Journal of Memory and Language **40,** 432–453 (1999)
Article ID jmla.1998.2623, available online at http://www.idealibrary.com on IDEAL®

Two-Process Models of Recognition Memory: Evidence for Recall-to-Reject?

Caren M. Rotello

University of Massachusetts at Amherst

and

Evan Heit

University of Warwick, Coventry, United Kingdom

According to two-process accounts of recognition memory, a familiarity-based process is followed by a slower, more accurate, recall-like process. The dominant two-process account is the recall-to-reject account, in which this second process facilitates the rejection of similar foils. To evaluate the recall-to-reject account, we reanalyzed two experiments from Hintzman and Curran (1994) in which subjects made word recognition judgments at different response deadlines, and we conducted two new recognition experiments using pairs of similar pseudowords. The new analyses included modeling at both the group and individual subject levels. The results did not provide any distinctive evidence for recall-to-reject. In addition to discussing this two-process account, we describe a one-process account of recognition, in which the nature of similarity information varies across the course of judgment.

Key Words: recall-to-reject; time course; recall; recognition; response signal.

Current research in memory is going through an exciting period, in which new dimensions are being investigated and new challenges are being created for models of memory. We refer in particular to recent research on the time course of memory, investigating how people make recognition judgments over time (e.g., Dosher & Rosedale, 1991; Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994). What kinds of processing take place during the first few hundred milliseconds of a recognition judgment? Likewise, what kinds of processing have a longer time course, possibly constituting optional elements of a recognition judgment? Historically, most research on recognition memory has focused on the end-product of recognition, that is, the final outcome of a recognition judgment. However, results showing the time course of recognition should be highly constraining on theoretical accounts of memory.

Generally speaking, models of recognition can be divided into one-process accounts and two-process accounts. Single process models, such as the *global memory models* (e.g., SAM, Gillund & Shiffrin, 1984; MINERVA 2, Hintzman, 1988; other multiple-trace models such as Estes, 1994; Heit, 1993; Nosofsky, 1988; composite-vector models such as CHARM, Metcalfe, 1982, and TODAM, Murdock, 1982; and

Caren Rotello has previously published under the name Caren M. Jones. We acknowledge the contribution of Michael Wenger to the implementation of Experiment 1, and we also thank Sonya Dougal for assistance with data collection. We thank Lewis Bott for contributions to implementing and running Experiment 2. We are grateful to Gordon D. A. Brown, Tim Curran, Scott Gronlund, Jason Hicks, Douglas Hintzman, Koen Lamberts, Neil Macmillan, Richard Marsh, Elizabeth Maylor, Doug Nelson, Michael Wenger, and anonymous reviewers for comments on this work. This research was supported by an Economic and Social Research Council grant to Evan Heit and by a Healey Endowment grant from the University of Massachusetts to Caren Rotello. Portions of these data were reported at the 38th Annual Meeting of the Psychonomic Society in Philadelphia.

Address correspondence to Caren Rotello, Department of Psychology, University of Massachusetts, Amherst, MA 01003-7710 or to Evan Heit, Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom. Electronic mail may be sent to caren@psych.umass.edu or to E.Heit@warwick.ac.uk.

0749-596X/99 $30.00

some connectionist networks: Kortge, 1990; Ratcliff, 1990), assume that recognition judgments, as well as judgments of frequency, are based on the target item's *familiarity* or its *total similarity* to the contents of memory (Jones & Heit, 1993). Although the global memory models have been widely and successfully applied to many results (see Clark & Gronlund, 1996; Raaijmakers & Shiffrin, 1992, for reviews), recent findings are discrepant with the core predictions of these models.

Specifically, the total similarity principle does not hold for all judgments. Subjects often reject new test items that are similar to a repeated study item (Hintzman, Curran, & Oppy, 1992; Jones & Heit, 1993). Jones and Heit (1993) found that prior presentations of words (e.g., *salmon*) did not influence frequency estimates or recognition judgments for unstudied similar words (e.g., *tuna*), even when the presented word had been shown as many as 18 times. In contrast, when both the old word and a similar word had been studied, the frequency estimates increased with the presentation frequency of either word. Similarly, Hintzman et al. (1992) had subjects study nouns that were either singular (e.g., *frog*) or plural, and then give frequency estimates for the words in either their original plurality (e.g., *frog*) or with the addition or deletion of an "s" (e.g., *frogs*). Hintzman et al.'s subjects produced frequency estimates for the similar distractor items that were bimodal: About half of the judgments were zero and about half of the judgments were roughly the frequency of the studied item. In both of these studies, subjects gave a surprising number of zero judgments for new items that were similar to old items which were presented many times.

These discrepant results (as well as those of others, e.g., Ratcliff, Clark, & Shiffrin, 1990; Ratcliff, Sheu, & Gronlund, 1992, see Clark & Gronlund, 1996, for a review) have recently led researchers (e.g., Hintzman & Curran, 1994; Jacoby, 1991) to suggest that one-process models of recognition memory, using a single familiarity measure, may be incomplete. It has been suggested that a second, recall-like process operates as well (see also Atkinson & Juola, 1973, 1974; Mandler, 1980). This second process is generally supposed to be slower than the initial, familiarity-based process, but it would be more accurate on fine discriminations between items. For example, the second process could lead people to state correctly that *tuna* has been presented zero times when it was another item, *salmon*, that actually had been presented. Likewise, the second process could help people to distinguish *frogs* from *frog*, though it might be expected to make some errors on such highly similar pairs.

Some recent findings on the time course of recognition judgments do suggest that familiarity-based information is used early in processing, whereas recall-like information would be used later in processing. Hintzman and Curran (1994) conducted experiments like those of Hintzman et al. (1992), wherein subjects were required to respond at various deadlines ranging from 100 to 2000 ms after the appearance of the test item. Early in processing, response rates on old items (e.g., *frog*) and similar, unpresented items (e.g., *frogs*) tracked each other. But later in processing, responses on the two kinds of items diverged, with the hit rate for old items increasing and the false-alarm rate for similar items decreasing somewhat. In some cases the false-alarm rates to similar items rose significantly during the early stages of processing and fell significantly later in processing. These results are consistent with the claim that an early, familiarity-based process would not distinguish well between *frog* and *frogs*, but later, recall-based information would make this discrimination possible.

There are several possible variants of two-process models and a number of issues that might be addressed within a two-process framework. One important issue is how people would use information from a second, recall-like process. The dominant, rather well-specified view is the *recall-to-reject* account (Clark, 1992, p. 241; Gronlund & Ratcliff, 1989, p. 857; Hintzman & Curran, 1994, p. 14; see Clark & Gronlund, 1996, pp. 56–57 for a review). Again, assume that the word *frog* has been studied but the word *frogs* has not. When a subject is tested on *frogs*, the second process leads to a *frog* memory trace being recalled, but no *frogs* memory trace. In the face of this negative evi-

dence, *frogs* is rejected and is given a "new" judgment regardless of the output of the first, familiarity-based process. That is, a recall-to-reject process would use information about a recalled item specifically to reject another, similar foil that cannot be recalled. The rationale is that recalling *frog*, and failing to recall *frogs*, leads the subject to attribute the familiarity of *frogs* to its similarity to the *frog* memory trace. Of course, the more time that is allowed for the recognition judgment to be made, the more likely it is that the recall process would have successfully retrieved *frog*, so the more likely it is that *frogs* would be rejected. We note that this recall-to-reject process would not affect judgments on true foils (e.g., *chair*), which are items that have not been seen and are not similar to any studied item. Presumably, true foils would not be familiar on the basis of the first process, so a second process would not be needed to reject them. The recall-to-reject account is consistent with the increasing ability of Hintzman and Curran's subjects to distinguish between old and similar, unpresented items late in processing, when the recall process might contribute to the judgments.

