Journal of Experimental Psychology:
Learning, Memory, and Cognition
1994, Vol. 20, No. 4, 763-785

0278-7393/94/$3.00

Empirical Generality of Data From Recognition Memory Receiver-Operating Characteristic Functions and Implications for the Global Memory Models

Roger Ratcliff, Gail McKoon, and Michael Tindall

The experiments presented in this article examined the slope of the z-ROC (receiver-operating characteristic) function for recognition memory. The slope was examined as a function of strength and the variables study time, list length, word frequency, and category membership. For normal distributions of familiarity, the slope of the z-ROC is the ratio of the new-item to old-item standard deviations. R. Ratcliff, C.-F. Sheu, and S. D. Gronlund (1992) found that the slope was constant within standard error as a function of strength of encoding, which is inconsistent with the predictions of the global memory models. The results presented here extend this finding: The slope was constant as a function of strength of encoding, list length, and the number of related items from a category in the study list. Word frequency did affect the slope, but within a frequency class the slope was constant as a function of strength. The implications of these data for the global memory models, the attention likelihood model, and variants of these models are discussed.

This article presents six new experiments designed to add to the archival database for recognition memory and to test current models of recognition retrieval processes. Each experiment tested recognition memory for words; lists of single words were studied, and each study list was followed by a list of test words. For each test word, subjects were asked to decide whether it had appeared in the study list and to indicate how confident they were of their decision. The confidence judgments were used to construct a receiver-operating characteristic (ROC) curve to show how discrimination of studied from nonstudied test words changes as a function of different criterion settings (i.e., different confidence levels). We report the effects on ROC curves of word frequency, category membership, study-list length, presentation rate, individual differences among subjects, and criterion shifts. The aim is to generalize and extend earlier empirical results and to examine theoretical implications of the results for the global memory models (e.g., Gillund & Shiffrin, 1984; Hintzman, 1986, 1988; Murdock, 1982; see also Ratcliff, Sheu, & Gronlund, 1992). Each of the reported experiments provides a specific test of one of the models as well as constraints for all of the models.

The global memory models (cf. Gillund & Shiffrin, 1984; Hintzman, 1986, 1988; Murdock, 1982) have been applied to a number of experimental procedures, including free recall, recognition, frequency judgments, and category judgments. Recognition memory is a particularly good domain in which to compare the models because they all make predictions about recognition performance. The models assume that a test item presented for recognition contacts all items in memory to determine the degree of match (familiarity) between it and memory. The familiarity value, in turn, determines the old–new (studied or nonstudied) judgment; the higher familiarity, the more likely an "old" response.

Roger Ratcliff, Gail McKoon, and Michael Tindall, Department of Psychology, Northwestern University.

This research was supported by National Institute of Mental Health Grants MH 44640 and MHK00871 to Roger Ratcliff and U.S. Air Force Office of Scientific Research Grant 90-0246 (jointly funded by the National Science Foundation) to Gail McKoon. We are grateful for extensive useful comments from Douglas Hintzman, Bennet Murdock, Kevin Murnane, and an anonymous reviewer.

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

z-ROC Functions

The experiments reported in Ratcliff et al. (1992) have previously addressed two major predictions of the global memory models. The first concerns how much variability there is in the values of familiarity for old versus new test items. The ratio of the standard deviations of familiarity can be obtained from standard signal detection theory by using confidence judgment data. The z transforms of the hit rate and false-alarm rate for each level of confidence are plotted against each other (z_h vs. z_{fa}) to produce a z-ROC curve. If the underlying distributions of familiarity values are normal, then the slope of the z-ROC is the ratio of the new-item standard deviation to the old-item standard deviation, σ_n/σ_o. The global memory models assume normal distributions (either directly or by the central limit theorem applied to sums of values under discrete assumptions), so for these models, the slope of the z-ROC provides a direct measure of the ratio of the standard deviations of old- and new-item familiarity. Two measures of d' are used in this article: $d'_1 = (\mu_o - \mu_n)/\sigma_n$, the standard definition, and $d'_2 = (\mu_o - \mu_n)/\sigma_o$, which is the intercept of the z-ROC equation, $z_h = (\sigma_n/\sigma_o)z_{fa} + (\mu_o - \mu_n)/\sigma_o$. Note that either can serve as a d' measure (e.g., McNicol, 1972).

The results presented by Ratcliff et al. (1992) showed a roughly straight-line z-ROC function with a slope of about 0.8 for both weakly encoded items and strongly encoded items. This constant value of the slope of the z-ROC across strength values is difficult if not impossible for the current global memory models to fit. The search of associative memory

(SAM) model (Gillund & Shiffrin, 1984) and the MINERVA 2 model (Hintzman, 1986, 1988) both predict that the slope should decrease with strength, whereas the theory of distributed associative memory (TODAM, Murdock, 1982) predicts the slope to be constant with a value near 1 regardless of strength. Equivalently, TODAM predicts that the old- and new-item familiarity distributions should have about the same standard deviations for all levels of familiarity, whereas SAM and MINERVA 2 predict that the standard deviation of old-item familiarity should increase in relation to the standard deviation of new-item familiarity as familiarity increases.

List-Strength Effect

The second major issue that the experiments in Ratcliff et al. (1992) have addressed is the list-strength effect. Most of the global memory models predict that performance on weakly encoded items will be hurt by including strongly encoded items with them in the study list (a mixed-strength list) as compared with performance for the weak items when there are no strong items in the study list (a pure list). This predicted decrement in performance has been labeled the *list-strength effect* (Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990), and although it is found in free recall, it is not found in recognition (Ratcliff et al., 1990).

The list-strength prediction for the global memory models boils down to a prediction about the standard deviation of the familiarity values of new items in pure lists. In a standard list-strength design, the strength of encoding of an item is manipulated by varying study time for the item or varying the number of repetitions of the item. There are two list types: pure lists with a single strength value for all of the items and mixed lists with one strength value for some items and a different strength value for other items. The statistic chosen to measure the list-strength effect is the ratio of ratios (Rr) of d' values, where Rr = ([mixed strong d']/[mixed weak d'])/([pure strong d']/[pure weak d']). This statistic is chosen because for most of the global memory models, the ratio reduces to a simple ratio of standard deviations (see Shiffrin et al., 1990), as follows: In the models, mean familiarity does not depend on the list composition, mixed or pure, so the mean familiarity of strong items in a mixed list equals the mean familiarity of strong items in a pure list and the mean familiarity of weak items in a mixed list equals the mean familiarity of weak items in a pure list. The standard deviation of the familiarity values for new items differs between a pure weak list and a pure strong list, but for a mixed list, the standard deviation for new items can have only one value. Thus the ratio of ratios reduces to Rr = (SD new pure strong)/(SD new pure weak). Because most of the global memory models predict that SD(new pure strong) > SD(new mixed) > SD(new pure weak), they predict that Rr will be greater than 1.

Data show that there is no list-strength effect, that is, that the ratio of ratios Rr is about 1, not greater than 1 (Murnane & Shiffrin, 1991; Ratcliff et al., 1990; Ratcliff et al., 1992; Yonelinas, Hockley, & Murdock, 1992; see also Shiffrin et al., 1990 for a discussion and presentation of a variant of SAM that does predict a ratio of ratios equal to 1). Simple artifactual explanations of the failure to find a list-strength effect such that the weak items in a mixed list get extra rehearsal time by borrowing from the strong items have been ruled out (see also Yonelinas et al., 1992).

The list-strength measure and the z-ROC slopes together give a picture of the behavior of the standard deviations in familiarity values for strong and weak items in pure and mixed lists under the normal distribution assumptions of the global memory models. The slope of the z-ROC curve provides the ratio of new- to old-item standard deviations, and the mixed–pure list comparison gives the ratio of the standard deviations of new items for pure weak and pure strong encoding conditions. For models other than the global memory models (e.g., Glanzer & Adams, 1990), predictions for the shape of the z-ROC function and the ratio of ratios can also be generated (see below).

Experimental Variables

The aim of the global memory models is to account for a wide range of kinds of data from a wide range of experimental procedures. The experiments that have examined list strength and z-ROC functions have focused mainly on the strength manipulation. The experiments in this article serve two main purposes: first, to extend the database on these phenomena and, second, to provide specific tests of models. To this end, the experiments were designed to manipulate presentation rate for words in the study lists, length of the study list, word frequency, and category membership.

Rate of presentation was varied by Yonelinas et al. (1992) in an examination of the list-strength effect. They used presentation rates varying from 50 ms up to 200 ms per item. Initially, they found a list-strength effect when the items of different strengths (different presentation rates) were randomly ordered in a mixed-strength study list, but when a blocked design was used, there was no list-strength effect. The differentiation model proposed by Shiffrin et al. (1990) predicts a list-strength effect at very rapid presentation rates. The experiments presented here further examined list-strength predictions and also included confidence judgments so as to produce z-ROC curves as a function of rate of presentation, allowing examination of their slopes at low learning levels.

The length of the study list is a central variable in the memory models. As length increases, performance decreases, and the models explain this as a result of increasing variability in the familiarity values of new test items (SAM and TODAM) or increasing forgetting of items studied early in the study list (TODAM). List length is also at the center of controversies surrounding the list-strength results and the resulting modifications of models designed to account for them (see Murnane & Shiffrin, 1991). For example, Murdock and Kahana (1993) presented a modification of the TODAM model in which a test item is matched not just against items from the immediately preceding study list but also against items from many preceding lists (a *continuous memory* assumption). This makes the standard deviations of new-item and old-item familiarity values almost independent of the composition of the current list and so correctly predicts that there will be no list-strength effect. However, as in many models, fixing one thing has the

possibility of breaking something else, and this continuous memory version of TODAM is tested later.

The materials variable word frequency has traditionally played an important role in memory research because it has large and opposite effects on recognition and recall performance and thus provides a benchmark against which to test models. Recognition of low-frequency words is usually better than recognition of high-frequency words, whereas recall of low-frequency words is usually worse than recall of high-frequency words (except in lists of mixed frequency; Gregg, Montgomery, & Castano, 1980). In general, in recognition, high-frequency new test words have a higher false-alarm rate than low-frequency new test words, whereas high-frequency old test words have a lower hit rate than low-frequency old test words. This means that the familiarity values of high-frequency test words are, in general, nearer to the decision criterion than the familiarity values of low-frequency test words. This symmetrical behavior has been termed the *mirror effect* by Glanzer and Adams (1985, 1990). The global memory models have difficulty predicting this effect, but Glanzer and Adams (1990; Glanzer, Adams, & Iverson, 1991) have proposed an alternative model, the likelihood ratio model.

To test this model, Glanzer and Adams (1990) extended the examination of word frequency to investigate the effect of frequency on the slopes of *z*-ROC curves. They found a systematic effect of frequency; for example, high-frequency slopes were nearer 1 than were low-frequency slopes. For the global memory models, one possible hypothesis is that decreasing word frequency affects familiarity in the same way as increasing study time; that is, it affects degree of match, and perhaps variability. To examine this hypothesis, the experiments presented here jointly manipulated strength through study time and word frequency to determine whether word frequency affects *z*-ROC curves in the same way as strength of encoding (the Glanzer & Adams, 1990, results and the Ratcliff et al., 1992, results suggest that this will not be the case).

So far, two ways to manipulate the degree of match between a test item and memory have been described: varying the strength of encoding for the item and varying word frequency. Another way is to vary the similarity of the test item to other items in the study list. This can be accomplished by using sets of words from the same semantic category (e.g., vehicles). A study list contains several words from the same category, and the test list contains both old and new words from the category. The new test words from the same category as studied items should have a higher familiarity value than other new words. We investigated whether this manipulation of familiarity has the same effect on *z*-ROC functions as other manipulations of familiarity in Experiment 6.

In some of the experiments here, we tested single subjects for a number of sessions to examine individual differences in *z*-ROC functions. Ratcliff et al. (1992) found that the slope of the *z*-ROC function was about constant as strength of the studied items varied, constant at a value of about 0.8. For modeling, it is important to know whether all individuals share this same constant value. If the value of the constant slope is different for different subjects (e.g., 0.9 for one subject and 0.6 for another), then the models must have the flexibility to cover a range of such individual differences.

Important Hedge

If the distributions of familiarity values underlying the ROC functions are normal, then the *z*-ROC function is linear and the slope equals the ratio of the new- to old-item standard deviation (σ_n/σ_o) and the intercept is a d' measure $(\mu_o - \mu_n)/\sigma_o$. Most of the analyses in this article are presented in terms of the slope and intercept of the fit of a straight line to an empirical *z*-ROC curve or in terms of the fit of the standard normal model to the raw confidence judgment scores. The global memory models assume the underlying familiarity distributions to be normal, and so the slopes and intercepts of the *z*-ROC functions relate directly to the means and standard deviations of the familiarity distributions of the models.

However, other distributions can also produce roughly linear *z*-ROC functions (e.g., Lockhart & Murdock, 1970), and these distributions may not carry the same implications for the standard deviations of old and new test item familiarity values as the normal distributions from the global memory models. Thus, finding that the *z*-ROC curves are linear does not necessarily mean that the underlying distributions are normal. But the results presented here in terms of slopes and intercepts can still be used to test alternative models that are based on nonnormal distributions by producing predictions from those models for the *z*-ROC functions. An analogy for using *z*-ROC curves to summarize data in this way is the use of a theoretical distribution (e.g., the convolution of normal and exponential distributions) to summarize the behavior of reaction time distributions (Ratcliff, 1978; Ratcliff, 1979; Ratcliff & Murdock, 1976), which has proved useful both empirically and theoretically. Confidence judgment data from all of the experiments and all individual subjects in Experiments 3 and 5 are presented in the Appendix.

Experiments 1 and 2

Rate of Presentation

One of the global memory models contradicted by failures to find a list-strength effect was TODAM, the model proposed by Murdock (1982). Yonelinas et al. (1992) argued that the failure to find a list-strength effect was due to a rehearsal strategy used by subjects during study: In a mixed list, they borrowed rehearsal time from strong items (items presented for longer study time) and used it for weak items (items presented for a shorter study time). Objecting to the various rehearsal control conditions used by Ratcliff et al. (1990), Yonelinas et al. claimed that a better control was to present items so fast that rehearsal would not be possible. Their initial experiments used rates of presentation as fast as 100 ms per word, and they did find a list-strength effect: Strong items in a mixed list were better recognized than strong items in a pure list, and weak items in a mixed list were more poorly recognized than weak items in a pure list. The experiments used a random presentation procedure so that study items with short and long presentation times were randomly intermixed. We thought this could lead to "reverse" rehearsal borrowing: When items were whizzing by at such fast rates, subjects might have given up on the fast items and used their time to rehearse the slow items. Experiments 1 and 2 tested this hypothesis (a

preliminary report was presented in Ratcliff & McKoon, 1991) by using presentation times of 50, 100, 200, and 400 ms, and our results were confirmed by Yonelinas et al. (1992, Experiment 6).

