as a function of word frequency, abstract/concrete, and reversed/standard letter order. Typically, slopes for the more

Table 5
Linear Regression Fits to the Z-Transformed ROC Curves for Experiment 3

Condition	Slope	Intercept	Slope *SD*	Intercept *SD*
Mixed strong	.752	1.128	.075	.061
Mixed weak	.829	.651	.058	.047
Pure strong	.718	1.263	.056	.051
Pure weak	.832	.721	.064	.052

Note. Weak items were studied for 1 s per pair and strong items for 5 s per pair. ROC = receiver operating characteristic.

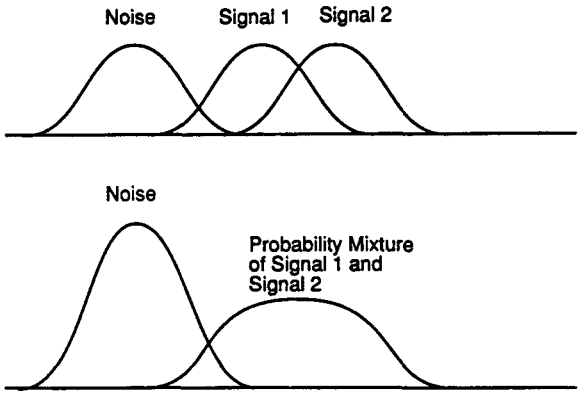

Figure 6. Illustration of the effect of averaging over study or test positions. (The top panel shows a noise distribution and two signal distributions. The bottom panel shows the result of averaging: a single noise distribution and a signal distribution with larger variance.)

extreme stimulus types (e.g., low frequency) were around 0.7, and for the less extreme stimulus types around 0.8. Thus, these data may show systematic differences in the slopes of the *z*-transformed ROC curves as a function of material type such that the materials with the highest *d'* values (e.g., low-frequency words) have the lowest slopes. However, the conditions with the lowest *d'* values appear to have their *d'* values near 0.5, which is the point at which the slope starts to rise toward 1.0 as discriminability falls to zero (see Figure 7). This suggests that, in fact, the differences may be smaller as a function of material type than those reported by Glanzer and Adams (1990).

Summary of the Empirical Results

There are three main results to constrain modeling. First, the *z*-transformed ROC curves presented earlier and by Murdock and Dufty (1972), Mandler and Boek (1974), Egan (1958), and Glanzer and Adams (1990) appear to be linear. This means that the functions are consistent with the assumption that the old- and new-item familiarity distributions are normal (though they do not imply the distributions are normal).

Table 6
Linear Regression Fits to the Z-Transformed ROC Curves for the Data of Murdock and Dufty (1972)

Subject	Slope	Intercept	Slope *SD*	Intercept *SD*
1	0.847	2.026	0.033	0.057
2	0.829	1.821	0.023	0.034
3	1.045	1.525	0.103	0.119
4	0.735	1.680	0.022	0.041
5	0.763	1.578	0.010	0.012
6	0.722	1.447	0.033	0.048
7	0.535	1.621	0.021	0.036
8	0.894	2.474	0.123	0.199
M	0.796	1.772		

Note. ROC = receiver operating characteristic.