Which response measures should be used to evaluate the recall-to-reject account? Generally speaking, a number of measures could be used to evaluate two-process accounts, and a number of results might be suggestive of two processes being at work. For example, the increasing then decreasing pattern of false alarms on similar items (*frogs*) in Hintzman and Curran (1994) is suggestive of a sequence of two processes. However, it is generally accepted that raw scores such as hit rates and false-alarm rates should be interpreted with caution, because they depend not only on signal strength but also on response bias. We would suggest that, in a typical response-signal experiment, in which a subject might sometimes have to respond after 100 ms and other times respond as much as 2000 ms later, it is extremely plausible that a subject may have different response biases and decision criteria at different deadlines. Hardly any information would be available early in processing, so a lenient criterion for saying "old" might apply. On the other hand, when further information is available a stricter criterion might be in effect.

Consistent with this proposal, there is a marked trend in Hintzman and Curran's results for the false-alarm rate on completely new items to decrease over the course of judgment, as if the response criterion were stricter at longer deadlines. (Similar patterns of decreasing false alarms across time have been observed in essentially all response-signal experiments.) These completely new items would not be familiar at all and likewise would never be recalled, so it is plausible that the decreased responses to these items merely reflect a change in criterion rather than some operation of a familiarity or recall-like process. We do consider it unlikely that each subject manages to keep a constant response criterion as the deadline, and thus the demands of the task, varies. Whereas Hintzman, Caulton, and Curran (1994) did not find significant changes in response criterion at different response deadlines, this issue has not been extensively considered, and it is not clear whether the finding would generalize to other studies. Indeed, measures of response criterion can be difficult to interpret when sensitivity (e.g., d') is also changing (e.g., Feenan & Snodgrass, 1990; Snodgrass & Corwin, 1988).

The traditional alternative to raw scores such as hit rate and false-alarm rate is a discrimination measure such as d' or d_L.[1] These measures are in essence the difference in responses to two items at a given deadline, so any change in response bias for that deadline should be canceled out.[2] For example, responses to an old

[1] d_L is a discrimination measure based on a logistic rather than normal distribution, which we adopt to maintain consistency with Hintzman and Curran (1994). It is equal to roughly 1.67 times d' and it is easier to compute than d'. It is computed as $d_L = \ln[HR(1 - FA)/FA(1 - HR)]$, where HR is the hit rate and FA is the false-alarm rate.

[2] Others have suggested that guessing rates or guessing processes might also change over different response deadlines (e.g., Meyer, Irwin, Osman, & Kounios, 1988). This point further supports our argument that hit rates and false-alarm rates should not be taken too literally, because they reflect other factors besides test item strength, such as guessing. In the present case, we use the framework of signal detection theory primarily as a means of data analysis, for measuring the difference between two distributions of response rates, rather than as a complete description of how recognition processes might change overtime.

item, *frog*, might be evaluated in terms of d_L at short and long deadlines. At each deadline, d_L would be computed by comparing *frog* hit rates to false alarms on completely new foils such as *chair*. If d_L on *frog* increases from the short deadline to the long deadline, there would be evidence for increased strength of old items which is independent of response bias.

Likewise, it is possible to analyze responses on similar, new items using difference scores. Here, responses on an unpresented item (e.g., *frogs*) which is similar to an old item would be compared to responses on a completely new foil (e.g., *chair*). Once again, the difference score could be expressed in discrimination units. Dosher (1984) used such an analysis for recognition judgments in the response-signal paradigm, referring to the measure as *pseudo-d′*; more recently, this measure has been used for similar experiments by Dosher and Rosedale (1991), Gronlund and Ratcliff (1989), and Ratcliff and McKoon (1989). Higher values of d' or d_L indicate a greater propensity to accept similar items as old, relative to completely new items. If the discrimination measure decreases for *frogs* at longer deadlines, there would be evidence of greater rejection of similar items independent of a possibly changing response criterion.

Critically, using a difference score to compare false alarms for similar items to false alarms for true foils is most appropriate for testing the recall-to-reject account because, according to this account, the recall-like process is used exclusively to reject items that are similar to what has been observed. The recall-to-reject process would affect similar foils, but, by definition, it would not affect true foil items that are not similar to studied items. Hence it is appropriate to use a difference score, comparing judgments on similar foils to judgments on true foils. At later points in the course of judgment, this difference score should reflect the differential operation of a recall-like process on the two kinds of stimuli. We emphasize that whether there is actually a change in response criterion at different lags, this difference score is the most direct, specific, and theoretically motivated measure of recall-to-reject processing.

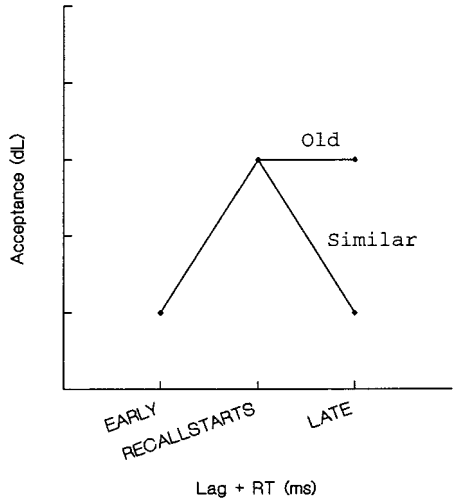

FIG. 1. Idealized predictions for the recall-to-reject account for old items and similar, unpresented, items at various response deadlines.

The (highly idealized) predictions of the recall-to-reject account are shown in Fig. 1. Early in processing, the discrimination measure (d_L) should increase for both old items and similar items, due to the influence of the familiarity-based process. But later in processing, when the recall process is providing useful information, the acceptance measure should decrease for similar items as these items are rejected. In the simplest case, d_L for old items would stay flat after the recall process becomes the predominant recognition mechanism, because the recall-to-reject process would not affect old items.

OTHER ISSUES

Before we turn to the available data in our effort to evaluate the recall-to-reject account, we briefly address some related issues. There are other questions that might be raised in the context of two-process memory models. For example, could the two processes overlap, so that at some points there is information from both processes becoming available? Our assumption is simply that information from the first process would begin to show its influence early in the course of processing and that the influence of the second process would appear relatively late in processing. We will be exam-

ining experiments that were designed so that foil items would be accepted by a familiarity-based process and rejected by a recall-to-reject process. Hence the recall-to-reject account would predict increasing acceptance (in terms of d_L) at initial lags, followed by decreasing acceptance at longer lags. None of our analyses make any strong assumptions about whether the two processes might overlap or even about when exactly the first and second processes would show their influences. For example, the modeling analyses will treat the onset times of the two processes as free parameters and allow for some period when both processes might be active.

A more far-reaching question is whether there are two processes at all. Some researchers (e.g., McClelland & Chappell, 1995; Mulligan & Hirshman, 1995; Shiffrin & Steyvers, 1997) have offered one-process accounts for results that seemingly indicated two processes. It should be noted that we do not presume to settle the one-process versus two-process debate in this paper. In the General Discussion we will consider how a one-process model might accommodate the results as well. However, next we continue with our main purpose, which is to contribute to the exposition of two-process models by analyzing data within this framework, in particular by assessing the recall-to-reject account.