In addition to testing the list-strength effect over a range of presentation times, we wanted to examine the slope of the z-ROC curve to determine when it begins to increase (as strength is decreased) from the constant value of 0.8 found by Ratcliff et al. (1992) to the value 1, which must be obtained when the hit rate equals the false-alarm rate at chance performance. To obtain ROC functions, we used a confidence judgment procedure. Subjects were required to make a recognition response on a 6-point scale with values ranging from *very sure old* (6) to *very sure new* (1). The response probabilities in each confidence category were used to construct the z-ROC curve by calculating cumulative probabilities successively from the right-hand side of the distribution and then plotting the z transforms of the hit cumulative values versus the z transforms of the false-alarm cumulative values (see McNicol, 1972).

Data Analysis

We used two methods to analyze z-ROC functions, each with some weakness, but each providing a check on the other. The first method was multiple regression. If the distributions of familiarity values for old and new test items are normal, then the slope and intercept of a single z-ROC curve can be estimated with simple linear regression, fitting z_h against z_{fa}. To generalize for the situation in which there is a different z-ROC curve for each experimental condition (e.g., strong vs. weak items), we used multiple regression, for which an equation was defined to represent the effect of each independent variable on the z_h scores in terms of z_{fa}. For a simple pure–mixed, strong–weak list design, the equation will be

$$z_h = b0 + b1\, z_{fa} + b2\, p + b3\, p\, z_{fa} + b4\, s + b5\, s\, z_{fa} + b6\, p\, s + b7\, p\, s\, z_{fa},$$

where $p = 0$ for a pure list, $p = 1$ for a mixed list, $s = 0$ for strong items, and $s = 1$ for weak items. The variables p and s are called dummy variables, and they allow for possible systematic effects of the independent variables on performance. For example, to test for an effect of strong versus weak on the intercept, the null hypothesis would be $b4 = b6 = 0$ (see Draper & Smith, 1966; Kleinbaum, Kupper, & Muller, 1988, for a discussion of this kind of use of multiple regression). F tests for the significance of the coefficients $b1$–$b7$ can be generated to examine the effects of the independent variables on slopes and intercepts. The data entered are the z values for the hit and false-alarm rates (the five pairs of z values for each condition, derived from the cumulative confidence judgment data). The problem with the multiple regression method is that it is based on the assumptions that the z_{fa} values are fixed and that all of the variability is in the z_h values. Thus, the F values obtained are not exact.

The other method we used is a maximum likelihood solution that assumes variability in both z_h and z_{fa}. This method uses an approximation of the normal distribution (the logistic distribution) to fit a distribution of raw confidence scores, producing estimates of the slope and intercept of the z-ROC, as well as standard errors in those estimates. The algorithm for this is called EPCROC and was presented by Ogilvie and Creelman (1968). Although this method has the appropriate statistical properties, it does have a practical limitation for use with the experimental designs presented in this article. The limitation is that the method produces estimates of the standard errors on the slope and intercept of the ROC function for each separate experimental condition, so comparison of the different conditions requires multiple comparisons of the estimated slopes and intercepts. Essentially the problem is like that of performing multiple t tests on one set of data without using a method for adjusting the significance level.

To draw conclusions from the experimental results for the ROC curves, we used a combination of the two methods. The multiple regression method was used to test hypotheses about the effects of the variables, and the EPCROC method was used to obtain parameter estimates and the standard errors in those estimates and to provide a check on the multiple regression method. In almost all cases, the two methods gave approximately equivalent results (e.g., within 1 standard error), thus we report the results from only one, the EPCROC estimates. The standard errors derived from EPCROC were consistent with the significance levels of the F tests, and the parameter estimates from multiple regression and EPCROC were in close agreement.

For the list-strength effect, we used an explicit F test derived from the multiple regression model given earlier to look for a mixed–pure by strong–weak interaction, testing the null hypothesis that $b6$ was zero. We also used EPCROC to determine the ratio of ratios of d' values. The ratio of ratios can be computed from d'_2, the intercepts of the z-ROC functions ([mixed strong intercept]/[mixed weak intercept])/([pure strong intercept]/[pure weak intercept]). The ratio of ratios can also be computed from d'_1 by using the intercept divided by its associated slope (σ_n/σ_o; i.e., the σ_o divides out of d'_2). We report both of these ratios of ratios for completeness.

Method

Subjects and materials. Subjects were paid volunteers from the Northwestern University undergraduate population. There were 16 subjects for Experiment 1 and 15 subjects for Experiment 2. Each participated in one 45-min session, with each session consisting of 20 study-test lists. The materials were words from a pool of 1,650 two-syllable common English words not more than eight letters long (an extended version of the Toronto word pool; e.g., Ratcliff & Murdock, 1976).

Procedure. Stimuli were presented on a Goldstar computer monitor with a fast P4 phosphor, and responses were collected on the keyboard of a PC computer. There were three kinds of study lists: pure weak lists, pure strong lists, and mixed lists. For Experiment 1, in a pure list, each of 32 words was presented once for an equal amount of time, 50 ms per word in a weak list or 200 ms per word in a strong list. In a mixed list, sequential blocks of words had different study times: the first block of 4 words at 50 ms, the next block of 12 words at 200 ms, the next block of 12 words at 50 ms, and the last block of 4 words at 200 ms (the first and last blocks were buffer words), or the reverse ordering of study times. For Experiment 2, the weak and strong study times

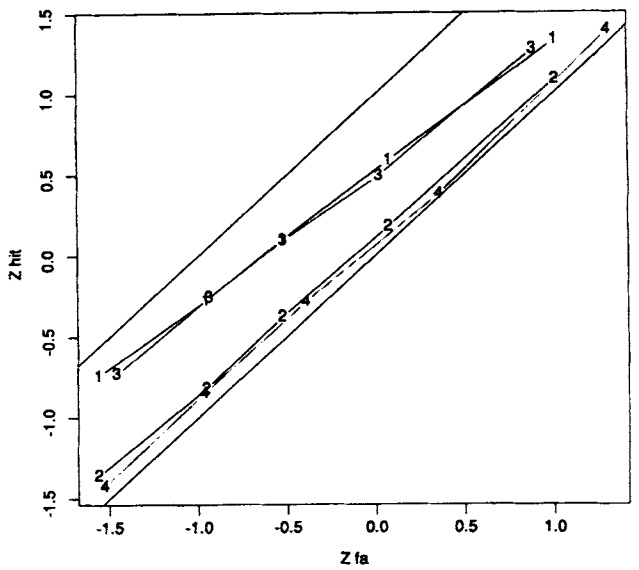

Figure 1. *Z*-transformed receiver-operating characteristic curves for the 50-ms and 200-ms groups in Experiment 1. Curve 1 = mixed strong condition, Curve 2 = mixed weak condition, Curve 3 = pure strong condition, and Curve 4 = pure weak condition. The diagonal straight lines are for comparison and have a slope of 1. fa = false alarm.

were 100 ms and 400 ms instead of 50 ms and 200 ms, respectively. There was no interstimulus interval.

A recognition test list followed each study list. There were 64 total test items, 32 old and 32 new presented in random order. Subjects were instructed to respond on a 6-point scale from *sure old* (6), *probably old* (5), *maybe old* (4), *maybe new* (3), *probably new* (2), to *sure new* (1). The keys on the keyboard used for the confidence judgments were the "x"

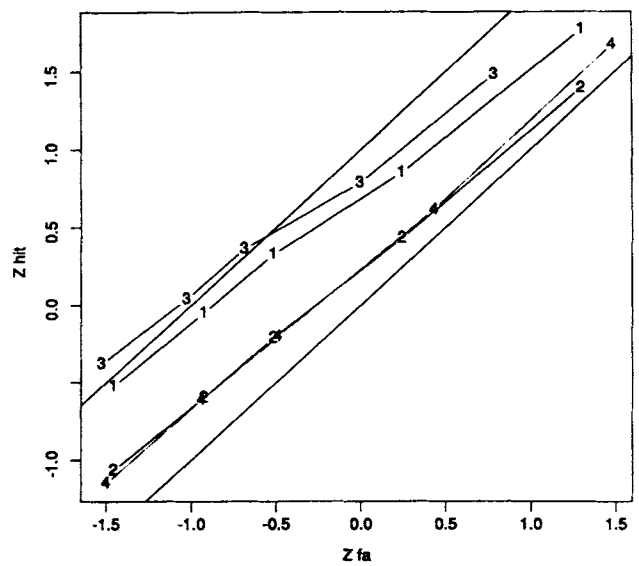

Figure 2. *Z*-transformed receiver-operating characteristic curves for the 100-ms and 400-ms groups in Experiment 2. Curve 1 = mixed strong condition, Curve 2 = mixed weak condition, Curve 3 = pure strong condition, and Curve 4 = pure weak condition. The diagonal straight lines are for comparison and have a slope of 1. fa = false alarm.

Table 1
Slopes and Intercepts for 50-ms and 200-ms Study Times in Experiment 1

Condition	Slope		Intercept	
	M	*SD*	*M*	*SD*
Mixed strong	0.794	0.022	0.591	0.032
Mixed weak	0.951	0.025	0.121	0.029
Pure strong	0.831	0.026	0.570	0.034
Pure weak	0.970	0.028	0.076	0.034

through "m" keys on the bottom row of the keyboard. There was a 300-ms delay between each response and presentation of the next test item. Subjects were instructed to try to distribute their responses over all of the judgment categories and to avoid using just one or two.

Results and Discussion

Responses faster than 250 ms and slower than 5,000 ms were eliminated from the analyses, as were responses to test items from the first and last four positions in the study lists (the buffer items) and the first position in the test list. Figures 1 and 2 show linear *z*-transformed ROC curves for the two experiments. The estimates of the slopes and intercepts and the standard errors in the estimates (*SD*) computed from EPCROC are shown in Tables 1 and 2.

For the items presented for 50 ms in the study list, subjects were hardly able to discriminate whether a test item was in the study list, as expected. The intercept of the ROC curve (a d' measure equal to μ_s/σ_s for normal distributions) was near zero (i.e., d' was about 0.1). Note, with d' near zero, the slope of the ROC curve must approach 1 because there is no discrimination and the hit rate must equal the false-alarm rate. Subjects reported that they could identify only about four or five words per study list. This suggests that encoding produced a probability mixture of a few weakly encoded words and many unencoded words. In this case, the slopes would be at 1 and d' values at 0 for the unencoded items, and these would be mixed with a few items with higher intercept and slope less than 1.

Subjects showed somewhat better discrimination with the 100-ms presentation rate in Experiment 2; the slopes of the ROC curves were about 0.9, and d' was about 0.25. For the 200-ms presentation time in Experiment 1, slopes were about 0.8 to 0.85, and d' was near 0.6. For the 400-ms rate, slopes were about 0.8, and d' was about 0.9.

For Experiment 1, the differences in slopes and intercepts due to the different rates of presentation were significant with

Table 2
Slopes and Intercepts for 100-ms and 400-ms Study Times in Experiment 2

Condition	Slope		Intercept	
	M	*SD*	*M*	*SD*
Mixed strong	0.813	0.024	0.795	0.033
Mixed weak	0.876	0.023	0.250	0.031
Pure strong	0.782	0.028	0.974	0.037
Pure weak	0.921	0.027	0.259	0.035

the general linear model described earlier, $F(2, 12) = 20.4$ and $F(2, 12) = 438.7$, respectively (all significant F values have $p < .05$). There was a marginal difference between pure and mixed slopes, $F(2, 12) = 2.8$, $p = .10$, and a significant difference between pure and mixed intercepts, $F(2, 12) = 16.3$. There was a significant effect of list strength (defined as an interaction between pure vs. mixed and strong vs. weak) on the intercept, $F(1, 12) = 11.8$.

For Experiment 2, strength had an effect on both slope, $F(2, 12) = 54.3$, and intercept, $F(2, 12) = 532.8$, and the effect of pure versus mixed study lists on slopes and intercepts was marginally significant, $F(2, 12) = 2.5$, $p = .13$, and $F(2, 12) = 1.9$, $p = .19$, respectively. There was a nonsignificant effect of list strength on the intercepts, $F(1, 12) = 1.5$.

For both experiments, the ratio of ratios from the mixed–pure list comparison was less than 1. When the ratios of ratios were based on the intercepts (μ_s/σ_s; from Tables 1 and 2), the values were 0.70 and 0.92, respectively, and when based on the intercepts divided by the slopes (μ_s/σ_n), the values were 0.69 and 0.84, respectively. The reason for the differences in the estimates of the ratios is that for very low d' values, a little random variability leads to a large change in the ratio (e.g., a d' difference from 0.1 to 0.2 can lead to a doubling of the ratio).

The results of Experiments 1 and 2 show that as study time was reduced, old–new discriminability (d') fell to near zero, and the slope of the z-ROC approached 1. As study time increased, the increase in d' was rather rapid, and by the time d' had increased to 0.5, the slope had approached the asymptotic value reported in Ratcliff et al. (1992). The ratio of ratios remained at or below 1.0 for these two experiments, replicating previous experiments and demonstrating no hint of a list-strength effect (which would require a value greater than 1). The values of the ratio were extremely low for Experiment 1, reflecting numerical instability because the d' values on which the ratios were based were near zero for the 50-ms condition (see also Loftus, 1974).

Rapid presentation rates also provide a test of the differentiation version of the SAM model proposed by Shiffrin et al. (1990; pointed out by R. M. Shiffrin, personal communication, January 1990). The differentiation model assumes that as strength increases, an item becomes differentiated from other items, and this is implemented as the residual strength of the item to all other items being reduced. Specifically, as study time is increased, context strength (the strength between a list context element and a studied item) increases and residual strength (preexperimental strength of connection between two items) increases up to some point at which residual strength begins to decrease as study time is increased more. The effect of residual strength decreasing counteracts the increasing context strength and produces the prediction that there will be no list-strength effect (see Shiffrin et al., 1990, Figure 1). Thus, at small values of study time, a list-strength effect should have been obtained because the residual strength is still increasing, but none was found. This result does not rule out the differentiation model, but it means that for the differentiation model to be correct, the rise of context strength must take place over a study-time range other than the 50–400 ms used in these experiments.

Experiment 3

List Length

The global memory models predict that increasing the length of a study list will increase the variability of old-item familiarity values. For example, in SAM, as list length increases, the number of images in memory increases, which leads to larger variance in both the old- and new-item familiarity values. The other models make similar predictions: List length increases the number of items in memory leading to increased variance in familiarity. For all of the models, the result of increased variance is a decrease in d'.