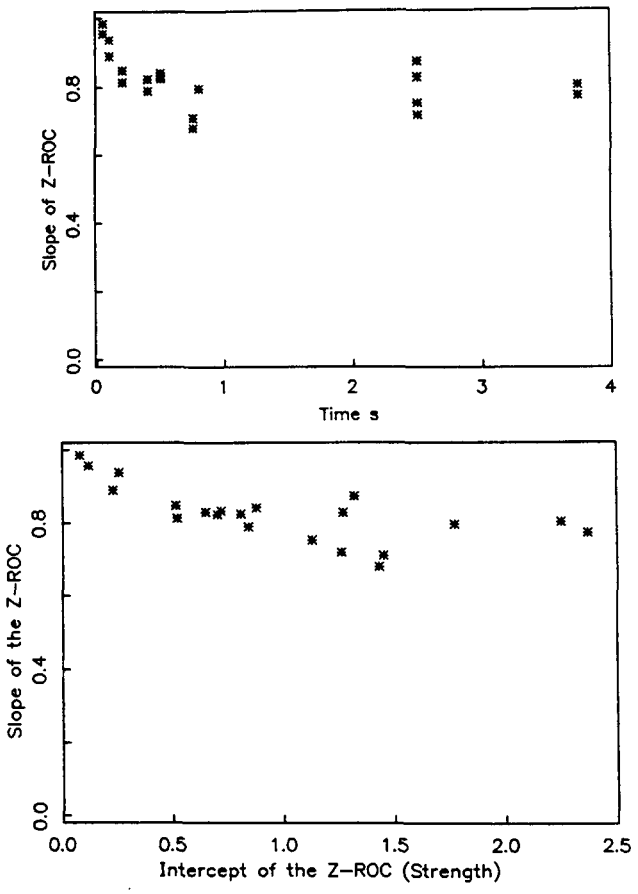

Figure 7. Slope of the *z* receiver operating characteristic (z-ROC) curve as a function of study time per item and as a function of strength, or intercept, for each condition. (For studied pairs, the study time per pair was divided by 2, and for multiple presentations, the time was the sum of the study times.)

The second major result is that the slopes of the *z*-ROC curves average about 0.8, independently of the strength of the old items. In other words, the ratio of the standard deviations in the new-item and old-item familiarity distributions does not change as a function of strength (assuming normal distributions). The results from all of the experiments can be summarized by plotting the slope of the *z*-ROC functions for each condition (all combinations of mixed and pure and weak and strong) of each experiment as a function of the study time per item and as a function of *d'* (the intercept of the *z*-ROC function) for that condition. These two plots are presented in Figure 7. It is clear that there are no systematic differences in the slope of the *z*-transformed ROC curves as a function of strength over the range of *d'* values from about 0.5 to 2.5. The value of the slope of the *z*-ROC is constant at about 0.79. Linear regression on the data points in Figure 7 (for slope of the *z*-ROC function plotted against intercept) including those from Egan, Mandler, and Boek (1974) and Murdock and Dufty (1972) and excluding points with intercept below 0.5, gives a slope (in the slope of the *z*-ROC vs. intercept) of 0.025, with a standard error in that slope of 0.025. Thus, statistically, there is no effect of strength of

the old items on the ratio of the standard deviations (σ_n/σ_s). Whatever the strength of old items, the variance of new-item familiarity is less than the variance in old-item familiarity in the ratio of 0.79. In contrast, SAM and MINERVA 2 predict a decreasing variance ratio as a function of strength and TODAM predicts a variance ratio of about one. (The values for the low strength values obtained using rapid presentation times and were from experiments in Ratcliff, McKoon, & Tindall, 1992; see also Ratcliff & McKoon, 1991).

The third result is the mixed/pure ratio of ratios result. The ratios of ratios of d' values for the three experiments are shown in Table 7. These values are consistent with those found by Ratcliff et al. (1990) and show that, in contradiction to most of the models (Shiffrin et al., 1990), the variance in new-item familiarity is about the same for strong and weak lists. The values for the ratios of ratios range from 1.13 to 0.94.

The whole pattern of results translated to familiarity distributions is shown in Figure 8 (which assumes the normal distributions predicted by the models). It is this pattern of results that constrains the existing models and provides important criteria for developing new models. Figure 8 shows that the noise distribution has the same standard deviation for pure weak, mixed, and pure strong lists. The standard deviation in the signal distribution is 1.25 times greater than the standard deviation in the noise distribution and is constant as a function of strength (if d' is greater than about 0.5) and constant as a function of list type (mixed or pure). Note that this is an empirical description of the data. There may be theoretical accounts in which theoretical familiarity is only one component of the empirical familiarity; the empirical familiarity distributions might arise from a combination of theoretical familiarity and some other factor (such as recall; personal communication, R. M. Shiffrin, January 1991).