PREVIOUS RESULTS

Our first step in evaluating the recall-to-reject account was to apply our new analyses to the results of two experiments by Hintzman and Curran (1994).[3] In these deadline-procedure experiments, subjects made recognition judgments on old items (e.g., *frog*), similar new items (e.g., *frogs*), and completely new items (e.g., *chair*). We would emphasize that Hintzman and Curran did not specifically address the recall-to-reject account in their own analyses, but instead looked more generally for evidence of biphasic patterns of responding over the time course of judgment. These well-conducted experiments have a useful design for evaluating recall-to-reject; therefore, we attempted to gain additional value from this data set by using it to test other, related issues.

Hintzman and Curran's (1994) Experiment 3

Hintzman and Curran's (1994) Experiment 3 used a wide range of response signal lags (100, 200, 350, 500, 750, 1200, 2000 ms). Twenty-six participants studied lists of words one or twice each and then made recognition judgments on old words, words that were new but similar to the studied words (differing only in plurality), and completely unrelated new foils.

First, we review the hit rate and false alarm results, presented in Hintzman and Curran's Fig. 4. There was an increasing trend for hits on old items as well as a decreasing trend for false alarms on new items, as the response lag increased. For the similar items, the false alarm rate increased up until the 350 ms response deadline and then decreased. The peak is fairly pronounced for the words similar to an item presented twice, but less pronounced for the words similar to an item presented once. On the basis of false alarms to similar items, Hintzman and Curran concluded that there was evidence for a recall-like process. In particular, they were looking for an inverted-U pattern of responding (a "peak"), with false alarms to similar items increasing early in processing and then decreasing later in processing. Using a quadratic polynomial contrast analysis, they found some evidence consistent with this pattern. However, even a finding of a significant quadratic component is not the strongest evidence for an inverted-U pattern, because a quadratic component merely indicates that the data show a nonlinearity (a "bend"); it does not imply that the data show a significant non-monotonicity. For example, a significant quadratic component is consistent with data that rise rapidly and then level off to an asymptote (Glass & Hopkins, 1984). To investigate this question further, we performed Tukey's tests and found that there were indeed significant drops. For words similar to an item presented once, there is a significant drop at the 2000-ms deadline, compared to the

[3] We are greatly indebted to Douglas Hintzman for providing data from these experiments. A small number of subjects' data points, recorded as 0.00, were replaced with 0.01 so that d_L could be computed.

350- and 500-ms deadlines. For words similar to items presented twice, there is a significant drop at the 750-ms deadline compared to the 350-ms deadline, and there is a significant drop at the 1200- and 2000-ms deadlines compared to the 350- and 500-ms deadlines.

The analyses on raw scores may provide suggestive evidence for a change in processing, but they do not specifically address the recall-to-reject account. Consequently, we conducted further analyses using derived d_L measures. These measures had three main benefits. First, as is conventionally accepted for discrimination measures, they were an effort to distinguish between two sources of saying "yes" to a test stimulus: signal strength and guessing or other response biases. Indeed, Hintzman and Curran reported a sharply decreasing false alarm rate on new items at increasing response deadlines. This unexplained finding calls into question the interpretation of false alarms on similar items. Second, our main analysis, using d_L to measure acceptance of similar items, was specifically aimed at finding evidence for a recall-to-reject process, which would operate on new items that are similar to old items. A recall-to-reject process would affect similar items but not true new items, causing the acceptance of similar items to decline as the recall process had more time to operate, but not affecting the acceptance of new foils. Finally, we fitted the discrimination measures with mathematical functions that either rise monotonically to an asymptote or initially rise and then decline as more time elapses. Thus, we can test directly whether the data show any significant non-monotonicities that would be consistent with the predictions of a recall-to-reject process.

For each of the deadlines, we computed d_L measures for old items and similar items by comparing acceptance rates on these items to acceptance rates on completely new items at the same deadline. The results of these analyses are shown in Fig. 2. In this graph and subsequent figures, the *x*-axis indicates the response deadline plus the average time to respond at this deadline (i.e., signal lag + response latency). The pattern of results is remarkably clear and is inconsistent with the recall-to-reject predictions. Early in processing (up until about 600 ms have elapsed), d_L increases for both old items and similar items. Later in processing, however, the acceptance rate (d_L) for old items continues to rise, whereas it stay roughly flat for similar items.

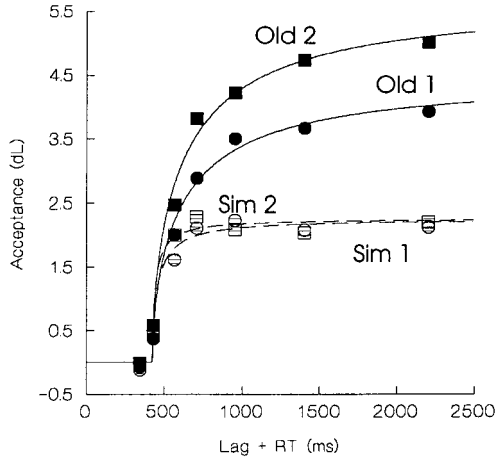

FIG. 2. Acceptance of stimuli (in terms of d_L) for Hintzman and Curran (1994), Experiment 3, at varying response deadlines.

We performed ANOVAs on these d_L measures, which supported our interpretation of Fig. 2. For the old items, there was a main effect of lag (100, 200, 350, 500, 750, 1200, or 2000 ms), $F(6,150) = 119.02$, $MSe = 1.45$, $p < .001$, and a main effect of presentation frequency (1 or 2), $F(1,25) = 76.53$, $MSe = .50$, $p < .001$. In addition, there was an interaction between these factors, $F(6,150) = 9.58$, $MSe = .23$, $p < .001$. For old items presented once, there were significant increases from the 100-ms deadline to the 200-ms deadline, from the 200-ms deadline to the 350-ms deadline, from the 350-ms deadline to the 500-ms deadline, from the 500-ms deadline to the 750-ms deadline, and from the 750-ms deadline to the 2000-ms deadline. The contrast analyses for the old items presented twice led to the same conclusions.

Next and more importantly, for the similar items, there was a main effect of lag, $F(6,150) = 52.75$, $MSe = .86$, $p < .001$, but the effect of frequency was not significant, $F(1,25) = 1.74$, $MSe = .30$. Also, there was no significant interaction between these factors, $F(6,150) =$

1.94, $MSe = .21$. Thus we performed contrast analyses collapsing over the frequency factor. The results of the contrast analyses are simple to describe. There is no significant increase from 100 to 200 ms, but the difference from both these cells to the remaining cells, 350, 500, 750, 1200, and 2000 ms, is significant. These latter cells did not differ significantly from each other. There was no evidence that responses to similar items, in terms of the d_L measure, either increased or decreased late in processing.

Finally, we conducted model-based analyses that most directly addressed whether there were decreasing responses to similar items at later deadlines. The data in Fig. 2 were fit to two separate functions. The first function represents monotonic growth to a limit (Ratcliff, 1978):

$$d_L(lag) = \frac{\lambda_1}{\sqrt{1 + \nu/(lag - \delta)}}, \qquad [1]$$

where *lag* is the response signal lag plus the latency to respond at that lag (i.e., total processing time), λ_1 is the asymptotic level of discrimination, ν is the rate of approach to that asymptote, and δ is the point at which discrimination rises above chance. The second function we fit to the data is a non-monotonic function of a similar form. It assumes that the data conform to Eq. 1 until a particular point in processing, *lag**, at which point discrimination begins to decline to a second asymptote. (Because we were looking for peaked functions, the second asymptote was constrained to be no higher than the first.) The non-monotonic function is given by

$$d_L(lag) = \frac{\lambda_2 + (\lambda_1 - \lambda_2)(lag^* - \delta)/(lag - \delta)}{\sqrt{1 + \nu/(lag - \delta)}} \quad \text{for lag} \geq \text{lag}^*. \qquad [2]$$

Equations 1 and 2 (and variants) have been widely used (e.g., Dosher, 1984; Gronlund & Ratcliff, 1989; Ratcliff, 1980; Wickelgren & Corbett, 1977).