To examine these predictions, Experiment 3 measured the slopes of z-ROC curves as a function of list length. Because list length is a between-lists variable and because subjects could easily become bored in the longer lists and not work to encode the items, we decided to run the experiment with a group of motivated subjects who would participate in 10 sessions each (they were motivated by payments that were based in part on performance). This design also provided the z-ROC slopes for individuals to determine whether it is constant or whether it differs for individual subjects.

Manipulation of list length provides for a test of a new version of TODAM proposed by Murdock and Kahana (1993) in which it is assumed that memory is continuous across the study lists of an experiment; memory is not reset after each study-test list as in earlier versions of TODAM. Because of the continuous memory assumption, the variability in the familiarity or match between a test item and memory is determined by the contents of all of memory and so is largely independent of the composition of the last studied list. Thus, this version of TODAM correctly predicts that there should be no list-strength effect. However, the continuous memory assumption also leads to the prediction that performance on an old-test item as a function of the lag between its study and test positions should not vary with list length. To produce the list-length effect, long lists have longer study to test lags on average than short lists. We tested the prediction that serial position functions for the same study test lags will overlay each other by examining performance as a function of study and test positions. (The new version of TODAM still predicts the slope of the z-ROC function to be close to 1.)

Method

Subjects. There were 7 subjects from the Northwestern University undergraduate population who were paid for participation in the experiment. Five subjects completed 10 sessions preceded by 1 practice session, and 2 subjects completed 7 sessions preceded by 1 practice session.

Procedure. The materials were the same as in Experiments 1 and 2. Study lists contained 8, 16, or 32 pairs of words (i.e., 16, 32, or 64 single words) presented at a rate of 1 s or 3 s per pair. Only pure lists were used in this experiment, so all pairs in a list were studied at the same rate. List length (long, medium, or short) was cued before study. Test lists consisted of 32, 64, or 128 single words with equal numbers of old and new test items in random order. Subjects responded on the same 6-point scale as in Experiment 1, with a 300-ms pause between a

response and the next test item. Pilot studies found that some subjects had very few responses in one or another of the response categories. To avoid this, feedback as to the number of responses per response category was presented after each list, and subjects were instructed to equate the number of responses per category as much as possible over the experiment (given, for example, that short lists would have fewer low-confidence responses than would long lists). There were 18 lists per session, 3 lists of each of the six types.

Results and Discussion

Data analysis was carried out as in Experiments 1 and 2, and the data from the first position in the test list were eliminated. To examine performance as a function of list length with the number of items intervening between study and test positions equated, we used data only from the last 16 studied words and Items 2–32 in the test list. Note that this means the number of observations for old items from longer lists is smaller than the number of observations for old items from shorter lists because, for the longer list, the first 32 test positions include other studied items besides the last 16 (and these other items were discarded from the analyses).

Figure 3 shows individual subject *z*-ROC curves for the six conditions. In general, they group as follows: The three upper-left lines represent strong items (3 s study time) and the lower three weak items (1 s study time). Within each group of three, generally, the upper left represents results from short lists, whereas the lower right represents results from long lists. Table 3 shows the results from the EPCROC fits to these data, which were based on the averages of the parameters from an EPCROC fit to the data from each individual subject. The

Table 3
Slopes and Intercepts for List Length (LL) in Experiment 3

Condition	Slope		Intercept	
	M	*SD*	*M*	*SD*
LL, 16 items				
Weak	0.769	0.033	1.545	0.036
Strong	0.770	0.049	2.274	0.045
LL, 32 items				
Weak	0.755	0.035	1.195	0.042
Strong	0.814	0.054	1.934	0.051
LL, 64 items				
Weak	0.792	0.046	0.971	0.053
Strong	0.969	0.071	1.718	0.058

Note. Study and test positions are equated for Items 2–32 in the test list and for the last 16 items studied. These values are averaged over the data of individual subjects, and the average standard deviation is obtained by averaging the standard deviation for each subject and dividing by the square root of the number of subjects (to give the standard error of the mean).

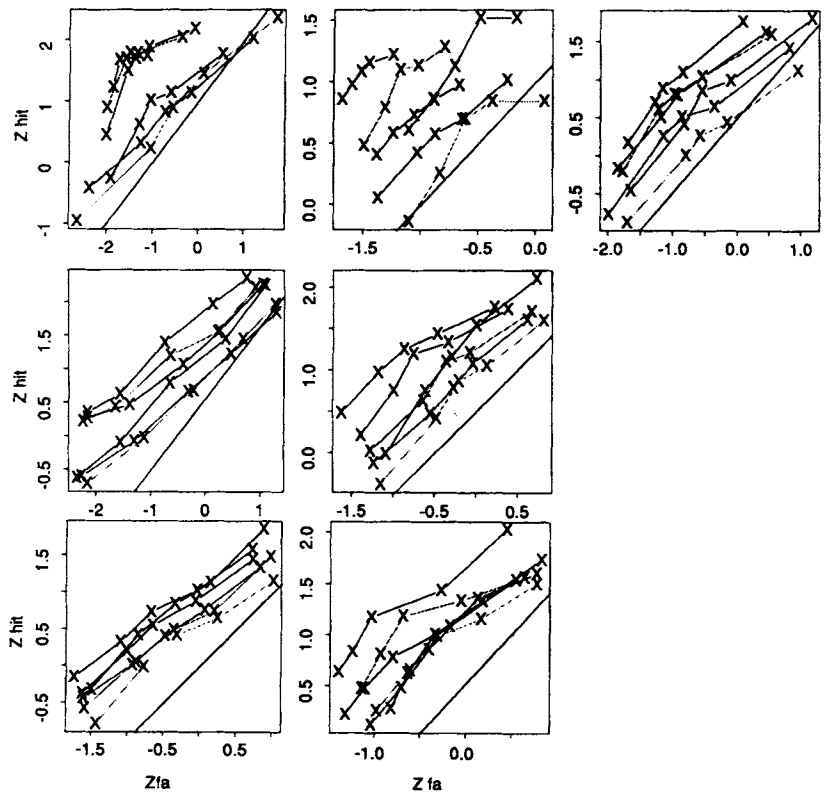

Figure 3. *Z*-transformed receiver-operating characteristic curves for individual subjects in Experiment 3. The six curves for each subject represent three list lengths crossed with two values of strength (study time). The same study and test position ranges are used for each list length. When the curves separate, the strong are at the upper left and the weak at the lower right. The order within the group of three (upper left, lower right) is upper left, List Length 16; middle, List Length 32, and bottom, List Length 64. The curves are presented to show separation and linearity. The diagonal lines have a slope of 1. fa = false alarm.

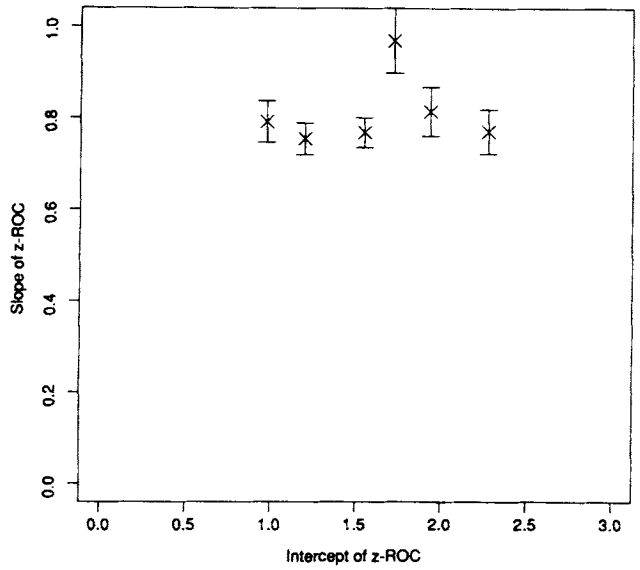

Figure 4. Slope of the *z*-transformed receiver-operating characteristic (ROC) plotted against the intercept for two strength values (the three right-hand points represent strong items, the three left-hand points represent weak items) and three values of list length (the left three points are weak, the right three points are strong, and within each group list lengths are 64, 32, and 16 from left to right). The error bars represent 1 standard error in the slope.

main result is that there is little effect of list length on the slope of the *z*-ROC function. Figure 4 shows this relationship: The slope values are plotted for the three list lengths and two levels of strength. Apart from one point (long study time for List Length 64), there is no effect of list length on slope, and none of the functions decrease as a function of strength (in fact each increases nonsignificantly).

Regression analyses showed that there was a significant effect of study time on intercept, $F(3, 18) = 29.2$, but not on slope, $F(3, 18) = 0.4$, and there was a significant effect of list length, $F(2, 18) = 12.1$, on intercept but not on slope, $F(2, 18) = 0.01$, (all significant $ps < .05$).

Because items were studied in pairs, it might be thought that a recall mechanism could have been used in conjunction with recognition; for example, when recognition familiarity was low, another list member might be recalled and used to increase confidence that the test item was old. However, response times in these experiments (especially with multisession subjects) were in the range 700 to 800 ms, and this seems too fast for a multistep process to take place (see Gronlund & Ratcliff, 1989).

Figure 5 shows smoothed serial position effects (using the 4[3RSR]2H method twice; Tukey, 1977, chapters 7 and 16) for the six conditions. The critical result for testing the new version of TODAM is that the serial position functions do not lie on top of each other; performance is not constant as a function of study-to-test lag. Averaging over the same first 32 test positions for each list length (as we did in the analyses and as shown in Figure 5), items from short lists were better recognized than items from long lists. The version of TODAM that assumes test items are matched not only against the immediately preceding study list but also against all other preceding test lists (Murdock & Kahana, 1993) is contradicted by this finding.

This experiment shows that the list-length manipulation (decreasing list length) appears to operate on the *z*-ROC functions in exactly the same way as a strength manipulation of study time or number of repetitions. For all of these variables, the slope of the *z*-ROC appears constant across levels of strength. In addition, the serial position functions appear to disconfirm a strong prediction of the version of TODAM that incorporates the continuous memory assumption.

Experiment 4

Word Frequency

Experiment 4 was designed to examine whether the slope of the *z*-ROC curve depends on word frequency and how word frequency interacts with strength. Glanzer and Adams (1990) found that *z*-ROC slopes were smaller for low-frequency words than for high-frequency words. In our experiments, there were two levels of word frequency (high and low) and two levels of strength (manipulated by two values of study time) in a mixed–pure list design. A mixed–pure design was used so that the list-strength effect could be examined separately for high- and low-frequency words, and the strength manipulation allowed us to determine whether the slopes of the *z*-ROC functions were consistently lower for low-frequency words as a function of strength. In Glanzer and Adams's (1990) previous examinations of word frequency effects, there was a possible problem: In general, high-frequency words have lower *d'* values than low-frequency words, and in Glanzer and Adams's study, some subjects had low *d'* values. The combination of these two factors could have tended to make the slope for high-frequency words higher than the slope for low-frequency

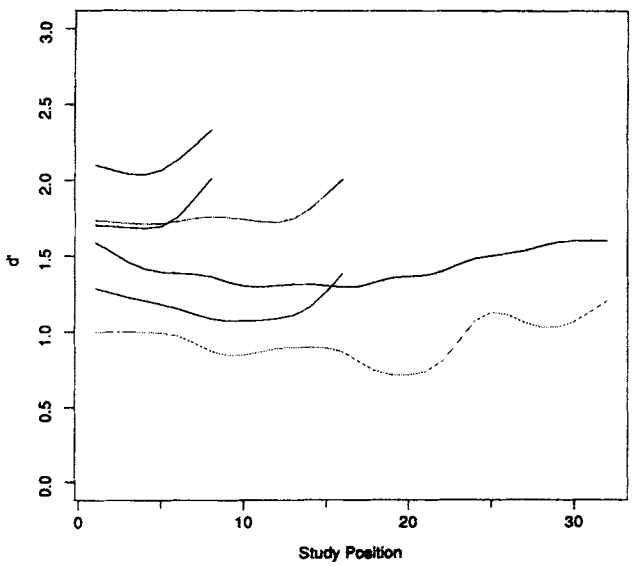

Figure 5. Smoothed serial position curves for three values of list length (the shorter curves represent shorter lists) and two values of strength (the lower of the two curves of equal length is the weaker list).

words because, as d' nears 0, the slope must approach 1 (see Experiments 1 and 2). Experiments 4 and 5 were designed to ensure that d' values were above 0.5 for all conditions.

Method

Materials. Two pools of words were formed from the Kucera and Francis (1967) word frequency lists. Words in the low-frequency pool had frequencies of either 4 or 5, and words in the high-frequency pool had frequencies between 78 and 10,601. The words varied from 4 to 10 letters in length. Words derived from other common words by adding suffixes (e.g., -ing, -ed, or -tion) were eliminated. In addition, no plurals or proper names were included, nor were any words that were deemed especially memorable or idiosyncratic in relation to the rest of the words. This resulted in a high-frequency pool of 815 words and a low-frequency pool of 871 words.

Subjects. There were 28 subjects from the Northwestern University introductory psychology class who received credit in the class for participation. Each subject participated in one session, for which there were 17 study-test lists, with the first list as a practice list.

Procedure. Study lists were composed of pairs of words to minimize the possibility of rehearsal trading strategies (see Ratcliff et al., 1990). In a pure list, each of 16 pairs was presented for the same amount of time, 2 s for weak or 5 s for strong items. In a mixed list, sequential blocks of pairs in the study list had different study times: the first 2 pairs at 2 s, the next 6 pairs at 5 s, the next 6 pairs at 2 s, and the last 2 pairs at 5 s, or the reverse ordering of presentation times. For both pure and mixed lists, within each middle block of 6 pairs, 3 pairs for which both words were high frequency and 3 pairs for which both words were low frequency were placed in random positions. The first and last 2 pairs in a list were buffer items, and one word of each buffer pair was high frequency and one low frequency. Subjects were instructed to learn the pairs for later cued-recall tests. In the 16 lists for a session, there were four of each type: pure weak, pure strong, and the two kinds of mixed lists.

There were 64 test items for each study list, with equal numbers of old and new test items in random order. Responses were recorded on the same 6-point scale used in the earlier experiments. After each response, there was a 250-ms blank interval followed by the next test item. For two randomly chosen 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 practice study-test list included a cued-recall test.