The critical issue is whether the various models are capable of fitting these sets of data, in other words, whether a model can simultaneously produce unequal standard deviations for old and new items (in the ratio of 0.8) and constant standard deviations in mixed and pure lists as a function of the strength of old items.

Table 7
Values of the Ratio of (Mixed Strong/Mixed Weak)/(Pure Strong/Pure Weak) for Experiments 1, 2, and 3

Experiment	Presentation time per item	Ratio of ratios
1	0.5, 2.5	0.99
2	0.75, 3.75	1.13
3	0.5, 2.5	0.94 (1.03)

Note. The values of the ratios of ratios are calculated from the slope and intercept of the receiver operating characteristic curves and represent $(\mu_s - \mu_n)/\sigma_n$. The value in parentheses is computed by taking a simple split in the confidence judgment range (old vs. new) and calculating the ratio of ratios from the computed d' values. The presentation time per item is computed from the average amount of time an item is presented (for a pair, the presentation time is divided by 2; for a repetition, the time for one presentation is multiplied by the number of repetitions).

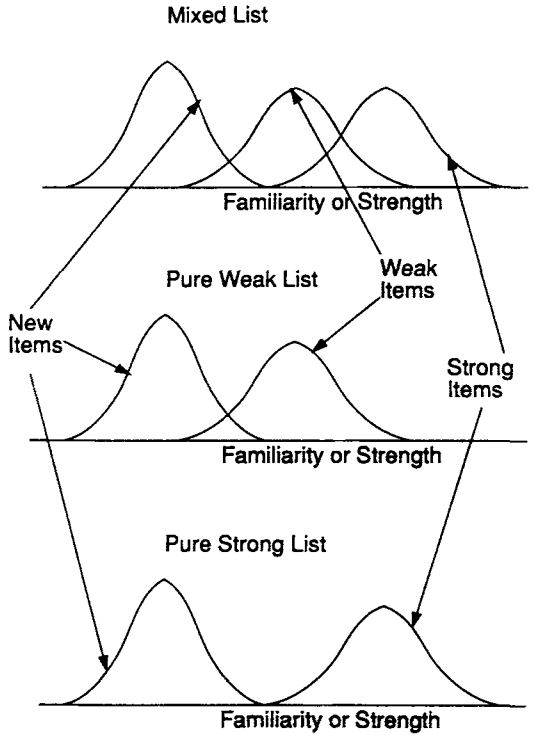

Figure 8. An illustration of the behavior of the strength distributions as a function of mixed versus pure list and as a function of strength differences. (The new-item strength standard deviation remains constant as a function of list type; the standard deviation of the old-item distribution remains constant as a function of strength and is larger than the standard deviation for new items.)

Theoretical Analyses

Hintzman's MINERVA 2 Multiple-Trace Vector Model

In the MINERVA 2 model, studied items are assumed to be represented separately in memory. A test item is compared with each stored item and an overall familiarity value is calculated by summing over these comparisons.

Specifically, the MINERVA 2 model proposes that items are vectors of features in which the features take values of −1, 0, or +1. Encoding proceeds by copying item vectors into memory vectors by using a probabilistic encoding process: For each feature, either its value or 0 is copied into the memory vector with some probability L, where L is larger for strong items (i.e., items with more study time or more repetitions). At test, the value of overall familiarity for a test item (echo intensity) is calculated as follows: If there are n vectors in memory and m features per vector, then the similarity of test vector $\mathbf{P(j)}$ to a memory vector $\mathbf{T(i, j)}$ is as follows:

$$S(i) = (1/N_R)\Sigma_j \mathbf{P(j)T(i, j)},$$

where N_R is the number of features relevant to the comparison, that is, the number that are nonzero in either $\mathbf{P(j)}$ or $\mathbf{T(i, j)}$. To get echo intensity (I), the similarity for each item is cubed to give an activation value, $A(i) = S(i)^3$, and these activation values

are summed over items, $I = \Sigma_i A(i)$. Echo intensity (familiarity) is the basis for a recognition decision through signal detection theory.