For each type of stimulus, distinct rates and asymptotes were estimated, but a single intercept (δ) was used for all the stimuli, and likewise a single *lag** was used for the non-monotonic fits. Model fitting was performed using SYSTAT's Simplex estimation method to iteratively adjust parameters so as to minimize the residual sum of squares (RSS). Consistent with the ANOVAs on these d_L data, we found that the non-monotonic function (R^2 adjusted = .996, RSS = .445)[4] did not fit the data significantly better than the monotonic function (R^2 adjusted = .997, RSS = .455; $\chi^2(3) = .622$, n.s.).[5] The best-fitting parameter values for both models are shown in the bottom half of Table 1; the predictions of the monotonic model are shown as the curves in Fig. 2. We found no significant decrease in the acceptance of similar items, relative to completely new foils, late in processing.

These model-based analyses focused on average group responses rather than individual subjects' responses. It is useful to look at group responses, particularly when not many data have been collected for each individual, but there is some risk that the averaged group-level responses may not capture all the characteristics of individual subjects' data. Hence, we repeated

[4] R^2 adjusted is the proportion of variance accounted for by the model, adjusted for the number of free parameters.

$$R^2_{adjusted} = 1 - \frac{\sum_{i=1}^{N} \frac{(d_i - \hat{d}_i)^2}{(N - k)}}{\sum_{i=1}^{N} \frac{(d_i - \bar{d})^2}{(N - 1)}},$$

where N is the number of data points (d_i), $\hat{d}_i$ is the predicted value, k is the number of free parameters, and $\bar{d}$ is the overall mean.

[5] These nested models were compared using the technique of Borowiak (1989). In brief, when model A is a nonlinear model with a free parameters estimated using a least-squares criterion, and B is a restricted version of this model with b free parameters, the likelihood ratio statistic is $\lambda = (RSS_A/RSS_B)\,(k/2)$, where RSS is the residual sum of squares of the model and k is the number of data points to be predicted (for this experiment, 20). Borowiak showed that $-2 \ln(\lambda)$ has a χ^2 distribution with $(a - b)$ *df*. (See Heit, 1998, and Lamberts, 1994, for other applications of this technique.) The models differed by 3 free parameters; hence, we used 3 *df*. We employed an $a = .05$ criterion.

TABLE 1

Best-Fitting Parameter Values for the Monotonic and Non-monotonic Models to the Data from Hintzman and Curran's (1994) Experiments 2 and 3

		Monotonic model			Non-monotonic model				
		Intercept	Asymptote	Rate	Intercept	Lag*	Asy. 1	Asy. 2	Rate
H&C, Expt. 2									
Old d_L	Freq. 1	473.6	3.57	230.9	473.6	—	3.57	—	230.9
	Freq. 2	473.6	3.75	923.5	473.6	—	3.75	—	923.5
Similar d_L	Freq. 1	473.6	1.74	67.7	473.6	473.6[a]	1.74	1.74	67.7
	Freq. 2	473.6	1.92	38.7	473.6	473.6[a]	1.92	1.92	38.0
H&C, Expt.3									
Old d_L	Freq. 1	428.7	4.48	424.1	428.1	—	4.49	—	429.5
	Freq. 2	428.7	5.70	435.0	428.1	—	5.71	—	439.9
Similar d_L	Freq. 1	428.7	2.26	80.7	428.1	454.0	2.27	−1.0	115.2
	Freq. 2	428.7	2.26	44.9	428.1	454.0	2.23	1.83	57.8

Note. H&C, Hintzman and Curran (1994); Freq., presentation frequency; Asy. 1, the earlier asymptote; Asy. 2, the later (lower) asymptote.

[a] Multiple values of this parameter make the same predictions. Sample value is shown.

our model-fitting analyses on the data from each subject individually. If subjects were individually showing recall-to-reject processing, then individual fits of Eqs. 1 and 2 should reveal that many (if not most) subjects' data were better fit by a non-monotonic function. Our analyses did not support this position, however. Of the 26 subjects, only 2 had data that were significantly better fit by the non-monotonic model than by the monotonic model. By chance alone, there would be a 38% probability of finding 2 or more significant results of 26. In addition, the incremental fit of the non-monotonic model, summed over all the subject's individual fits, did not significantly differ from chance, $\chi^2(78) = 57.58$, n.s.

In sum, the raw scores reported by Hintzman and Curran for this experiment were suggestive of a change in processing at later lags, but the ANOVA and model-based analyses of d_L did not provide any distinctive evidence for recall-to-reject.

Hintzman and Curran's Experiment 3 also included another condition, the Inclusion condition, in which subjects were told to respond positively to both old items and similar, unpresented items. Hintzman and Curran themselves computed d_L for this condition, comparing responses on old items to responses on new items as well as comparing responses on similar items to responses on new items. The results were compatible with those of our other analyses. That is, early in the time course of judgment, the difference measures for old items and similar items tracked each other, whereas later in the course of judgment, the measures for the two kinds of items diverged, with greater acceptance for old items than for similar items. However, this Inclusion condition does not address the recall-to-reject hypothesis directly, because subjects were not instructed to reject similar items. Put together, the pattern of false alarms in both conditions of Experiment 3 could be considered as a kind of suggestive evidence for a second process in recognition, but they do not represent a distinctive test of the recall-to-reject account.

Hintzman and Curran's (1994) Experiment 2

We applied these analyses to Hintzman and Curran's (1994) Experiment 2, which was similar to their Experiment 3 but included a smaller range of response signal lags (100, 250, 500, 800, and 1200 ms). Our conclusions based on these data were the same as those described for Experiment 3, except that there was no evidence

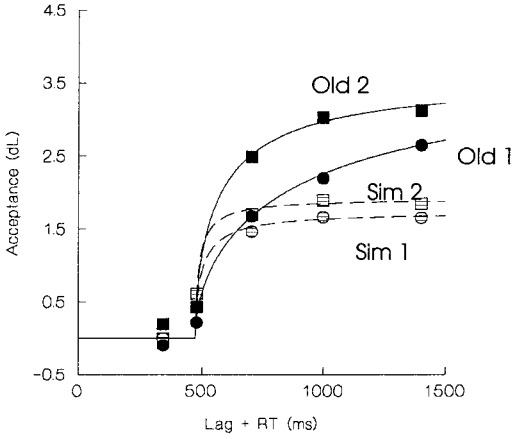

FIG. 3. Acceptance of stimuli (in terms of d_L) for Hintzman and Curran (1994), Experiment 2, at varying response deadlines.

of a significant rise-then-fall pattern in the false-alarm rate to similar foils. The discrimination (d_L) scores computed from these data are shown in Fig. 3, along with the curves from the best-fitting monotonic functions. The fit of the monotonic model was excellent (R^2 adjusted = .998, RSS = .076) and the fit of the non-monotonic model was not significantly better (R^2 adjusted = .997, RSS = .076, $\chi^2(3) = 0.0$, n.s.). (The best-fitting parameter values for both models are shown in Table 1.) Moreover, fits of the models to the individual data revealed that only 3 of the 26 subjects in their Experiment 2 made judgments that were significantly better fit by the non-monotonic function than by the monotonic function; the improvement in fit with the non-monotonic function, summed over subjects, was not reliable: $\chi^2(78) = 57.60$. Consistent with Hintzman and Curran's Experiment 3 data, Experiment 2 also revealed no distinctive evidence for the recall-to-reject account.