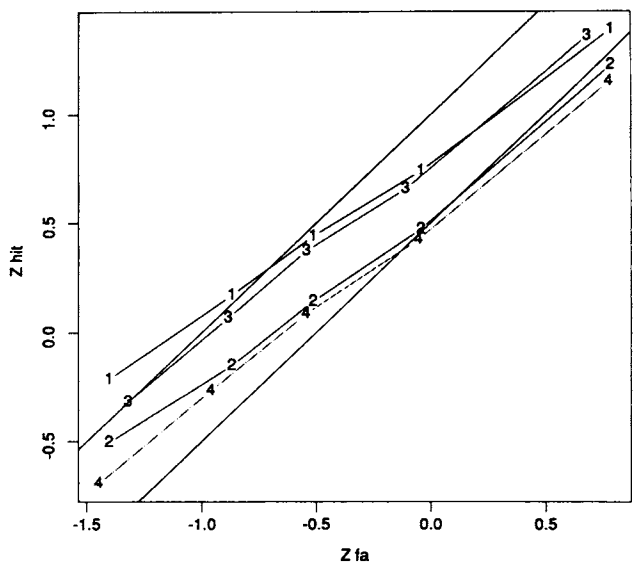

Figure 6. Z-transformed receiver-operating characteristic curves for high-frequency words for the 2-s and 5-s groups in Experiment 4. Curve 1 = mixed strong condition, Curve 2 = mixed weak condition, Curve 3 = pure strong condition, and Curve 4 = pure weak condition. The diagonal straight lines are for comparison and have a slope of 1. fa = false alarm.

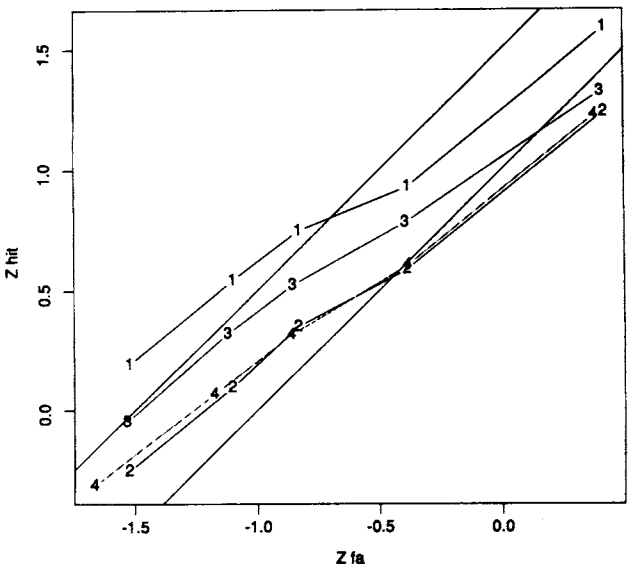

Figure 7. Z-transformed receiver-operating characteristic curves for low-frequency words for the 2-s and 5-s groups in Experiment 4. Curve 1 = mixed strong condition, Curve 2 = mixed weak condition, Curve 3 = pure strong condition, and Curve 4 = pure weak condition. The diagonal straight lines are for comparison and have a slope of 1. fa = false alarm.

Results and Discussion

Data analyses excluded responses with reaction times less than 250 ms and greater than 5,000 ms. Figures 6 and 7 show z-ROC curves for high- and low-frequency words, respectively. The estimated slopes and intercepts and the standard deviations in the estimates obtained from EPCROC by using the confidence judgment data pooled over subjects are shown in the first eight lines of Table 4. (Note that linear regression on the averages of z scores for individual subjects produced essentially the same results.) Figures 6 and 7 show parallel z-ROC functions that do not differ systematically from linearity. The slopes are in the 0.7 to 0.8 range, all significantly different from 1. The frequency manipulation significantly affected intercept, $F(4, 24) = 154.1$, and marginally affected slope, $F(4, 24) = 1.8$, $p = .16$. The study-time (strength) manipulation significantly affected intercept, $F(4, 24) = 76.8$, and marginally affected slope, $F(4, 24) = 2.0$, $p = .13$. The difference in slopes between weak and strong items was 0.021 (weak minus strong), and the difference in slopes between low-frequency and high-frequency items was 0.065 (high fre-

Table 4
Slopes and Intercepts for Word Frequency and Strength in Experiment 4

Condition	Slope		Intercept	
	M	*SD*	*M*	*SD*
HF				
Mixed strong	0.716	0.029	0.950	0.041
Mixed weak	0.765	0.028	0.631	0.038
Pure strong	0.825	0.037	0.879	0.044
Pure weak	0.801	0.033	0.566	0.043
LF				
Mixed strong	0.703	0.035	1.550	0.046
Mixed weak	0.734	0.030	1.072	0.041
Pure strong	0.693	0.035	1.291	0.047
Pure weak	0.719	0.034	1.091	0.046
HF fa vs. LF fa				
Mixed	0.923	0.028	−0.345	0.027
Pure strong	0.997	0.038	−0.287	0.038
Pure weak	0.929	0.040	−0.357	0.039

Note. High-frequency (HF) hits are scaled against high-frequency false alarms (fa) and low-frequency (LF) hits are scaled against low-frequency false alarms.

quency minus low frequency). The next experiment, Experiment 5, shows that the difference in slope as a function of strength does not replicate, so the marginally significant effect is probably due to the one exceptionally low data point in this experiment, the mixed strong, high-frequency condition (which had a slope of 0.716).

The results of Experiment 4 show a strong mirror effect (see Figures 6 and 7): The false-alarm rates for low-frequency words were lower than the false-alarm rates for high-frequency words, and the hit rates for low-frequency words were higher than the hit rates for high-frequency words (this pattern held for almost the whole range of confidence judgments). The mirror effect, combined with the low-frequency and high-frequency difference in slopes (0.065), replicates the results of Glanzer and Adams (1990) and Glanzer et al. (1991). With several variables in addition to frequency, they showed that, in general, item types with higher d' values had higher hit rates and lower false-alarm rates and also had lower ROC slopes than item types with lower d' values. Implications of our results for the Glanzer and Adams model are taken up in the General Discussion.

The analyses just described compared high- and low-frequency slopes by comparing high-frequency hits with high-frequency false alarms and low-frequency hits with low-frequency false alarms to produce *z*-ROC slopes for each. The slope represents the ratio of the standard deviations of the new-item familiarity distribution to the old-item familiarity distribution (assuming normal distributions). So the analyses just described reflect new compared with old high-frequency distributions, and new compared with old low-frequency distributions. It is also of interest to compare low-frequency new distributions to high-frequency new distributions (see Glanzer & Adams, 1990), which for normal distributions would provide the ratio of the standard deviations of high-frequency new-item distributions to low-frequency new-item distributions. Results for these comparisons are shown in the last three rows of Table 4. The results show that the slope is less than 1, indicating a larger standard deviation for the low-frequency new-item distribution, and that the intercept is less than zero. This replicates the results presented by Glanzer and Adams (1990), in which the slopes for various comparisons are ordered so that the more extreme the performance, the lower the slope of the *z*-ROC.

The larger standard deviation for the low-frequency distribution is plausibly explained by assuming that the familiarity of low-frequency words is more variable for a given subject than the familiarity of high-frequency words. A low-frequency word like *muse* might be unfamiliar to one subject but quite familiar to another, whereas most high-frequency words are uniformly familiar across subjects. Thus, what we labeled a low-frequency word on the basis of the Kucera and Francis (1967) statistics might actually be high frequency for one subject and very low frequency for another subject.

The ratios of ratios of d' values (the intercepts of the *z*-ROC functions) provide a measure of the list-strength effect. Calculating the ratios of ratios from the intercepts of the *z*-ROC curves produced a value of 0.969 for high-frequency words and 1.219 for low-frequency words. Computing the ratios from the intercepts divided by the slopes (to give d' values based on σ_n instead of σ_s in the denominator, assuming normal distributions) gave values of 1.066 for high-frequency words and 1.227 for low-frequency words. The low-frequency words appeared to show a slight list-strength effect, but it was not significant, $F(4, 24) = 0.86$, and the high-frequency words showed no list-strength effect.

Experiment 5

Word Frequency With Multisession Single-Subject Data

Because the joint behavior of strength and word frequency is important in testing attention likelihood theory (Glanzer & Adams, 1990), we decided to repeat Experiment 4 by collecting

Table 5
Slopes and Intercepts for Word Frequency and Strength in Experiment 5

Condition	Slope		Intercept	
	M	*SD*	*M*	*SD*
HF				
Mixed strong	0.721	0.021	1.468	0.025
Mixed weak	0.708	0.016	1.028	0.023
Pure strong	0.682	0.020	1.444	0.027
Pure weak	0.710	0.018	1.091	0.025
LF				
Mixed strong	0.659	0.024	2.264	0.031
Mixed weak	0.612	0.017	1.571	0.027
Pure strong	0.582	0.023	2.225	0.034
Pure weak	0.633	0.021	1.711	0.029
HF fa vs. LF fa				
Mixed	0.866	0.020	−0.501	0.031
Pure strong	0.878	0.017	−0.516	0.025
Pure weak	0.875	0.018	−0.515	0.027

Note. High-frequency (HF) hits are scaled against high-frequency false alarms (fa) and low-frequency (LF) hits are scaled against low-frequency false alarms. These values are averaged over the data of individual subjects, and the average standard deviation is obtained by averaging the standard deviation for each subject and dividing by the square root of the number of subjects (to give the standard error of the mean).

enough data to allow the performances of individual subjects to be modeled. We collected data from 11 subjects who each had from 7 to 11 sessions. This also enabled us to examine subject differences; discussion of individual subject differences is taken up in the General Discussion.

Method

The method was the same as in Experiment 4 with one change: The presentation times per pair were 1.5 s for weak pairs (to produce larger weak–strong performance differences) and 5 s for strong pairs. The 11 subjects provided a total of 97 sessions, after one practice session per subject was eliminated.

Results and Discussion

The data were analyzed as in Experiment 4. The confidence judgment data from individual subjects were fitted by EP-CROC and then the slopes and intercepts for the individual subjects were averaged to provide the results displayed in Table 5. The *z*-ROC curves for each individual subject are presented in Figures 8 and 9 and show mainly linear functions but with large individual differences.

Analyses from the general linear model showed that word frequency significantly affected the intercepts of the *z*-ROC curves, $F(4, 24) = 54.1$, and marginally affected the slopes, $F(4, 24) = 2.7, p = .06$. The effect of word frequency on slope was marginally significant in both Experiments 4 and 5, and combining the data for the two experiments, the effect reached significance. In Experiment 5, strength affected the intercepts, $F(4, 24) = 30.8$, but not the slopes, $F(4, 24) = 0.7$ (all significant ps $< .05$). The difference in slopes as a function of strength was 0.002 and as a function of word frequency (high frequency minus low frequency) was 0.109. The mirror effect

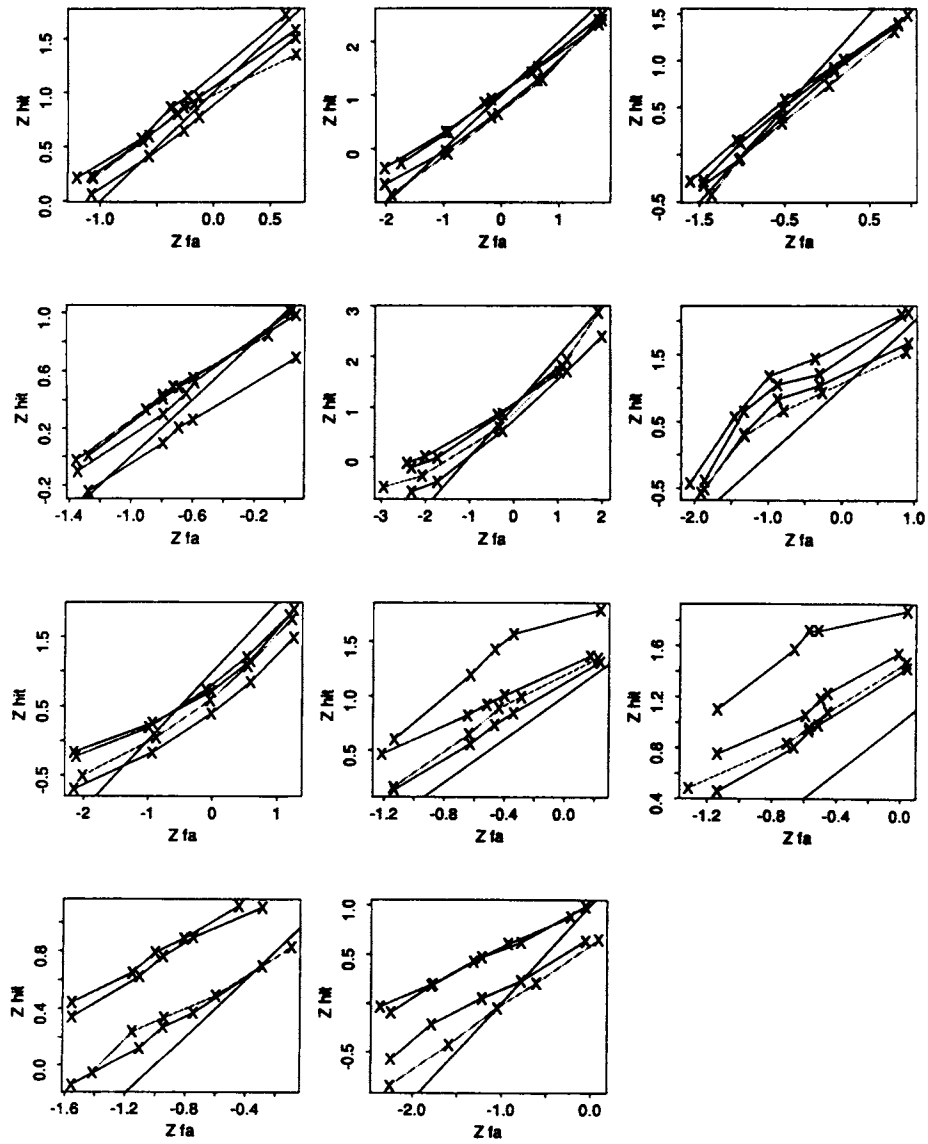

Figure 8. *Z*-transformed receiver-operating characteristic curves for high-frequency words for individual subjects in Experiment 5. The four curves represent mixed strong, mixed weak, pure strong, and pure weak conditions. The diagonal straight lines are for comparison and have a slope of 1. fa = false alarm.