The model makes strong predictions about the relative values of the standard deviations of echo intensity for old and new items. The standard deviations of the similarities of old items are comparable with the standard deviations of the similarities of new items. However, the standard deviations are not comparable for activation values. For new items, mean similarity is near zero, whereas for old items it is greater than zero. Thus, when similarity is cubed, the standard deviations in the resulting activation values for new items are attenuated relative to the standard deviations for the old-item activation values. Typically, the mean and variance for activation in the single memory item that matches an old test item are comparable with the sums of the means and variances of activation values for new test items that do not match any memory item. To compute means and variances for echo intensity, the means and variances of the activation values are summed over all items in memory; thus, the means and standard deviations in echo intensity are larger for old items than for new items.

As study time per item increases, the similarity of old test items to matching items in memory increases, resulting in larger means and larger standard deviations in activation values and echo intensities for the old items. When the number of repetitions of a study item increases, the activation values for the repeated items are added into the calculation of echo intensity, and the variances for old test items matching repeated studied items in memory dominate over the nonmatching variances for new items. Thus, with increases in either study time or number of repetitions, the model predicts that the standard deviation in echo intensity is larger for old items than for new items.

Sample values of echo intensity from the model (for reasonable parameter values) are shown in Table 8. (In addition, inspection of Figure 4, Hintzman, 1988, illustrates the increasing variance with repetition.) The results in Table 8 were produced by using a recursive exact solution for Hintzman's model (Sheu, in press). First, the distributions of echo intensity are approximately normal, which fits well with the linear *z*-ROC functions. Second, as the encoding probability increases (learning rate in Table 8), both σ_n and σ_s increase but their ratio decreases (until the learning rate gets near one, when the standard deviation for old items begins to decline to zero); this decrease is inconsistent with the empirically obtained constant slope of the *z*-ROC functions. Moreover, because σ_n increases, the model predicts that the difference in d' for mixed lists will be greater than the difference in d' for pure lists; that is, the model predicts that the ratio of ratios will be greater than one, which is inconsistent with the empirical ratios. Third, as the vector length increases, the ratio of σ_n/σ_s decreases. This means that increasing vector length will only make the predictions for the ratio of the standard deviations in echo intensity worse. Fourth, the only conditions that give the right range of values for the ratio of standard deviations are for low vector lengths, long list lengths, and extreme learning rates (low or high). However, the ratio of standard deviations changes as a function of $(\mu_s - \mu_n)/\sigma_n$, which is inconsistent with the data.

Thus, one can conclude that in its current formulation, MINERVA 2 has significant problems in dealing with the behavior of the standard deviations of the echo intensity distributions as a function of strength of items in lists.

Murdock's TODAM Distributed, Single-Trace Vector Model

In the TODAM model, items are represented as vectors of attributes (or features). Memory is represented as a single vector that is a weighted sum of all studied items. Each feature of an input vector is sampled from a feature distribution that is normal, with a mean of zero and a variance of P/N, where P is the

Table 8
Sample Predictions of Hintzman's (1986) MINERVA 2 Model

Vector length	Learning rate	List length	σ_s	σ_n	σ_n/σ_s	μ_s/σ_s
20	.2	32	.033	.0228	.685	0.48
20	.4	32	.094	.0469	.498	0.92
20	.6	32	.171	.0686	.401	1.46
20	.8	32	.225	.0872	.387	2.40
20	.9	32	.213	.0952	.447	3.51
20	.99	32	.120	.0953	.794	8.11
20	.2	16	.029	.0161	.553	0.55
20	.4	16	.088	.0332	.377	0.98
20	.6	16	.164	.0485	.296	1.52
20	.8	16	.216	.0617	.287	2.51
20	.9	16	.202	.0673	.333	3.70
20	.99	16	.099	.0674	.678	9.79
40	.4	32	.054	.0158	.293	1.39
40	.8	32	.151	.0307	.203	3.48
40	.99	32	.066	.0361	.549	14.80
40	.4	16	.053	.0111	.210	1.42
40	.8	16	.150	.0217	.145	3.51
40	.99	16	.060	.0255	.423	16.10