Evaluation

We did find the lack of evidence so far for recall-to-reject to be surprising, considering the high similarity within pairs of words in the Hintzman and Curran experiments. Such pairwise similarity might be expected to facilitate the recall process. That is, if a recall process is operating, when *frogs* is tested the subject ought to be able to recall a *frog* memory trace easily. However, these stimuli do represent such an extremely high degree of similarity, both in terms of word meaning and orthography, that people might have some difficulty discriminating between old words and similar words. Even a slow, fairly accurate recall-like process might confuse *frog* and *frogs.* Therefore it is hard to tell by intuition to what extent Hintzman and Curran's stimuli actually encouraged recall-based processing.

We conducted two new experiments with the primary purpose of trying to replicate the basic Hintzman and Curran (1994) findings. For the most part, we followed Hintzman and Curran's methodology. The main change was that we used pseudoword pairs such as *PRUMIR-PRAMAD* which have orthographic similarity but not semantic similarity.[6] Although we did not discover direct evidence for recall-to-reject in the previous experiments, we speculated that the outcome might be different with other kinds of stimuli. On the other hand, if the two new experiments also failed to give direct evidence for recall-to-reject, there would then be a fairly strong case against this account.

EXPERIMENT 1

Our first experiment examined the time course of orthographic similarity effects and was similar to those of Hintzman and Curran (1994) in that subjects made recognition judgments about old items, similar items, and completely new items. The stimuli were also similar to those of Hintzman and Curran; the paired stimuli differed only in terms of one or a few letters. However, we varied the position of the differing letters, unlike Hintzman and Curran, who varied only the final "s." As we shall suggest in the General Discussion, there may be distinctive characteristics associated with the

[6] A third experiment used synonym pairs (*ATTORNEY–LAWYER*) as stimuli. It led to the same conclusions as our Experiments 1 and 2. We have chosen not to report those data because subjects in that experiment made fewer recognition judgments than is typical in a response-signal paradigm; thus the data were relatively noisy.

final "s," such as slower processing, that could be responsible for a particular pattern of results.

Method

Subjects. Ten University of Massachusetts undergraduates participated in four 1.5 h sessions and were paid $30; one additional undergraduate participated in three sessions and received $22.50.

Stimuli. Subjects studied lists of pseudowords and were asked to make recognition judgments on old items, new items that were similar to old items, and completely new items.

The 1080 pseudoword pairs were generated as follows: Pronounceable pseudowords were created for each subject from a set of 3 basic stimulus frames (CCVCVC, CVCCVC, or CVCVCC). For each pseudoword (e.g., *PRUMIR*), either a low similarity or a high similarity partner was generated. High similarity pseudowords differed by one vowel (e.g., *PRAMIR*); low similarity pseudowords differed by both vowels and the final consonant (e.g., *PRAMAD*). The interpair similarity was minimized by use of a wide variety of consonant combinations within each basic frame.

Each pseudoword in each pair was presented 0, 1, or 3 times. Presentation frequencies were assigned to stimulus pairs so that there were 21 pairs with each of the symmetric target-similar pseudoword frequencies 0–0 (neither pseudoword presented), 1–1 (both pseudowords shown once), and 3–3. Because of the symmetric nature of these frequencies, testing either pseudoword in the pair provided a data point in an *X–X* frequency condition; thus there were 42 testable pseudowords (2 from each pair) in each *X–X* condition. An additional 42 pairs were assigned to each of the target-similar pseudoword frequencies 1–0, 3–0, and 3–1. These word pairs had asymmetric frequencies, so testing one pseudoword in the pair provided a data point in an *X–Y* condition and testing the other pseudoword provided a *Y–X* data point. Thus, there were 42 pseudowords (one from each pair) in each *X–Y* or *Y–X* condition.

There were also 42 pairs at each of the target-similar pseudoword frequencies 0–0, 0–1, and 0–3 for which only one pseudoword was tested. For these single-test pairs, only the pseudowords designated as targets appeared on the recognition tests. This somewhat elaborate assignment of presentation frequencies ensured that there would be equal numbers of studied and unstudied pseudowords on the recognition tests. The assignment of stimulus pairs to frequencies, and the selection of the "target" pseudoword for each pair, were randomized for each subject.

The pairs of pseudowords at each frequency combination were evenly divided into 24 study lists, each of which contained a total of 96 stimulus presentations: 24 pseudowords shown with a frequency of 1 and 24 shown three times. Also, there were 42 pseudowords that had a frequency of 0 and were not studied. The order of the stimuli was randomized for each list and for each subject, with the constraint that repetitions of a pseudoword be separated by at least 3 other stimuli.

The recognition test following each study list consisted of 72 items: 18 pseudowords that had been presented once, 18 that had been presented three times, and 36 distractors (24 similar pseudowords and 12 completely new pseudowords) with presentation frequency 0. At each signal lag (30, 90, 270, 500, 700, and 1000 ms) on each test, there was a target pseudoword with each target-similar word frequency. For example, there was a test stimulus that had been studied three times and whose partner had been studied once (i.e., the target from a 3–1 pair) at each signal lag. The test order was randomized for each subject and for each list.

Procedure. At the beginning of the first session, subjects learned to respond in a response-signal paradigm. For each of the 48 training trials, they were shown an orienting cue (- ->+<- -) centered on the monitor of a computer for 500 ms. The cue was replaced with the word YES or NO, which appeared for a variable amount of time (30, 90, 270, 500, 700, or 1000 ms) before being replaced with a specific mask (- -***- -). At the same time the mask appeared, a 25-ms signal tone sounded, signaling that the subject should press "z" if the word YES had been shown and "/" if NO had been shown. They were instructed to respond after the signal, but within 100–400 ms; they received accuracy

TABLE 2

Hit Rates (HR) and False-Alarm Rates (FA) for Experiment 1

	Response deadline (ms)					
Stimulus	30	90	270	500	700	1000
HR						
Old 1	.55	.58	.66	.66	.67	.68
Old 2	.61	.63	.75	.86	.87	.89
FA						
Similar 1	.57	.59	.57	.54	.47	.44
Similar 3	.63	.62	.68	.63	.60	.60
New	.59	.55	.50	.29	.23	.21

and latency feedback after each trial. There were eight training trials at each of the six signal lags; half were YES and half were NO.

Following the training trials, subjects were presented with a series of six 96-stimulus presentations. For each of the six study lists, the words were shown one at a time, centered on the screen, at a 4-s rate. Immediately following each list, subjects made speeded recognition decisions about items in that list. Each test trial began with an orienting cue, which was replaced by a test word that remained on the screen for variable amount of time (30, 90, 270, 500, 700, or 1000 ms) before being replaced by a randomly generated letter mask. At the same time as the mask appeared, a 10-ms signal tone sounded, indicating that the subject should press "z" or "/" to indicate that the word was old or new. Subjects were instructed to respond after the tone, within 100–400 ms. They received latency and accuracy data after each response; in addition, summary data were provided at the end of each test list. Subjects were told to emphasize timing over accuracy. Breaks were allowed between study lists.