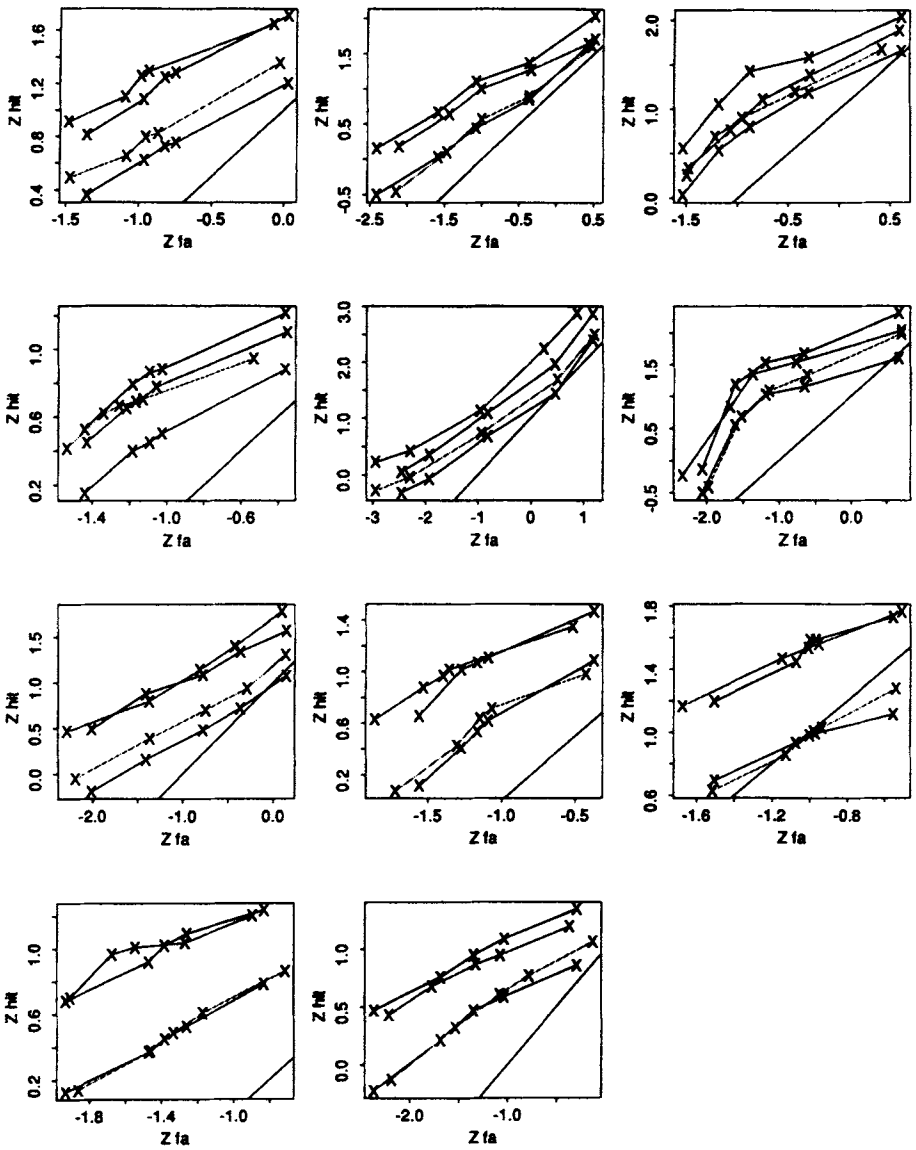

Figure 9. Z-transformed receiver-operating characteristic curves for low-frequency words for individual subjects in Experiment 5. The four curves represent mixed strong, mixed weak, pure strong, and pure weak conditions. The diagonal straight lines are for comparison and have a slope of 1. fa = false alarm.

was obtained for all of the subjects except one who had hit rates inconsistent with a mirror effect. To compare high- and low-frequency words, the slopes and intercepts for the *z*-transformed, high- and low-frequency false-alarm rates are presented in Table 5.

The list-strength effect was not significant ($F < 1$). The ratio of ratios for high-frequency words based on the intercept (d'_2) was 1.079 and based on the intercept divided by the slope (d'_1) was 1.018. For low-frequency words, the ratio of ratios based on the intercepts was 1.108 and based on the intercepts divided by the slopes was 0.946.

Overall, the results essentially replicate those of Experiment 4 except that the d' (intercept of the *z*-ROC) was much higher for Experiment 5. This could be a result of practice effects in conjunction with better motivated (paid) subjects.

It is interesting to note that in a session-by-session analysis, the frequency advantage (d') for low-frequency words in relation to high-frequency words was maintained across the 10 sessions; that is, after 10 presentations, 1 in each session, low-frequency words had not become equivalent in performance to high-frequency words.

The results of Experiments 4 and 5 replicate the results of Glanzer and Adams (1990) and extend them by showing that the slope of the *z*-ROC varies as a function of word frequency but not as a function of item strength. Thus, low-frequency words have roughly constant *z*-ROC slopes as a function of strength, and high-frequency words have a higher slope that is constant as a function of strength. The implications of these results are presented in the General Discussion.

Experiment 6

Category Manipulations

Experiments 1 through 5 manipulated the degree of match between a test item and memory by varying encoding time, which varies the strength of an individual trace, and by varying word frequency, which is a variable intrinsic to the item. Another way to manipulate the degree of match between a test item and memory is to vary the similarity of the test item to studied items. We did this by including in the study lists multiple words from the same semantic category. The aim of Experiment 6 was to determine whether varying the similarity of items within the list (compared with dissimilar new items) has an effect on the slope of the *z*-ROC. The global memory models predict the effect of similarity on the *z*-ROC to be the same as the effect of the study-time manipulation, so TODAM would predict a slope of 1 and SAM and MINERVA 2 would predict the slope decreasing as a function of the degree of match.

In this experiment, subjects were presented with mixed and pure study lists, and study time per item was varied. In each study list, there were two sets, with four pairs of words in each set and with all eight words of a set from the same category. At test, new words from the two studied categories were tested along with new words unrelated to the categories. To examine the slope of the *z*-ROC as familiarity increased, we used the responses to new unrelated test items as a single baseline against which to scale all of the other conditions (except for the word-pool items). The models all predict that the ordering of conditions in terms of strength or d' is unrelated new items, related new items, unrelated old items, and related old items. Thus we should see the slope of the *z*-ROC become constant as a function of strength and category condition once discriminability exceeds about 0.5 (see Experiments 1 and 2).

Method

Materials and subjects. The 48 categories of words used in the experiment were selected from the Battig and Montague (1969) category norms. Proper name categories, snakes, and names of a state, college, city, and building for religious services were excluded. The first 16 nonoverlapping single words from each category were used. An extra pool of words was selected from the same pool as in Experiment 1. Seventeen subjects from the Northwestern University introductory psychology pool participated for credit in a psychology course.

Procedure. Subjects were presented with 16 study-test lists (8 mixed lists and 8 pure lists). In a pure list, each of 16 pairs of words was presented for the same amount of time, 2 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 block of 2 pairs at 2 s, the next block of 6 pairs at 5 s, the next block of 6 pairs at 2 s, and the last block of 2 pairs at 5 s (the first and last blocks were buffers), or the reverse ordering of presentation times. The category structure of a study list was as follows: In each list, 4 pairs of words from each of two categories were presented. Words from these two categories are referred to as the *category* condition. Three of the pairs from a category were placed in one of the middle blocks (the first three positions or the last three), and 1 of the pairs was placed in a buffer block. Three of the pairs from the other category were placed in the other middle block in the same way, and the 4th pair was placed in the other buffer. There were also 4 pairs of words for which each word was selected from a different category (i.e., the words came from eight different categories). These were labeled the *random* condition. Three of these pairs were placed in one of the middle blocks, and 1 pair was placed in a buffer block. Finally, eight words from the extra word pool were used to make up the remaining pairs in the study list, and this was termed the *word-pool* condition. Across the 16 study lists of the experiment, the different types of pairs (category, random, and from the extra word pool) appeared equally often at each serial position in the middle blocks for each mixed–pure study-time condition (using a Latin square design). When a category was used for the category condition, no words from that category were used in any other study or test list in the experiment.

A test list was made up of the 32 studied items, plus 8 new items from one of the categories that was used to make up 4 study pairs, 8 new items from the other category that was used to make up 4 study pairs, 8 new items from categories for which no item appeared in the study list, and 8 new items from the extra word pool. These test items appeared in random order. Note that the 16 categories that were not used to make up word pairs for the category condition were reused across the 16 study-test lists, but individual items from those categories were not repeated. The experimental lists were preceded by 2 practice lists.

Subjects were instructed to study each pair of words for a cued-recall test. Three such tests were given, one each after the 2nd, 6th, and 10th lists. After each study list, subjects performed the same confidence judgment recognition memory test as in the preceding experiments.

Results

Z-ROC curves were constructed for responses from the category, random, and extra word-pool items as a function of mixed and pure lists and weak and strong study conditions, and fits are shown in Table 6. The category hits were scaled against the random false alarms, the random hits were scaled against random false alarms, and the false alarms for new test items from the studied categories were scaled against the false alarms for the random new items. (The word-pool hits were scaled against the word-pool false alarms because they are different words from the category members in the other conditions.) These comparisons represent two manipulations of degree of familiarity: study time and whether a test item matches other studied items from the same category. The *z*-ROC functions are shown in Figure 10.

For old-test items, the statistical analyses were carried out only on the category responses and the random responses as these were the main focus of the experiment. Strength (study time) had a significant effect on intercept, $F(4, 24) = 9.7$, and no effect on slope, $F(4, 24) = 0.54$. The pure–mixed list variable had no significant effect. Item type (category vs. random) had a significant effect on intercept, $F(4, 24) = 9.7$, and a marginal effect on slope, $F(4, 24) = 2.6, p = .059$. There was a significant list-strength effect, $F(2, 24) = 3.8$, but we assume this is spurious because there was no list-strength effect when the computation was based on the slope divided by the intercept (all significant $ps < .05$).

The ratio of ratios for the list-strength effect was computed for three comparisons. For the category items, the ratio of ratios based on the intercept (d'_2) was 1.030 and based on the intercept divided by the slope (d'_1) was 1.255. For the random items, the ratio of ratios based on the intercept (d'_2) was 1.270 and based on the intercept divided by the slope (d'_1) was 0.921.

Table 6
Slopes and Intercepts for Experiment 6: Match to Category Items and Strength

Condition	Slope		Intercept	
	M	*SD*	*M*	*SD*
Category				
Mixed strong	0.529	0.059	2.507	0.095
Mixed weak	0.656	0.064	2.186	0.085
Pure strong	0.615	0.066	2.424	0.092
Pure weak	0.625	0.062	2.167	0.085
Random				
Mixed strong	0.576	0.079	2.568	0.125
Mixed weak	0.416	0.048	1.676	0.104
Pure strong	0.520	0.067	2.246	0.115
Pure weak	0.518	0.060	1.873	0.105
Word pool				
Mixed strong	0.495	0.058	1.962	0.107
Mixed weak	0.565	0.058	1.600	0.095
Pure strong	0.477	0.055	1.944	0.107
Pure weak	0.604	0.061	1.521	0.093
Category false alarms				
Mixed strong	0.709	0.053	0.233	0.062
Mixed weak	0.777	0.054	0.248	0.061
Pure strong	0.702	0.054	0.176	0.064
Pure weak	0.694	0.051	0.325	0.061

Note. Category items are items studied along with other items from the same category. Random items are items from the categories but there is no other list item from that category. Category and random hits are scaled against random false alarms. Word-pool hits are scaled against word-pool false alarms. Category false alarms are scaled against random false alarms.

For the word-pool items, the ratio of ratios based on the intercept (d'_2) was 0.959 and based on the intercept divided by the slope (d'_1) was 0.864. These data were somewhat noisier than the data from the other experiments, but there were no systematic trends in the ratios of ratios.

The category false alarms scaled against the random false alarms present a minor puzzle. The EPCROC fits shown in Table 6 deviated systematically from the linear regression slopes (which were 0.84, 0.88, 0.78, and 0.83 as opposed to the EPCROC slopes 0.70, 0.69, 0.71, and 0.78; see Table 6). The reason for this is differential weighting as a function of number of observations in the two methods. For the false alarm–false alarm comparison, there were few high-confidence old responses, which means that EPCROC will not weight this category much compared with the high-confidence new category that had over half of the responses. The linear regression analysis on the other hand weights all categories equally. For the old–new comparisons, this is not a problem because when there are small numbers of observations in one category for false alarms, say, there are large numbers in that category for hits leading to roughly equal weighting for both methods. The values of the intercept (d') were close to zero, but the slope was not close to 1 (cf. Experiments 1 and 2 in which low d' resulted in slopes near 1).

The category false alarm versus the random false alarm comparison provides an important result for modeling. The finding that the slopes were less than 1 suggests that unlike the case of repetitions of a single item, presentations of related items increase the variance of the familiarity distributions (assuming normal distributions): For the category new items for which there were eight other members of the category, the variability in familiarity values was greater than for category new items for which there was no other member of the category.

For the word-pool items, slope decreased as strength of the old items increased. This was the only effect of strength on slope in all of the experiments, and we assume it is spurious. Only one of the effects was significant (the pure weak to pure strong comparison). The average slope was much lower in Experiment 6 than in the earlier experiments. However, the range of values certainly fall within the range of individual differences (see the General Discussion section for a review of individual differences), and so the difference was probably the result of the particular group of subjects in this experiment.

These results show that manipulating strength by using study time and including similar items in the study list had remarkably similar effects. The slope of the *z*-ROC was constant as a function of strength but was marginally affected by the category–random manipulation (whether there were other similar items in the study list). A similar result was found for the category versus random false alarms, a slope different from 1. This means that (under the assumption of normal distributions) the effect of other items from the same category on the list as the test item increased the standard deviation in the match value for both old items and new items, which is a prediction of the global memory models. However, as noted by Ratcliff et al. (1992), the models did not predict the behavior of the slope of the *z*-ROC as a function of strength manipulated by study time.

Nonlinear z-ROC Functions

Inspection of Figures 3, 8, and 9 shows that some subjects have systematically curved *z*-ROC functions, several convex and a few concave. There are two possible reasons for this. The first is that the underlying distributions are not normal and do not mimic normal distributions. Investigation of this possibility would require the consideration of alternative distributions in the context of some larger model. The second possibility is that the curving is artifactual, caused, for example, by some proportion of trials for which responding was random or systematically different from other trials (e.g., changing decision confidence cutoffs systematically).