Note. The expected value of μ_n is 0.

power of the vector and N is vector length (P is usually set to 1, so the variance is 1/N). At retrieval, a test vector is matched with the memory vector by computing the dot product. The equation for encoding is as follows:

$$\mathbf{M}_j = \alpha \mathbf{M}_{j-1} + p\mathbf{f}_j,$$

where p represents probabilistic encoding, α is a forgetting parameter, $\mathbf{f}$ is the new input vector, j is the number of the item in the list being encoded, and $\mathbf{M}$ is the memory vector. The equation for recognition matching is as follows:

$$\mathbf{f} \cdot \mathbf{M} = s,$$

where s is a scalar quantity denoting degree of match. When a pair of items is studied, each item of the pair is stored in the memory vector, and an association (another vector) formed by convolving the members of the pair of vectors is also stored.

The predictions for the standard deviations of old- and new-item familiarity values (denoted as s in the last equation) can be complicated but tractable (Murdock, 1982, 1983; Weber, 1988). To derive predictions, it is assumed that the stored vectors are independent. Then each stored item adds a component to the overall variance. In the simple case, with no forgetting and other parameters set to 1, each stored item that does not match the test item adds 1/N to the variance, and each item that matches the test vector adds 2/N (the 1/N and 2/N are the expected values of the variances). For example, for study lists with a length of 32 words, for a new test item the variance will be 32/N. For an old test item matching 1 studied item and not matching the other 31, the variance will be 33/N. Thus, the ratio of the variances is 32/33 and this is close to 1, that is, old- and new-item familiarity standard deviations are about the same. Thus, for long lists, the contributions of items that do not match an old test item dominate the contribution from the single matching item (in contrast with the Hintzman MINERVA 2 model, in which a single matching item can dominate all nonmatches). To explain learning in this model, Murdock (1989) introduced probabilistic encoding (as in MINERVA 2), such that the probability that a feature is encoded at study is a function of presentation time. This does not alter the qualitative predictions of the model.

The predictions of the model can be illustrated with the conditions of Experiment 1. In that experiment, a list of 16 pairs of words was studied, and single words were tested. We used a full version of the model with several additional parameters, such as weighting at retrieval, attention weights at encoding, and probabilistic encoding (Murdock, 1989). Moreover, several simplifying assumptions allow tractability (these do not affect the conclusions; for detailed analyses, see Shiffrin et al., 1990): Strong and weak pairs alternated at study, the tested item was encoded at study serial position 8, the attention weight γ was set to .7, the attention weight Ω was set to .9, and the retrieval weights q and w were set to 1.0. Using vector length 181, we were able to obtain d' values near those found in Experiment 1 for the pure strong condition (the probabilistic encoding parameter $p = 0.9$) and for the pure weak condition ($p = 0.35$); these are shown in Table 9. Also shown are results for increased vector length and for the forgetting parameter set to one. This allows evaluation of how the predictions would change as a function of these parameters.

The main predictions of the model under these conditions are as follows: First, the distributions of familiarity are approximately normal, which is consistent with the linear z-ROC functions in the data. Second, the old- and new-item standard deviations are nearly equal. This is inconsistent with the ratio from the data of about 0.8. Third, the model predicts that the difference between strong and weak d' values for mixed lists should be greater than that for pure lists; thus, the ratio of ratios of the d' values should be about 1.61 (see Table 9), much larger than the empirical values. As with MINERVA 2, the model makes this prediction because the variance in the new-item familiarity distribution changes from a pure weak list to a pure strong list (whereas for the mixed list, there is only one new-item familiarity distribution) and because the mean match values are the same for both pure and mixed conditions.

It should also be noted that Eich's (1982) model, as well as Pike's (1984) and Anderson's (1972, 1973) composite matrix and vector models, make the same prediction as TODAM: nearly equal standard deviations for old- and new-item familiarity values.