The second through fourth experimental sessions were conducted like the initial session, except that no training trials were included. The time between sessions was minimized for each subject, with the constraint that only one session be run each day.

Results

Data from the initial training session for each subject were discarded. All analyses were based on only those responses that occurred between 100 and 350 ms after the response signal. Roughly 18% of the data were excluded.[7] Preliminary analyses did not reveal any significant differences with regard to stimulus frame or similarity level (one or three letters changed), so results are reported collapsed over those variables.

For completeness, we provide the hit rates and false alarm rates for old, similar, and new items in Table 2. These values were adjusted by adding 0.5 to the number of YES ("old") responses and dividing by the total number of responses + 1.0, so that d_L could be calculated later (Hintzman & Curran, 1994; Snodgrass & Corwin, 1988). As can be seen in Table 2, positive responses to studied items increased over processing, and positive responses to completely new items decreased. These data, like those from Hintzman and Curran's Experiment 3, also show at least some suggestion of a non-monotonicity in the false alarms to similar items. Tukey's tests showed that the false alarms to pseudowords that were similar to items studied three times showed a peak at 270 ms: False alarms to those items increased from the 90- to the 270-ms lag and then decreased from 270 to 500 ms. However, acceptance of pseudowords that were similar to items studied once showed a general decreasing trend, with significant drops from the 90-ms lag to the

[7] We also analyzed the data using a 100- to 500-ms inclusion criterion. That analysis, from which only 5% of the data were excluded, led to the same conclusions.

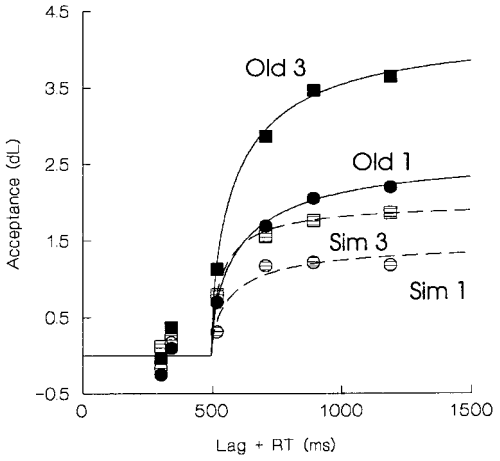

FIG. 4. Acceptance of stimuli (in terms of d_L) for Experiment 1, at varying response deadlines.

500-ms lag and again from the 500-ms lag to the 700-ms lag.

However, the most striking finding in Table 2 is the marked decrease in false alarms to completely new items as the test lag increases. This result reinforces our argument that the responses on old and similar items must be interpreted relative to some baseline which may itself vary across the time course of recognition.[8]

The results of the d_L analyses for Experiment 1 are shown in Fig. 4. The pattern of results is similar to that of the other experiments, showing a general increasing trend over time and also suggesting an effect of presentation frequency as well as an overall difference between old and similar items.

We first performed an ANOVA on the d_L measures for old items. There was a main effect of lag (30, 90, 270, 500, 700, or 1000 ms), $F(5,50) = 42.83$, $MSe = .911$, $p < .001$, and a main effect of presentation frequency (1 or 3), $F(1,10) = 127.3$, $MSe = .176$, $p < .001$. In addition, there was an interaction between these factors, $F(5,50) = 16.33$, $MSe = .114$, $p < .001$. For old items presented once, responses at the 30- and 90-ms lags did not differ significantly from each other, but there was significant increase from 90 to 270 ms. There was also a significant increase from 270 to 500 ms, but the 500-, 700-, and 1000-ms cells did not differ significantly from each other. For the old items presented three times, there was a significant increase from 90 to 270 ms and a significant increase from 270 to 500 ms, but the 500-, 700-, and 1000-ms cells did not differ significantly from one another.

More critically, for the similar items, there was a main effect of lag, $F(5,50) = 25.47$, $MSe = .407$, $p < .001$, and a main effect of presentation frequency, $F(1,10) = 53.02$, $MSe = .109$, $p < .001$. There was also a marginally significant interaction between these factors, $F(5,50) = 2.33$, $MSe = .094$, $p < .06$. For similar items presented once, there was a significant increase from the 270-ms deadline to the 500-ms deadline. For the similar items presented three times, there were reliable increases from the 30-ms deadline to the 270-ms deadline and from the 270-ms deadline to the 500-ms deadline. Most importantly, there were no significant differences among the 500-, 700-, and 1000-ms cells for the pseudowords that were similar to stimuli shown either once or three times. The results are similar to those of Hintzman and Curran's experiments, in that there was no evidence that acceptance of similar items decreased late in processing.

We also conducted model-based analyses on the d_L data. Specifically, we fit the data in Fig. 4 with both monotonic and non-monotonic functions. As in our analyses of Hintzman and Curran's data, each function was applied to the d_L data for old and similar items. The fit of the monotonic model was quite good, (R^2 adjusted = .989, RSS = .456) and the fit of the non-monotonic model was not significantly better (R^2 adjusted = .986, RSS = .450; $\chi^2(3) = .318$, n.s.). Estimated parameter values for both models are shown in Table 3, and the predic-

[8] We also calculated the logistic bias (Snodgrass & Corwin, 1988) for both of our own experiments and both of Hintzman and Curran's (1994). For all four experiments, logistic bias for similar items compared to new items varied significantly as a function of response lag, with subjects becoming markedly more conservative at longer lags. In addition, for both of our experiments, the bias measure for old items versus new items showed significant differences between various lags.

TABLE 3

Best-Fitting Parameter Values for the Monotonic and Non-monotonic Models to the Data from Experiment 1

		Monotonic model			Non-monotonic model				
		Intercept	Asymptote	Rate	Intercept	Lag*	Asy. 1	Asy. 2	Rate
Old d_L	Freq. 1	496.9	2.62	263.7	496.5	—	2.57	—	241.2
	Freq. 3	496.9	4.27	228.1	496.5	—	4.24	—	222.2
Similar d_L	Freq. 1	496.9	1.46	189.6	496.5	525.8	1.41	1.41	207.6
	Freq. 3	496.9	1.99	103.8	496.5	525.8	1.92	1.36	108.9

Note. Freq., presentation frequency; Asy. 1, the earlier asymptote; Asy. 2, the later (lower) asymptote.

tions of the monotonic model are shown as the curves in Fig. 4.

Finally, we fit individual subjects' data with the monotonic and non-monotonic models. None of the 11 subjects provided data that were significantly better fit by the non-monotonic model than by the monotonic model. Looking at the summed improvement of fit for the non-monotonic model, it did not appear that the total improvement was significant across subjects, $\chi^2(33) = 1.230$, n.s. That is, none of the subjects individually showed evidence in favor of recall-to-reject processing, and there was not a significant improvement due for the non-monotonic model when information was aggregated across subjects.

Discussion

There was not much evidence for recall-to-reject found in this experiment. Although our subjects were well able to discriminate between studied and unstudied items and were increasingly able to reject completely new foils across time, they did not show any evidence of an increasing rejection rate for similar foils relative to completely new foils. Consistent with Hintzman and Curran (1994), the d_L analyses indicated a relatively flat propensity to accept similar items at longer lags (see Fig. 4), and there was no evidence for a drop at either the group or individual subject level.