Contaminations like these raise complicated issues that will require considerable further research. The aim here is to alert readers to the possibility of mixtures and contaminations and to point out what these effects might look like. It is easy to demonstrate that contamination of the signal or noise distributions or both by a small proportion of data from trials on which decision criteria shift leads to convex *z*-ROC functions. Two examples of such contamination were examined with simulations. Both examples assumed normal distributions of signal (old-item familiarity) and noise (new-item familiarity; with the noise distribution $M = 0$ and $SD = 0.8$, and the signal distribution $M = 2.0$ and $SD = 1.0$). These distributions produce a *z*-ROC line $z_h = 0.8z_{fa} + 2.0$ (i.e., slope 0.8 and intercept 2.0). Confidence judgment criterion values were set at −0.9, 0.6, 1.3, 1.8, and 2.9, and from the cumulative distributions, the proportion of counts in each confidence category was determined for each of the ranges above the highest criterion, below the lowest criterion, and between the criteria. To mimic what

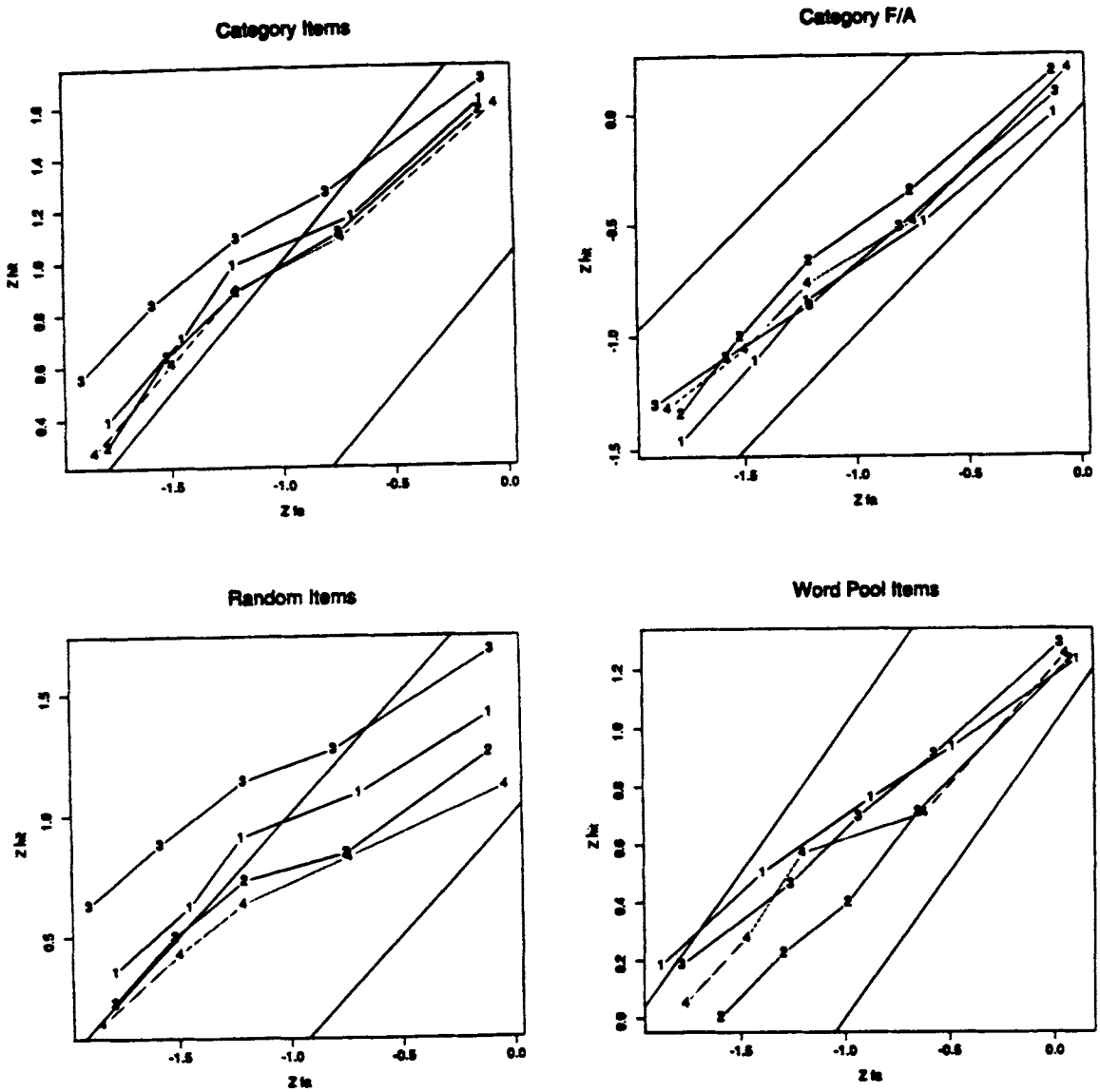

Figure 10. *Z*-transformed receiver-operating characteristic curves for Experiment 6 for the following comparisons: category hits (old items from a studied category) versus random (items from other nonstudied categories) false alarms (fa), random hits (studied items from a category with no other category members in the list) against random false alarms, category false alarms (words from a studied category but not studied in the list) versus random false alarms, and word-pool hits versus word-pool false alarms. Curve 1 = pure strong condition, Curve 2 = pure weak condition, Curve 3 = mixed strong condition, and Curve 4 = mixed weak condition. The diagonal straight lines are for comparison and have a slope of 1.

would happen when a subject varied some of these criteria for some proportion of test items, contaminated distributions were obtained by changing the two extreme criteria, moving −0.9 to 0 and 2.9 to 2.0. Then 95% of the proportion of counts for the uncontaminated distribution was added to 5% of the counts for the contaminated distribution, leading to a simulated distribution of counts in the confidence categories. Then the counts were transformed back to a distribution function. For the *z*-ROC curve obtained from these contaminated distributions, the slope was affected little in relation to the uncontaminated distribution, 0.816, but the intercept was reduced by about 25% to 1.576. The *z*-ROC and the linear fit are shown in the top panel of Figure 11 and show an almost linear function.

The second simulation used the same distributions as the first and the same criteria, but contamination was due to spurious data introduced into all of the confidence judgment categories. The spurious data came from a uniform distribution of counts in the confidence judgment categories, and 5% of the counts from this contaminated distribution was added to 95% of the counts for an uncontaminated distribution. The slope of the *z*-ROC curve was 0.847 and the intercept was

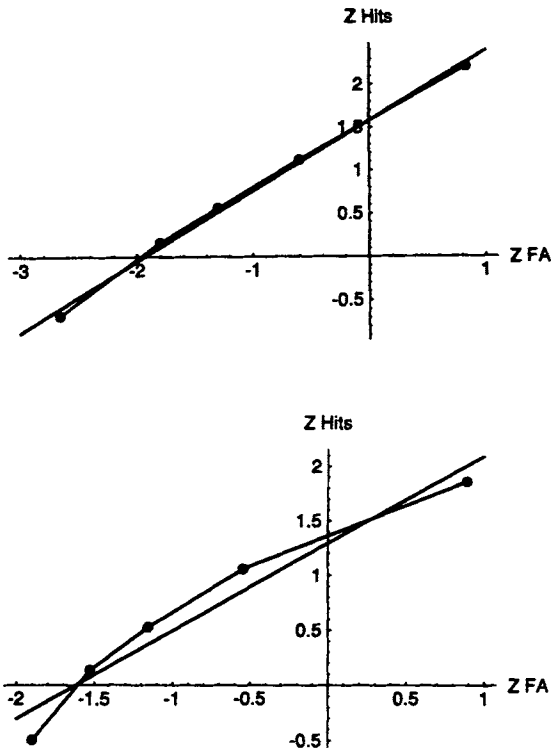

Figure 11. Plots of *z*-transformed receiver-operating characteristic functions: for top panel, altering the two extreme confidence judgment criteria for 5% of the observations; for bottom panel, adding in 5% uniform noise (i.e., equal numbers of observations in each confidence category). The diagonal straight lines are for comparison and have a slope of 1. FA = false alarm.

1.289, so that the intercept was changed a lot, but again the slope was affected by less than 10%. In contrast to the first simulation, this *z*-ROC function is convex and looks like the convex data shown for some of the subjects in Figures 3, 8, and 9.

These examples are quite simple, but they make the important point that variability in criterion settings and noise in the data can have large effects, including producing nonlinear *z*-ROC functions (note the examples do not deal with convexity, which might have a similar explanation). It might be thought that a nonlinear *z*-ROC is a signature of nonnormal distributions, but Figure 11 shows that it might also be the result of noise added to normal distributions, and that the first investigation of nonlinear *z*-ROC functions should be to see whether noise could be contaminating the data. Although this is only a nonsystematic initial attempt to look at the problems of averaging and contamination by random data in the *z*-ROC analysis, it is clear that nonlinearity in the shape of the *z*-ROC does not necessarily mean nonnormal distributions (or distributions that mimic the normal).

General Discussion

To summarize the empirical results, we begin by listing the results that replicate the findings of Ratcliff et al. (1992). The most basic result is that the *z*-ROC curves appear to be linear, consistent with the assumption of the global memory models that the distributions of familiarity values for old and new test items are normal. Given linearity, the *z*-ROC curves can be used to test predictions of the global models.

The second major result contradicts predictions of the global memory models: the slopes of the *z*-ROC curves average about 0.8, independent of the strength of encoding of studied test items. For the global memory models, the slope is the ratio of the standard deviation of new test item familiarity to the standard deviation of old test item familiarity, and the models predict that this ratio should be about 1 (Murdock, 1982) or that it should decrease as a function of strength (Gillund & Shiffrin, 1984; Hintzman, 1986). Ratcliff et al. (1992) considered the possibility that the constant slope resulted from an averaging artifact: If studied items of different strengths are averaged together (e.g., from different study and test positions), the distribution of familiarity for old items becomes wider and the slope of the *z*-ROC must decrease below 1. They rejected this possibility empirically by performing analyses in which the data were broken down by study and test position to show that there were no significant differences for weaker items (early study and late test) compared with stronger items (late study and early test). Theoretically, this artifact can be ruled out because the difference in strength among studied items required to produce a 0.8 slope was too large to be plausible. For the experiments reported here, we again considered the averaging artifact, and again, analyses based on study and test position showed no systematic differences.

The third result is that the standard deviation for the new-item familiarity value is about the same whether the new items are tested following a pure weak or pure strong encoding list or a mixed list. There is no significant list-strength effect (the ratios of ratios, Rr, were always about 1), in contradiction to the models' predictions (outlined in the introduction).

Five out of the six experiments contained a list-strength manipulation, and Table 7 shows a summary of the ratios of ratios for each experiment and condition. The overall result is that the average ratio of ratios was 1.03 when calculated from the intercept (d'_2) of the *z*-ROC and 0.98 when calculated from the intercept divided by the slope (d'_1). This result extends yet further the generality of the finding of no list-strength effect (Murnane & Shiffrin, 1991; Ratcliff et al., 1990; Yonelinas et al., 1992).

The experiments reported here add to and generalize previous results in several ways. First, the rapid presentation rates for study items used in Experiments 1 and 2 show how the *z*-ROC function changes as d' approaches zero. With rapid presentation rates, subjects cannot easily redistribute rehearsal across study items; they cannot use study time for slow items to rehearse fast items (see Ratcliff & McKoon, 1991; also Yonelinas et al., 1992). The results show approximately constant slopes for d' values between 0.5 and 2.5. Below 0.5, the slope quickly approaches 1 (as it must) as d' approaches 0.

In the experiments reported by Ratcliff et al. (1992), the familiarity of a test item matched against memory was manipulated only by strength of encoding; study items were presented for either a longer time or a shorter time. The slope of the *z*-ROC did not change as a function of strength, and the failure to find a list-strength effect held constant across strength values. In the experiments reported here, we added two more

manipulations of familiarity—list length and similarity to other items in the study list—and combined them with study time. With a longer list length, the familiarity of a studied item decreased, but the slope of the *z*-ROC remained constant. With the familiarity of new test items increased by taking them from the same semantic categories as studied items, the slope of the *z*-ROC still held constant. And the list-strength effect predicted by the global memory models was not obtained with any of these manipulations.

The pattern of results that is translated to familiarity distributions, assuming normal distributions, is shown in Figure 12 (copied from Ratcliff et al., 1992). The constraints on the models provided by the data are shown in the figure. The standard deviation of the familiarity values for new test items is the same for mixed lists, pure strong lists, and pure weak lists, for which strength can be manipulated by study time, number of repetitions, list length, or similarity. The standard deviation of familiarity values for old test items is 1.25 times greater, as dictated by the 0.8 slope of the *z*-ROC curves, and it is constant as a function of strength. This figure represents the simplest description of the data under the assumption of normal distributions. It could be that theoretical familiarity distributions might be only one component of several processes determining the shapes of the distributions illustrated here. But a more complex model of this kind has yet to be developed.

The critical issue is whether the global memory models can accommodate the patterns shown in Figure 12. Since Ratcliff et al. (1992), there have been two suggestions about how this might be done. One is the differentiation version of SAM discussed by Shiffrin et al. (1990). In this model, as study time increases, the context strength between the studied item and context increases, whereas the residual item strength is assumed to increase up to some short encoding time and then decrease counteracting the increase in context strength. This makes the overall familiarity of a new test item constant as a function of strength of old test items, and constant familiarity gives constant standard deviation and the correct prediction that there will be no list-strength effect. The initial increase in residual item strength leads to a predicted list-strength effect at short presentation durations, but we did not find this in Experiments 1 and 2, although the rates of presentation we used may have missed the critical region. But, as discussed in Ratcliff et al. (1992), the model still predicts that the slope of the *z*-ROC function will decrease as a function of strength or *d'*.

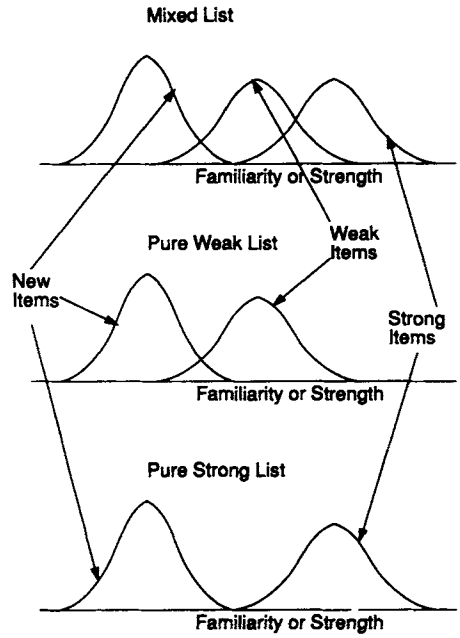

Figure 12. An illustration of the behavior of the strength distributions as a function of mixed versus pure list and as a function of strength differences. The new-item strength standard deviation remains constant as a function of list type; the standard deviation of the old-item distribution remains constant as a function of strength and is larger than the standard deviation for new items. Reprinted from "Testing Global Memory Models Using ROC Curves" by R. Ratcliff, C.-F. Sheu, and S. D. Gronlund, 1992, *Psychological Review, 99,* p. 530. Copyright 1992 by the American Psychological Association.

Table 7
Ratio of Ratios (Rr) for Experiments 1, 2, 4, 5, and 6

Experiment and condition	Rr based on intercept of the *z*-ROC	Rr based on the intercept divided by the slope of the *z*-ROC
Exp. 1 Study Time	0.70	0.69
Exp. 2 Study Time	0.92	0.84
Exp. 4, HF	0.97	1.07
Exp. 4, LF	1.22	1.23
Exp. 5, HF	1.08	1.02
Exp. 5, LF	1.11	0.95
Exp. 6, Category	1.03	1.26
Exp. 6, Random	1.27	0.92
Exp. 6, Word pool	0.96	0.86
Average	1.03	0.98

Note. Rr refers to the ratio of mixed strong to mixed weak divided by the ratio of pure strong to pure weak. *z*-ROC = *z*-transform of the receiver-operating characteristic; HF = high frequency; LF = low frequency.