Gillund and Shiffrin's Search of Associative Memory Model (SAM)

Gillund and Shiffrin developed the SAM model to account for performance in recall and recognition. In the SAM model, items to be encoded are placed in a short-term buffer, and strengths between each of the items as a cue and each as an "image" in memory are increased as a function of time spent in the buffer. What is stored is how strongly each item, when it is presented as a cue (or test) item, is related to items in memory. There are three encoding parameters: b is the interitem strength between an item in the short-term buffer as a cue and images in memory of the other items in the buffer; c is the self strength between each item in the buffer as a cue and its own image in memory; and a is the context strength. Each parameter determines the amount of strength accumulated per unit time, so that in t seconds, at units of context cue to buffer item strength accumulate, ct units of self strength accumulate for each item in the buffer, and bt units of strength accumulate between each pair of items in the buffer. In addition, there is a parameter d that represents the preexperimental residual strength assumed to exist between any cue and any image in memory. Variability is introduced into the model by assuming that if the mean strength value computed from the encoding process is X, then the value of strength encoded is set according to the following probabilities:

$$.5X \text{ with } p = 1/3$$

$$X \text{ with } p = 1/3$$

$$1.5X \text{ with } p = 1/3.$$

It is this assumption that makes variability greater as strength becomes larger: If $X = 1$, the values stored are .5, 1, and 1.5, whereas if $X = 4$, the values stored are 2, 4, and 6, showing a greater absolute range and hence greater variability. This assumption (necessary for fitting the model to some data sets) is

Table 9
Sample Predictions of Murdock's (1982, 1989) TODAM Model

Condition	α	N	σ_s	σ_n	σ_n/σ_s	μ_s/σ_s
ps	.98	181	.386	.377	.977	1.388
pw	.98	181	.239	.235	.985	0.885
ms	.98	181	.324	.314	.969	1.654
mw	.98	181	.318	.314	.988	0.656
ps	1.0	181	.415	.406	.978	1.516
pw	1.0	181	.257	.253	.985	0.954
ms	1.0	181	.349	.339	.969	1.803
mw	1.0	181	.341	.339	.991	0.717
ps	.98	361	.273	.267	.979	1.962
pw	.98	361	.169	.167	.986	1.232
ms	.98	361	.229	.222	.969	2.337
mw	.98	361	.225	.222	.992	0.926

Note. The expected value of μ_n is 0. α = forgetting parameter; N = vector; ps = pure strong; pw = pure weak; ms = mixed strong; mw = mixed weak.

fundamentally responsible for the prediction that the variance in familiarity for old items is larger than for new items.

For recognition, the overall match between the cue (in practice, the combination of the context cue and the test item) and memory is computed; this is called *familiarity* and serves as the basis for a recognition decision. For images $1, \ldots, k$ in memory and cue I_i, the familiarity is defined as the following:

$$F(C, I_i) = \Sigma_k S(C, I_k)^{W_c} S(I_i, I_k)^{W_I},$$

where the sum is over all images $(1, \ldots, k)$ in memory, C represents the context cue, W_c is the weight given to the context cue, and W_I is the weight given to the item cue.

First, the resulting distributions of familiarity are approximately normal, which is consistent with the linear z-ROC functions. Second, to derive predictions from this model for Experiment 1 for pure lists (in which paired associates are studied), we make some standard assumptions about the encoding process (cf. Gillund & Shiffrin, 1984). It is assumed that the encoding buffer only contains the single pair of items under study and that interitem strength is accumulated between those two items only. Thus, the expression for d' can be derived (both from notes by R. M. Shiffrin, 1986, and our own calculations), as follows:

$$d' = K\{[(b/d)t + 1]^{1/2} + [(c/d)t + 1]^{1/2} - 2\}.$$

For lists of N items (N/2 pairs), the expression for the ratio of variances for noise to signal is as follows:

$$\sigma_n/\sigma_s = Nd/(Nd + bt + ct)$$
$$= 1/(1 + Lt),$$

where $L = (b + c)/(Nd)$. According to this expression, the ratio of standard deviations must change as a function of t (time spent in the buffer). The ratio cannot be 0.8 for both of two values of time that differ significantly, for example, $t = 1$ s, and $t = 5$ s. To demonstrate this, we performed the following computations: For a set of values of the parameters b/d and c/d, we set $t = 1$ for weak items and found a value of t that gave a value of d' two times larger for strong items. (Note that it is possible to reparameterize the expressions in terms of b/d and c/d so that these are the only model parameters that enter the expressions for d' and the ratio of variances.) Then we substituted the values of t into the expression for the ratio of standard deviations and found that over a range of parameter values, there were no values for b/d and c/d that gave nearly constant standard deviation ratios of about 0.8 (e.g., for varying values of b/d and c/d, values for the ratio of standard deviations for the strong (s) and weak (w) items were s = 0.48, w = 0.73; s = 0.53, w = 0.76; and s = 0.59, w = 0.80). Thus, the model cannot produce the constant ratio of standard deviations as a function of strength of old items found in the data. Third, as with TODAM and MINERVA 2, the model predicts that the difference in d' values between strong and weak items in a mixed list is greater than the difference in d' values between two pure lists, contrary to data (as documented by Shiffrin et al., 1990).

Differentiation Variant of SAM

To deal with the results from mixed/pure list experiments, Shiffrin et al. (1990) introduced a differentiation version of the SAM model. In this variant, it is assumed that the better encoded an item, the more clear are the differences between it and nonmatching test items. Thus, instead of the residual (preexperimental) strength of a distractor item to an image remaining constant, it will decrease the stronger the image is encoded into memory. This can be quantified with the assumption that the residual strength is an inverse function of context strength, $d = k/at$. With this assumption, the variance in the new-item distribution is independent of strength of old items; thus, the difference between d' values for weak and strong items is predicted to be the same in mixed and pure lists (i.e., the ratio of ratios is predicted to be one).

The ratio of variances for old- and new-item matches again provides strong predictions. The prediction for the ratio of variances is similar to that in the original SAM model, as follows:

$$\sigma_n/\sigma_s = Nk/(Nk + abt^2 + act^2)$$
$$= 1/(1 + Mt^2),$$

where

$$M = a(b + c)/Nk.$$

Comparing this expression with the earlier one for the original SAM model shows that the new version of SAM predicts a

variance ratio similar to the original model, with time squared in this case. For parameter values that produce a 2:1 difference in d' values, the ratio of the standard deviations is about the same as the original model. Thus, although this modification to the SAM model can account for the empirical mixed/pure list ratio of ratios, it is unable to produce adequate predictions of the ratio of standard deviations for the familiarities for old- and new-item distributions. It produces predictions that are essentially the same as the predictions of the old version of SAM.

General Discussion

In this article, an old issue in the application of signal detection theory to recognition memory is reopened. ROC curves have been used in the past to draw conclusions about recognition memory (Egan, 1958; Wickelgren & Norman, 1966, for short-term memory; see Murdock, 1974, for further discussion), but there has been no application of the ROC methodology to the recent global memory models. The three experiments presented in this article, as well as results from four previous studies in the literature, provide a consistent and simple pattern of results. First, the z-ROC functions do not deviate significantly from linearity. Thus, any model that predicts normal distributions will be consistent with this aspect of the data, although the distributions underlying the data are not necessarily normal because many distributions are consistent with linear z-transformed ROC curves (Lockhart & Murdock, 1970; Murdock, 1974). Second, the slope of the z-transformed ROC curve is constant at an average value of about 0.8 as a function of strength of the studied items. Assuming that the signal and noise distributions are normal, then this constant means that the noise distribution standard deviation is 0.8 times the signal standard deviation on average, varying from 0.5 to 1.0 for individual subjects. Third, the standard deviation in the noise distribution does not change as a function of strength of the old items in the study list.

To summarize (under the assumption that the underlying distributions are normal), (a) the new-item familiarity standard deviation is independent of the strength of old items; (b) the old-item familiarity standard deviation is independent of the strength of old items; (c) the new-item familiarity standard deviation is about 0.8 the value of the old-item familiarity standard deviation; and (d) the z-transformed ROC functions appear to be linear.