This experiment did differ in another way from those of Hintzman and Curran. Whereas in Hintzman and Curran's study phase, only one word was presented from a pair of similar items (such as *frog* or *frogs*), we sometimes presented both items in the study list. More specifically, for 27% of our word pairs we presented both items (e.g., *PRUMIR* and *PRUMAD*) during study. We did so to provide a slightly more realistic situation that did not encourage specialized strategies on the part of the subject. That is, in recognition memory tasks outside of the laboratory, people are often faced with a situation where they have seen two items which are similar to each other. A study list that does not contain any similar pairs of items might lead subjects to use an idiosyncratic recall-based strategy which could not possibly be used in a more general situation. For example, if a subject knows that either *frog* or *frogs* was presented but not both, then recalling *frog* would indicate with certainty that *frogs* was not presented. However, the data we report for Experiment 1 are comparable to Hintzman and Curran's in that we analyzed only the judgments for word pairs for which just one word had been presented. Nonetheless, we compared our design directly to Hintzman and Curran's in our next experiment. Half of the subjects in Experiment 2 never studied more than one item from a similar pair; the remaining subjects studied both members for about half of the pseudoword pairs.

EXPERIMENT 2

This experiment had two main purposes, in addition to generally being another attempt to look for recall-to-reject processes. First, we examined the possible effects of list composition

on recognition strategy. If the second, recall-like process is optional, then it may be under strategic control and sensitive to the organization of the training set. In particular, in the paired condition, subjects often studied both members of a pair of similar items (for 36% of training stimuli, a similar item also appeared in the study list). In the unpaired condition, as in the Hintzman and Curran (1994) experiments, subjects never studied both items from a pair of similar items. If subjects' recognition strategies are sensitive to list composition, then a recall-to-reject process might be more likely in the unpaired condition than in the paired condition. The actual stimuli that were involved in pairings in the paired condition were only tested as old items. Hence the similar foils and true foils were comparable for the paired and unpaired conditions, and it was only the design for the old items that differed.

Also, this experiment collected a more extensive set of data per subject compared to previous experiments, with seven signal lags from 100 to 2000 ms, and a total of 13 testing sessions, facilitating analyses at the level of individual subjects.

Method

Subjects. Ten University of Warwick students participated in 13 sessions of approximately 1 h each and were paid 75 pounds sterling. Five subjects were chosen randomly for the paired condition and the remaining five were in the unpaired condition.

Stimuli. The subjects studied lists of pronounceable pseudowords, generated as in Experiment 1, and were asked to make recognition judgments on old items, new items that were similar to old items, completely new items, and filler items. Each study list consisted of two (untested) primacy items, 18 critical pseudowords shown once each, 18 critical pseudowords shown three times each, 6 filler items, and 2 (untested) recency items.

In the unpaired condition, all 36 critical pseudowords and the 6 fillers were unrelated to one another. Thus, subjects in this condition always studied only one member of a similar stimulus pair. Six repeated and six unrepeated critical stimuli were randomly chosen to serve as old items on the recognition test; the unstudied similar partner of 12 other pseudowords (6 repeated and 6 unrepeated) served as the similar foils on the test, and 12 completely new pseudowords served as the new foils. The 6 filler items (4 presented once each; 2 presented three times) appeared on both the study lists and recognition tests.

In the paired condition, subjects studied 6 repeated pseudowords and 6 unrepeated items that were unrelated to one another. The unstudied similar partners to each of these stimuli served as the similar foils on the recognition test. Thus, the similar test items were exactly comparable in the paired and unpaired conditions. In addition, subjects in the paired condition studied both members of 6 stimulus pairs that were shown once and both members of 6 stimulus pairs that were shown three times. One member of each of these 12 pairs was randomly selected to serve as a studied ("old") pseudoword on the recognition test; 12 additional pseudowords were selected to serve as completely new foils. There was also 6 fillers items (4 shown once each; 2 shown three times) on both the study and test lists.

Six critical stimuli of each type (frequency one or three; old, similar, new) were tested in each study–test block, and there were seven response signal lags (100, 200, 350, 500, 750, 1200, and 2000 ms) on each list. A Latin square design was used to assign items to response deadlines such that over a set of seven study-test blocks there were six items of each type tested at each lag. The filler items were assigned to lags such that there were always six test trials at each lag in each block. The presentation and test orders of the critical stimuli were randomized for each list for each subject.

Procedure. The procedure was the similar to that of Experiment 1, except that each of the sessions included a total of seven lists of 66 stimulus presentations, and each study list followed a test with 42 items. Also, responses were collected with a button box, with two buttons labeled YES and NO. As in Experiment 1, the first session was preceded by a training exercise for the response-signal task. To increase the

TABLE 4

Hit Rates (HR) and False-Alarm Rates (FA) for Experiment 2

	Response deadline (ms)						
Stimulus	100	200	350	500	750	1200	2000
Paired condition							
HR							
Old 1	.44	.49	.61	.64	.65	.66	.65
Old 3	.44	.50	.69	.77	.80	.83	.84
FA							
Similar 1	.41	.41	.44	.39	.33	.33	.32
Similar 3	.40	.47	.41	.47	.36	.37	.41
New	.36	.40	.34	.29	.24	.24	.22
Unpaired condition							
HR							
Old 1	.42	.43	.57	.61	.61	.60	.60
Old 3	.36	.51	.66	.76	.77	.73	.71
FA							
Similar 1	.41	.40	.42	.38	.37	.34	.33
Similar 3	.41	.44	.47	.42	.44	.39	.29
New	.38	.44	.38	.33	.27	.25	.21

yield of available data within the response range of 100 to 350 ms, subjects were instructed to try to respond within 300 ms of the signal tone.

Results and Discussion

Data from the first session for each subject were discarded. All analyses were based on only those responses that occurred between 100 and 350 ms after the response signal. About 5% of the data were excluded.[9]

We provide the hit and false alarm rates to old, similar, and new items in Table 4. Compared with our Experiment 1, and with Hintzman and Curran's (1994) Experiment 3, these subjects appear to have made their earliest responses with a more stringent response criterion: The positive response rate to all types of items is rather low at the earliest lags. However, these subjects clearly were able to discriminate studied from new items. The hit rate to studied items increased across processing, and the false-alarm rates to both similar and completely new items decrease over time, with false alarms to completely new items dropping more rapidly than those to similar foils. These subjects, like those in Hintzman and Curran's Experiment 2, show little evidence of a peak in their false alarms to similar items, except perhaps for the foils that were similar to items studied three times. Tukey's tests did not reveal any significant increases or decreases, at different lags, in false alarms for similar items.

The results of the d_L analyses for Experiment 2 are shown in Fig. 5. Figure 5A shows the data from the paired condition, Fig. 5B shows the data from the unpaired condition, and Fig. 5C shows the pooled data from all 10 subjects. The pattern of results is similar to that of the other experiments, showing a general increasing trend for old items and a rather flat curve for similar items. We would note that the d_L values for similar items are somewhat lower than those in our previous analyses. Although the false-alarm rates to the similar items were roughly comparable to the analogous data in Hintzman and Curran's Experiment 2, the false-alarm rates for completely new foils in this experiment did not

[9] These exclusions also included a very small number of occasions on which the computer recorded both response buttons being pressed, and, in the case of one subject, responses to eight items were dropped due to an error in a stimulus file.