Murdock and Kahana (1993) proposed a new variant of TODAM with a continuous memory assumption. According to this assumption, the items in memory against which a test item is matched are not only the items from the immediately preceding study list but also all of the items from all earlier study lists. With this assumption, TODAM predicts correctly that there will be no effect of list strength. Shiffrin, Ratcliff, Murnane, and Nobel (1993) criticized this model on several grounds, including the problem that it has not been tested against recognition memory phenomena other than the list-strength effect and that it cannot (in any obvious way) account for list discrimination effects (Anderson & Bower, 1972) or the effects of repeating new test items (Ratcliff & Hockley, 1980). In the discussion of Experiment 3, we pointed out that it incorrectly predicts equal performance across list lengths for test items with equivalent study-test lags. In addition, the model cannot account for the constant 0.8 value of the *z*-ROC slope; it predicts a slope near 1.

Besides accounting for the pattern of data leading to the theoretical distributions (based on the assumption of normal distributions) shown in Figure 12, new versions of the global memory models will also have to allow for individual differ-

Table 8
Slopes and Intercepts per Subject Averaged Across Conditions in Experiment 3 (List Length) and Experiment 5 (Word Frequency)

Subject	Slope		Intercept	
	M	*SD*	*M*	*SD*
Experiment 3, List Length				
1	0.847	0.058	2.415	0.060
2	0.687	0.036	1.616	0.061
3	0.763	0.029	0.982	0.041
4	0.897	0.079	1.766	0.060
5	1.041	0.049	1.425	0.042
6	0.885	0.047	1.467	0.045
7	0.787	0.034	1.388	0.046
Experiment 5, Word Frequency				
1	0.742	0.025	1.550	0.029
2	0.642	0.022	1.336	0.028
3	0.560	0.015	1.348	0.029
4	0.533	0.021	1.672	0.033
5	0.699	0.017	1.406	0.027
6	0.630	0.016	1.587	0.030
7	0.773	0.033	1.692	0.035
8	0.433	0.017	1.441	0.038
9	0.791	0.026	1.409	0.032
10	0.833	0.029	2.038	0.035
11	0.694	0.039	2.122	0.041

Note. Slopes and intercepts are averaged over strength and frequency manipulations for word frequency and over list length and strength for list length. The standard deviations are for the straight line fit to the data averaged over all the mixed–pure and strong–weak conditions for that subject.

ences in the value of the slope of the *z*-ROC curve. In Experiments 3 and 5, individual subjects were tested for large numbers of sessions, and Table 8 shows slopes and intercepts averaged over high- and low-frequency words and all strength values for Experiment 5, and averaged over all list lengths and strengths for Experiment 3. The slopes vary from a low of 0.433 to a high of 1.041 (with small standard errors in the slopes), a range that corresponds to that obtained from Murdock and Dufty's (1972) individual subjects (reported in Ratcliff et al., 1992). Thus the models must be capable of producing slopes of the *z*-ROCs that vary for individuals between 0.5 and 1.0.

Another critical problem is presented by the mirror effect and the effects of word frequency on recognition memory performance. The global memory models could explain the change in slope of the *z*-ROC curves as a function of word frequency. The assumption would be that the standard deviation of low-frequency words is greater than the standard deviation of high-frequency words. This assumption seems intuitively reasonable; high-frequency words are probably highly familiar to all subjects, but some low-frequency words are unfamiliar to some subjects. However, the models do have the problem noted earlier that the old and new distributions of familiarity are nearer to the criterion for high-frequency compared with low-frequency words, and none of the models have satisfactorily accounted for this result. The mirror effect has been addressed by a different kind of model from the global memory models, Glanzer and Adams's (1990) attention *likelihood model* (see also Glanzer, Adams, Iverson & Kim, 1993).

Glanzer and Adams's (1990) Attention Likelihood Theory

The attention likelihood model was developed primarily to account for the mirror effect in recognition memory. The model assumes that each item in memory is represented by a list of N features, where these features can be marked or not marked. Before an item is encoded, some proportion of its features (p[new]) is already marked. At encoding, some additional (typically) small proportion is marked, for a total proportion marked of p(old). The proportion marked at encoding is a function of the item's type: the more attention evoking the item (e.g., a low-frequency word in contrast to a high-frequency word), the greater the proportion of marked features. At retrieval, the subject examines some (again, typically small) number n of features for a test item, and then decides whether the number of these that are marked (x) is likely to represent an old item or a new item. The number n sampled at retrieval is a function of an item's type just as at encoding. The decision rule compares a likelihood ratio computed for a test item to a criterion (or set of criteria if a confidence judgement procedure is used). The likelihood ratio is the probability of the observed number of marked features given the item is old divided by the probability of the observed number of marked features given the item is new, that is, the probability of x-marked features from a binomial distribution with total number of (observed) features n and probability parameter p(old) divided by the probability of x-marked features from a binomial with n features and probability parameter p(new). Therefore, to find the probability that a test item is old, the subject has to know the values of p(old), p(new), and n for that kind of test item. (Note that the total number of features N, does not enter this calculation except through p[old] and p[new].) To explain the mirror effect, higher values of p(old) and n are assumed for low-frequency words than for high-frequency words. But the attentional mechanism that gives these higher values is unspecified, and the model provides no insight into what features of the stimulus in memory give a low-frequency word extra attention in relation to a high-frequency word.

Glanzer and Adams (1990; see also Glanzer, et al., 1993) attempted to show that the attention likelihood model could account qualitatively for the behavior of the slopes of *z*-ROC curves as a function of word frequency and other materials variables (e.g., concreteness). To do this, Glanzer and Adams assumed that the slope of a *z*-ROC curve was the ratio of the standard deviations for the new- and old-item likelihood distributions. However, this assumption is incorrect because the relationship between slope and standard deviation ratio only holds for normal distributions (as in the global memory models), not for the likelihood distributions in the attention likelihood model. If the model is correctly fit to the data, that is, the parameter p(new) is set to produce standard deviation ratios in the 0.6 to 0.7 range obtained in empirical data, then the model predicts *z*-ROC slopes that are much larger than those in real data (e.g., standard deviation ratios of 0.6 to 0.7 give *z*-ROC slopes of around 0.9). To understand this, consider the distributions in Figure 13. To produce standard deviation ratios in the 0.6 to 0.7 range (Glanzer & Adams, 1990), the p(new) value must be in the range of 0.05 (for n around 60). But when p(new) is set to 0.05, then the new (left-hand) distribution has a truncated left tail (i.e., a high probability of

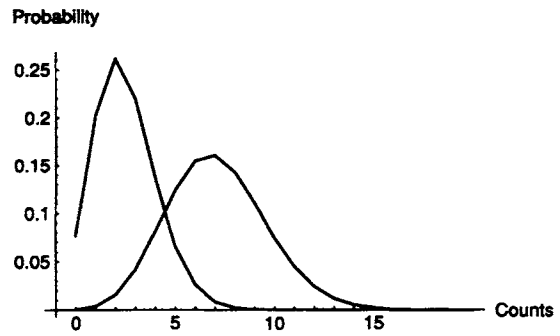

Figure 13. Binomial distribution for the Glanzer and Adams (1990) model. The parameters are p(new) = .05, p(old) = .107, and n = 60.

zero counts, or marked features), as shown in the figure (the probability of zero counts, the minimum, is at 0.07). Figure 13 also shows the distribution for p(old), set at a value of 0.1 (which would produce d' values in the range of the experimental data). To obtain the ROC functions from these distributions, cumulative proportions are obtained moving from the right-hand side of the figure, and these probabilities can then be transformed to z scores to give z-ROC functions. To understand the discrepancy between the ratio of standard deviations and the slope of the z-ROC, consider the result if the left-hand tail of the new distribution was not truncated in Figure 13. Then, the z-ROC would be about the same as for the truncated tail case, but the standard deviation for new-item distribution would be much larger. The conclusion is that for these nonnormal distributions, the slope of the z-ROC cannot be computed from the standard deviations; instead it must be computed directly from the ROC functions.

To predict slopes of the empirically obtained values from the ROC functions, the attention likelihood model would have to use extreme values of p(new) and p(old). To produce a slope of about 0.8, the values would have to be p(new) = 0.01 and p(old) = 0.05 with n = 50, the distributions shown in Figure 14. The problem with these distributions is that they are quite different from those assumed by Glanzer and Adams (1990). In particular, of the hundreds of features for an item, the decision mechanism would be provided with only zero, one, or two marked features for a new item (typically) against only zero to six marked features for an old item (typically), which is a very small sample on which to base decisions. Any ROC function with parameter values near these would have only three or four distinct points corresponding to nonzero counts in the new-item distribution in Figure 14, which is contrary to the data. Given these problems, it is difficult to know whether the Glanzer and Adams model could be reworked to produce empirically adequate fits to the z-ROC data as well as hit and false-alarm rates for the mirror effect.

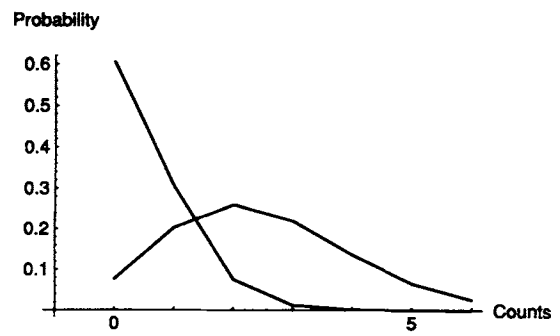

Figure 14. Binomial distribution for the Glanzer and Adams (1990) model. The parameters are p(new) = .01, p(old) = .05, and n = 50.

A key feature of the attention likelihood model is the transformation embodied in the scale of the decision axis, the transformation from an absolute strength criterion to a likelihood ratio criterion. This transformation does not, by itself, affect the shape or slope of the z-ROC curve; for example, for two overlapping distributions with a criterion set somewhere in the middle, stretching the scale (as likelihood theory does) leaves the proportion of each distribution above the criterion the same no matter how much the scale is stretched or shrunk on the right or left (note, however, that the ratios of standard deviations would be affected by such stretching). Thus the slopes and shapes of the z-ROC functions are independent of the particular decision rule adopted (likelihood or strength criterion) and instead are determined by the distributional assumptions.

Glanzer et al. (1993) made clear predictions about how the study-time variable is modeled in attention likelihood theory. A mixed-list design in which some of the items are strong and some weak is the best design for testing the model because the value of p(new) is fixed and common to weak and strong old items, in contrast to pure lists in which it might change as a function of list type. To examine predictions from attention likelihood theory, we generated z-ROC curves for parameter values p(new) = 0.01, n = 50, and p(old) = 0.02, 0.03, 0.05, and 0.07 (to represent four degrees of strength). The predicted intercept values for the z-ROC curves were 0.614, 1.051, 1.701, and 2.211, and the slope values were 0.925, 0.881, 0.825, and 0.788, so that the slope fell as the strength of the items increased, contrary to data. Thus, under the assumption that subjects cannot adjust criteria on the basis of strength in a mixed-list design, the attention likelihood model fails to account for the pattern of empirically obtained z-ROC curves in precisely the same way as the models of Gillund and Shiffrin (1984) and Hintzman (1986, 1988), as discussed by Ratcliff et al. (1992).

Another problem for the attention likelihood model is that it assumes that only the representation of the test item is accessed at test, not the representations of other items in memory. Experiment 6 shows that a test item does contact other items because category and random test items are from the same pool of category members and what differentiates the two classes is whether other items of the same category were presented with them in the study list. Hit rates for an old item and false-alarm rates for a new item were higher when other items from the same category were studied in relation to the cases in which there were no other members of the category in the study list.

The results reported in this article cause problems for attention likelihood theory. But attention likelihood theory is currently the most comprehensive mechanism available for dealing with the mirror effect, and it is hoped that the results presented here will provoke further development of this model.

Conclusion

The data presented in this article extend the experimental results of Ratcliff et al. (1992). The slope of the *z*-ROC function was affected only by type of materials (e.g., high- vs. low-frequency words) and not by strength manipulations, such as amount of encoding, list length, and similarity of other study and test items and that is counter to the predictions of the global memory models. The findings of individual differences among subjects both in slope and shape of the *z*-ROC provide additional problems for the models. The models must be able to predict individual differences in the slope from 0.5 to 1.0, and shape differences must be ruled out by appealing to averaging effects as discussed earlier or by predicting or assuming alternative distribution shapes. These data provide empirical findings to add to the database for developing and extending the global memory models. An important challenge both theoretically and empirically is to understand why few manipulations have an effect on the slope of the *z*-ROC functions (within the standard errors reported here) and what this means in the global memory models for the shapes and behaviors of the distributions of familiarity underlying recognition memory.

References

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

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication of the Connecticut Category Norms. *Journal of Experimental Psychology Monograph, 80,* 1–46.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis.* New York: Wiley.

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.

Glanzer, M., Adams, J. K., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 81–93.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567.

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

Gronlund, S. D., & Ratcliff, R. (1989). The time-course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15,* 846–858.

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.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods.* Boston: PWS-Kent.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

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

Loftus, G. R. (1974). Acquisition of information from rapidly presented verbal and nonverbal stimuli. *Memory and Cognition, 2,* 545–548.

McNicol, D. (1972). *A primer of signal detection theory.* London: Allen & 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., Jr., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology, 94,* 284–290.

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689–697.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855–874.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

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

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86,* 446–461.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list strength effect: 1. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178.

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: Nonmonotonic lag functions. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 555–573). Hillsdale, NJ: Erlbaum.

Ratcliff, R., & McKoon, G. (1991). Using ROC data and priming results to test global memory models. In S. Lewandowsky & 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., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190–214.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Shiffrin, R. M., Ratcliff, R., & Clark, S. (1990). The list strength effect: 2. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195.

Shiffrin, R. M., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1445–1449.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). A demonstration of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345–355.