Overall, the data give a clear idea of the behavior of signal and noise distributions in recognition memory (assuming normal distributions). The pattern of results is summarized in Figure 8. We believe that these data are extremely constraining for the global memory models and also for connectionist models of memory and that they should provide some of the initial data to be tackled in any new modeling attempts. Of course, if a model does not predict normal distributions of familiarity, then the z-transformation of the predictions can be used to compare with the results reported here.

The current global memory models make strong predictions about the relative standard deviations of the signal and noise distributions. The composite vector models (and composite matrix models) such as TODAM make the prediction that the signal and noise distributions have about equal variance. This is because, although a match of an old test-item vector with another identical vector in memory has a higher variance than the match of a new test item against an item in memory, the old-item match is swamped by nonmatches from the comparisons with all of the other items in memory. The SAM model (and the SAM model with the differentiation assumption) has the problem that as mean familiarity of old items increases, the variance in familiarity must also increase because noise is introduced through multiplication and match values are summed over items. The MINERVA 2 model makes similar predictions to the SAM model: It predicts that as item strength increases, the variance in the signal distribution increases. This is because cubing large values of similarity (for matches) leads to larger results (activations) with larger variances than cubing small values (for nonmatches) of similarity.

The question now is where to go from here. It is possible that some new variants on the global memory models may have some success with the ROC data, but significant changes to the models will be required as well as considerable effort in refitting the models to the data bases from which they were developed. The class of connectionist models offers some promise simply because these models are relatively unexplored, and there are many possible architectures from which to choose. However, our initial attempts with the adaptive resonance theory architecture (Carpenter & Grossberg, 1987) and with backpropagation-based connectionist models (e.g., Kortge, 1990; Ratcliff, 1990) have not proved successful. Much more work is required on these connectionist models because the models have not been applied to the range of available memory data and because, unlike the global memory models, it is impossible to gain any understanding of the behavior of the variances in the match distributions as a function of experimental variables without simulations. Coupled with this problem is the possibility of myriads of variants on any individual model, any one of which might solve the problem. The challenge is to understand them.

The prescription for further theoretical development is to account for such central empirical effects as list length and presentation time and, at the same time, abide by the constraints on the variances of the familiarity distributions provided by the ROC methodology. Any model that produced predictions consistent with the ROC data naturally (and not just as a consequence of ad hoc assumptions) would be particularly appealing.

To conclude, we have models that apply over a wide range of paradigms with much success. Although they are flexible, powerful, and account for a range of data, the results we have presented here and found in the literature provide tight constraints and falsify the current versions of the models. The challenge is to use these results to modify the existing models or to develop the next generation of models, or both.

References

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences, 14,* 197–220.

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

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.

Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58-51). Hearing and Communication Laboratory, Indiana University.

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

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review, 92,* 1–38.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 19,* 1–65.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition, 13,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

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

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 *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 764–771). Hillsdale, NJ: Erlbaum.

Lockhart, R. S., & Murdock, B. B., Jr. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109.

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

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

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

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

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. (1989). Learning in a distributed memory model. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Floweree Symposium on Cognition* (pp. 69–106). Hillsdale, NJ: Erlbaum.

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

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., & 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., & McKoon, G. (1991). Using ROC data and priming results to test global memory models. In S. Lewandowsky and W. E. Hockley (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock, Jr.* (pp. 279–296). Hillsdale, NJ: Erlbaum.

Ratcliff, R., McKoon, G., & Tindall, M. (1992). An empirical analysis of ROC functions in recognition memory. Manuscript in preparation.

Sheu, C.-F. (in press). A note on the multiple-trace memory model without simulation. *Journal of Mathematical Psychology.*

Shiffrin, R. M., Ratcliff, R., & Clark, S. (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.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316–347.

Received March 18, 1991
Revision received September 30, 1991
Accepted October 30, 1991 ■