Appendix

Counts per Confidence Category

Condition	Confidence category					
	− − −	− −	−	+	+ +	+ + +
Rapid presentation times, Experiment 1						
MS new	788	1585	1160	656	555	299
MS old	172	355	345	283	308	432
MW new	788	1585	1160	656	555	299
MW old	261	551	414	280	223	165
PS new	477	776	527	321	258	184
PS old	192	401	290	267	316	442
PW new	235	649	719	442	254	156
PW old	151	496	480	350	221	142
Rapid presentation times, Experiment 2						
MS new	466	1468	1399	597	496	345
MS old	69	285	314	267	323	537
MW new	466	1468	1399	597	496	345
MW old	148	452	451	254	235	256
PS new	519	683	601	224	211	151
PS old	125	263	256	227	293	631
PW new	167	629	848	329	254	159
PW old	85	399	556	268	259	223
List length, Experiment 3						
Subject 1						
W, LL 16, New	136	197	58	24	33	13
W, LL 16, Old	18	42	13	53	158	185
S, LL 16, New	404	42	0	0	5	15
S, LL 16, Old	19	3	0	0	29	413
W, LL 32, New	52	205	95	60	45	4
W, LL 32, Old	5	25	18	40	67	78
S, LL 32, New	274	117	10	5	15	10
S, LL 32, Old	5	4	0	2	34	197
W, LL 64, New	20	200	125	69	72	2
W, LL 64, Old	1	7	12	24	45	18
S, LL 64, New	236	147	26	10	18	10
S, LL 64, Old	2	2	2	3	35	90
Subject 2						
W, Ll 16, New	46	73	127	64	16	3
W, LL 16, Old	4	20	46	104	63	85
S, LL 16, New	75	73	105	54	14	5
S, LL 16, Old	3	5	19	60	31	207
W, LL 32, New	35	79	106	97	29	4
W, LL 32, Old	5	12	22	42	30	41
S, LL 32, New	60	76	106	68	12	4
S, LL 32, Old	2	7	8	32	12	86
W, LL 64, New	34	48	112	91	37	5
W, LL 64, Old	2	4	14	21	20	19
S, LL 64, New	50	84	86	85	22	5
S, LL 64, Old	1	4	7	15	6	51
Subject 3						
W, LL 16, New	107	135	102	46	49	24
W, LL 16, Old	35	54	48	61	105	164
S, LL 16, New	108	132	109	53	45	19
S, LL 16, Old	27	44	37	66	88	202
W, LL 32, New	91	122	95	63	55	24
W, LL 32, Old	22	32	29	37	42	79
S, LL 32, New	71	112	96	75	58	24
S, LL 32, Old	17	39	20	41	58	68
W, LL 64, New	73	119	105	75	68	36
W, LL 64, Old	14	15	9	19	31	24
S, LL 64, New	85	114	87	77	59	30
S, LL 64, Old	4	12	9	17	36	46
Subject 4						
W, LL 16, New	240	21	14	12	8	27
W, LL 16, Old	53	10	12	15	20	210
S, LL 16, New	290	11	3	3	3	15
S, LL 16, Old	36	4	5	8	10	258
W, LL 32, New	196	47	23	13	22	28
Subject 4 (*continued*)						
W, LL 32, Old	23	13	6	8	21	77
S, LL 32, New	255	20	11	8	9	22
S, LL 32, Old	14	4	1	11	14	95
W, LL 64, New	152	58	27	22	22	44
W, LL 64, Old	14	0	3	11	11	31
S, LL 64, New	184	39	24	18	17	44
S, LL 64, Old	5	0	5	5	6	56
Subject 5						
W, LL 16, New	163	131	67	30	36	38
W, LL 16, Old	19	23	12	51	88	269
S, LL 16, New	192	124	59	34	31	24
S, LL 16, Old	18	17	14	28	69	316
W, LL 32, New	122	112	42	46	82	49
W, LL 32, Old	13	21	18	25	56	108
S, LL 32, New	113	125	49	44	71	45
S, LL 32, Old	10	16	5	30	52	116
W, LL 64, New	99	115	65	48	91	59
W, LL 64, Old	6	10	5	16	34	38
S, LL 64, New	105	120	53	52	59	62
S, LL 64, Old	2	5	7	12	32	56
Subject 6						
W, LL 16, New	136	128	101	36	19	43
W, LL 16, Old	29	36	37	47	44	273
S, LL 16, New	153	130	114	20	12	38
S, LL 16, Old	10	25	21	38	28	341
W, LL 32, New	102	106	97	48	42	69
W, LL 32, Old	12	7	24	26	30	119
S, LL 32, New	118	117	105	32	22	57
S, LL 32, Old	13	7	6	20	24	149
W, LL 64, New	104	102	89	55	47	77
W, LL 64, Old	7	6	4	11	14	63
S, LL 64, New	89	98	90	42	28	90
S, LL 64, Old	5	6	8	12	16	73
Subject 7						
W, LL 16, New	153	177	54	27	40	15
W, LL 16, Old	24	45	29	43	120	199
S, LL 16, New	216	156	34	11	26	21
S, LL 16, Old	18	45	24	24	89	261
W, LL 32, New	89	188	72	29	45	10
W, LL 32, Old	19	44	11	23	93	53
S, LL 32, New	140	193	60	25	35	18
S, LL 32, Old	12	20	13	13	67	90
W, LL 64, New	81	186	74	33	80	21
W, LL 64, Old	14	21	7	11	33	20
S, LL 64, New	53	187	74	39	69	22
S, LL 64, Old	4	14	4	16	38	36
Word frequency, Experiment 4						
High frequency						
MS new	738	1026	611	378	381	273
MS old	105	187	127	130	194	526
MW new	738	1026	611	378	381	273
MW old	147	278	166	156	176	408
PS new	433	518	285	184	165	163
PS old	102	197	116	139	177	438
PW new	353	488	299	200	152	119
PW old	151	254	157	168	181	294
Low frequency						
MS new	1157	1063	501	229	242	218
MS old	72	154	65	85	166	736
MW new	1157	1063	501	229	242	218
MW old	143	225	112	131	181	528
PS new	543	488	237	102	107	99
PS old	125	167	112	99	195	644

(*Appendix continues on next page*)

Appendix (*continued*)

Counts per Confidence Category

Condition	Confidence category					
	− − −	− −	−	+	+ +	+ + +
Low frequency (*continued*)						
PW new	573	481	251	121	116	78
PW old	133	196	124	117	179	454
Word frequency, Experiment 5						
High frequency						
Subject 1						
MS new	300	407	66	144	184	179
MS old	28	54	12	39	64	283
MW new	300	407	66	144	184	179
MW old	32	74	19	39	65	251
PS new	167	211	37	58	95	72
PS old	21	60	13	43	63	280
PW new	149	215	36	70	78	92
PW old	43	47	12	38	60	280
Subject 2						
MS new	605	324	38	40	144	129
MS old	78	62	11	13	75	241
MW new	605	324	38	40	144	129
MW old	118	73	10	21	64	194
PS new	311	151	11	29	80	57
PS old	76	69	13	26	76	220
PW new	349	142	13	19	61	56
PW old	96	54	10	18	67	235
Subject 3						
MS new	136	223	292	402	202	20
MS old	14	46	54	77	79	206
MW new	136	223	292	402	202	20
MW old	33	63	71	106	90	115
PS new	78	116	155	189	91	11
PS old	17	38	57	92	80	196
PW new	72	118	141	187	107	14
PW old	19	49	67	97	102	146
Subject 4						
MS new	784	203	72	46	95	76
MS old	65	24	18	21	48	304
MW new	784	203	72	46	95	76
MW old	117	54	18	28	50	213
PS new	427	76	32	21	41	38
PS old	64	26	13	21	34	322
PW new	344	120	65	31	29	50
PW old	98	52	27	18	56	229
Subject 5						
MS new	52	285	391	340	185	27
MS old	3	28	55	95	126	173
MW new	52	285	391	340	185	27
MW old	4	44	86	114	110	122
PS new	29	164	202	130	89	26
PS old	5	33	57	86	110	189
PW new	30	122	184	192	94	18
PW old	4	44	77	132	128	95
Subject 6						
MS new	30	116	625	456	40	13
MS old	0	13	80	150	38	199
MW new	30	116	625	456	40	13
MW old	4	17	123	187	32	117
PS new	26	70	317	213	9	5
PS old	0	22	74	143	25	216
PW new	18	70	313	227	11	1
PW old	1	16	116	178	38	131
Subject 7						
MS new	358	200	40	49	120	113
MS old	12	7	6	13	51	235
MW new	358	200	40	49	120	113
MW old	32	35	11	19	53	186
Subject 7 (*continued*)						
PS new	194	100	19	19	66	50
PS old	29	24	7	9	39	228
PW new	185	90	25	31	60	57
PW old	30	24	9	24	60	189
Subject 8						
MS new	469	233	95	66	22	11
MS old	51	33	16	34	34	144
MW new	469	233	95	66	22	11
MW old	95	53	25	38	48	101
PS new	263	105	37	25	13	4
PS old	64	28	21	29	31	163
PW new	193	110	51	38	18	5
PW old	81	50	31	46	43	61
Subject 9						
MS new	175	237	205	128	71	64
MS old	27	31	43	45	52	126
MW new	175	237	205	128	71	64
MW old	27	36	53	58	37	125
PS new	75	115	120	74	40	24
PS old	23	29	42	54	58	130
PW new	95	125	95	66	28	39
PW old	32	46	46	52	48	112
Subject 10						
MS new	172	426	183	95	58	30
MS old	6	34	11	39	137	121
MW new	172	426	183	95	58	30
MW old	18	39	20	66	121	115
PS new	105	223	100	45	27	10
PS old	7	22	17	64	146	127
PW new	94	210	90	61	33	14
PW old	24	43	30	50	125	108
Subject 11						
MS new	424	189	18	26	111	112
MS old	10	4	0	5	25	280
MW new	424	189	18	26	111	112
MW old	26	29	4	12	38	227
PS new	226	77	7	14	67	57
PS old	21	16	3	9	27	260
PW new	217	86	18	20	65	42
PW old	24	23	10	11	38	230
Low frequency						
Subject 1						
MS new	625	364	26	50	102	112
MS old	21	27	3	16	33	380
MW new	625	364	26	50	102	112
MW old	55	53	4	16	44	308
PS new	338	189	9	16	44	44
PS old	24	23	3	15	22	392
PW new	327	190	14	20	44	45
PW old	42	56	4	21	26	329
Subject 2						
MS new	821	264	19	24	55	95
MS old	54	37	2	10	41	336
MW new	821	264	19	24	55	95
MW old	90	57	9	9	45	269
PS new	406	139	9	11	23	49
PS old	64	40	10	8	33	323
PW new	450	112	11	9	18	39
PW old	82	35	4	7	33	316
Subject 3						
MS new	566	256	178	176	72	28
MS old	28	14	23	25	57	328
MW new	566	256	178	176	72	28

Appendix (*continued*)

Counts per Confidence Category

Condition	Confidence category					
	− − −	− −	−	+	+ +	+ + +
Subject 3 (*continued*)						
MW old	68	46	38	59	66	203
PS new	296	131	80	78	47	7
PS old	18	21	22	41	52	325
PW new	286	108	101	89	46	9
PW old	46	38	32	51	82	229
Subject 4						
MS new	1010	125	24	16	56	33
MS old	52	13	7	12	32	358
MW new	1010	125	24	16	56	33
MW old	102	38	12	13	44	262
PS new	518	47	25	9	12	17
PS old	55	15	2	5	34	360
PW new	486	72	19	11	26	20
PW old	93	37	18	19	43	265
Subject 5						
MS new	390	432	279	108	61	10
MS old	10	31	24	57	89	269
MW new	390	432	279	108	61	10
MW old	21	74	64	75	99	147
PS new	215	190	134	52	37	11
PS old	24	25	27	50	80	274
PW new	203	207	129	55	35	10
PW old	26	63	48	84	104	155
Subject 6						
MS new	153	260	602	229	26	9
MS old	1	11	54	108	57	249
MW new	153	260	602	229	26	9
MW old	4	32	82	135	46	180
PS new	122	133	276	101	6	1
PS old	0	6	56	100	34	283
PW new	73	122	333	105	6	1
PW old	3	19	90	136	44	187
Subject 7						
MS new	569	190	14	18	37	52
MS old	24	21	3	4	34	250
MW new	569	190	14	18	37	52
MW old	45	42	9	15	36	177
PS new	313	96	3	8	14	14
PS old	30	22	4	8	25	247
PW new	299	85	8	13	24	19
PW old	55	25	8	25	45	178
Subject 8						
MS new	550	208	53	37	33	8
MS old item	32	18	12	19	34	245
MW new	550	208	53	37	33	8
MW old	61	26	12	29	52	128
PS new	286	96	20	24	11	6
PS old	38	19	7	19	26	223
PW new	229	95	32	32	20	6
PW old	44	24	15	32	55	140
Subject 9						
MS new	236	308	164	63	49	55
MS old	7	12	7	23	47	239
MW new	236	308	164	63	49	55
MW old	16	21	31	27	62	166
PS new	123	152	72	37	33	31
PS old	10	18	17	30	49	212
PW new	151	148	73	26	20	30
PW old	16	23	23	20	51	202
Subject 10						
MS new	239	472	127	58	32	19
MS old	4	12	6	20	166	170
MW new	239	472	127	58	32	19
MW old	18	24	10	47	139	107
PS new	121	272	66	19	19	5
PS old	8	16	9	43	146	157
PW new	117	245	69	31	19	12
PW old	9	25	18	40	157	129
Subject 11						
MS new	625	109	7	15	66	58
MS old	14	5	0	6	14	297
MW new	625	109	7	15	66	58
MW old	43	9	1	4	22	245
PS new	311	61	6	14	35	21
PS old	13	7	1	3	17	295
PW new	316	54	5	15	29	29
PW old	34	17	2	13	23	247
Categorized materials, Experiment 6						
Category						
MS new	494	245	118	61	44	101
MS old	34	46	26	48	66	578
MW new	450	285	97	82	55	99
MW old	37	59	38	66	93	501
PS new	532	198	119	70	65	76
PS old	36	46	29	54	86	551
PW new	450	224	115	100	73	92
PW old	44	62	43	59	96	487
Random						
MS new	309	123	63	23	11	7
MS old	14	22	16	26	37	285
MW new	281	135	61	26	19	10
MW old	52	28	22	27	41	231
PS new	315	116	57	22	13	10
PS old	27	23	16	30	34	274
PW new	296	126	60	23	12	10
PW old	37	37	19	32	43	224
Word pool						
MS new	268	140	59	33	20	9
MS old	31	28	21	29	40	247
MW new	265	141	78	21	16	13
MW old	40	56	18	43	36	207
PS new	279	120	66	43	17	6
PS old	34	24	26	32	51	236
PW new	239	150	63	40	14	16
PW old	37	50	40	25	34	202

Note. Number of counts in each confidence category from sure new (− − −) to sure old (+ + +) by experiment. Experiments 1, 2, 4, and 6 have group data (subjects ran one session each) whereas Experiments 3 and 5 have individual subject data because subjects ran in multiple sessions. The experiments have responses with long reaction times (e.g., more than 9 s) eliminated. These data are used to construct the *z*-ROC (receiver operating characteristic) functions shown in the figures in the experimental results. LL = list length; W = weak; S = strong; MS = mixed strong; MW = mixed weak; PS = pure strong; PW = pure weak.

Received July 15, 1992
Revision received June 14, 1993
Accepted August 23, 1993 ■