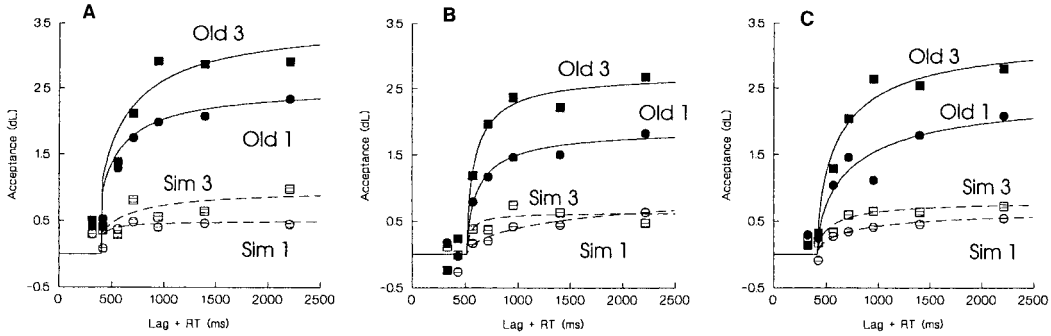

FIG. 5. Acceptance of stimuli (in terms of d_L) for Experiment 2, at varying response deadlines; (A) the data from the paired condition; (B) the data from the unpaired condition; and (C) the overall means.

decline as rapidly as in their study. As a result, our subjects showed relatively low similar–new discrimination (d_L) measures.

We first performed an ANOVA on the d_L measures for old items. Condition (paired or unpaired) was not a statistically significant variable and it did not significantly interact with other variables. There was a main effect of lag (100, 200, 350, 500, 750, 1200, and 2000 ms), $F(6,48) = 39.41$, $MSe = .43$, $p < .001$, and a main effect of presentation frequency (1 or 2), $F(1,8) = 8.98$, $MSe = .78$, $p < .05$. In addition, there was an interaction between these factors, $F(6,48) = 3.87$, $MSe = .20$, $p < .01$. Hence we compared mean responses at different lags, pooled over condition. For old items presented once, responses at the 100- and 200-ms lags did not differ significantly from each other, but there was a significant increase from these two to the 350-ms lag. The 350-, 500-, 750-, and 1200-ms lags did not differ significantly from each other, but there was a significant jump from 350 to 2000 ms. For the old items presented three times, responses at the 100- and 200-ms lags did not differ significantly from each other, but there was a significant increase from these two to the 350-ms condition. From the 350-ms condition there is a significant increase to the 500-ms condition, and finally the increase from the 750- to the 2000-ms condition is statistically significant.

More critically, for the similar items, condition (paired or unpaired) was not a statistically significant variable and it did not significantly interact with other variables. There was a main effect of lag, $F(6,48) = 5.07$, $MSe = .17$, $p < .001$, and a main effect of presentation frequency, $F(1,8) = 11.65$, $MSe = .08$, $p < .01$, but the interaction between these two variables was not significant, $F(6,48) = .75$, $MSe = .06$. Hence we looked at mean responses at different lags, pooled over the condition and presentation variables. There was clearly an increasing trend overall, but the only significant pairwise difference was the increase from the 200- to the 2000-ms deadline. There was no evidence for decreased acceptance of similar items at longer lags.

We also applied Eqs. 1 and 2 to the data from each condition and to the results pooled over the two conditions. For the paired condition, the monotonic model gave a good fit (R^2 adjusted = .970, RSS = 1.185) and the non-monotonic model did not fit significantly better (R^2 adjusted = .964, RSS = 1.185; $\chi^2(3) = 0.0$, n.s.). Indeed, the best fit for the non-monotonic model actually assumed that the second asymptote would the same as the first asymptote, i.e., there would be no drop. Requiring the second asymptote to be lower only made the non-monotonic model fit worse. The best-fitting parameter values for both models are shown in Table 5, and the predictions of the monotonic model are shown as the curves in Fig. 5A. For the unpaired condition, the fit of the monotonic model was again quite good (R^2 adjusted = .982, RSS = .462), and the non-monotonic model did not fit significantly better (R^2 adjusted =.980, RSS = .434, $\chi^2(3) = 1.751$, n.s.). The estimated parameter values for the both

TABLE 5

Best-Fitting Parameter Values for the Monotonic and Non-monotonic Models to the Data from Experiment 2

		Monotonic model			Non-monotonic model				
		Intercept	Asymptote	Rate	Intercept	Lag*	Asy. 1	Asy. 2	Rate
Paired condition									
Old d_L	Freq. 1	404.7	2.53	354.3	404.7	—	2.53	—	354.3
	Freq. 3	404.7	3.54	529.7	404.7	—	3.54	—	529.7
Similar d_L	Freq. 1	404.7	.50	118.3	404.7	404.7[a]	.50	.50	118.3
	Freq. 3	404.7	.98	510.7	404.7	404.7[a]	.98	.98	510.7
Unpaired condition									
Old d_L	Freq. 1	520.6	1.88	272.8	515.8	—	1.98	—	375.2
	Freq. 3	520.6	2.72	192.2	515.8	—	2.71	—	195.8
Similar d_L	Freq. 1	520.6	1.03	2756.0	515.8	716.0	.52	.39	1077.9
	Freq. 3	520.6	.63	85.9	515.8	716.0	.37	.28	.1
Pooled data									
Old d_L	Freq. 1	416.8	2.39	768.7	416.8	—	2.39	—	768.7
	Freq. 3	416.8	3.26	502.6	416.8	—	3.26	—	502.6
Similar d_L	Freq. 1	416.8	.66	806.3	416.8	416.8[a]	.66	.66	806.3
	Freq. 3	416.8	.80	328.2	416.8	416.8[a]	.80	.80	328.2

Note. Freq., presentation frequency; Asy. 1, the earlier asymptote; Asy. 2, the later (lower) asymptote.
[a] Multiple values of this parameter make the same predictions. Sample value is shown.

models are shown in Table 5, and the predicted responses are shown as the curves in Fig. 5B. In addition, we applied these models to pooled data from all 10 subjects. The fit of the monotonic model was very good (R^2 adjusted = .977, RSS = .692) and the non-monotonic model did not fit significantly better (R^2 adjusted = .973, RSS = .692, χ^2 (3) = 0.0, n.s.). The estimated parameter values are shown in Table 5 and as the curves in Fig. 5C. Again, the best-fitting version of the non-monotonic model made essentially the same predictions as the monotonic model, assuming that the two asymptotes would be identical.

Finally, we fit individual subjects' data with the two models. In this way, we assessed the evidence for recall-to-reject over a range of individual performance levels, such as varying levels of discrimination and bias for different subjects. However, we found that none of the 5 subjects in the unpaired condition showed statistically significant evidence for the use of a recall-to-reject process. Of course, this condition was expected to have the greatest chance of revealing the contribution of recall-to-reject. Similarly, we found that only 1 of the 5 subjects in the paired condition showed significant evidence for recall-to-reject. Putting together these analyses, the summed improvement due to the non-monotonic model for the 10 subjects was not statistically significant, $\chi^2(30) = 28.46$, n.s. Across these four experiments, then, we have found only 6 of the 73 subjects to have provided data that were significantly fit better by the non-monotonic model. There is a 10.4% probability of 6 or more significant results of 73 by chance alone.

In sum, the findings for Experiment 2 were similar to those of the other experiments. The corrected d_L analyses showed a significant increasing pattern of acceptance for similar items (as well as for old items), without any evidence for decreasing acceptance of similar items at longer lags. That is, the distinctive prediction for recall-to-reject was not supported at the group or individual subject level.

GENERAL DISCUSSION

Summary

It is always difficult to demonstrate that a null result indicates the true absence of an effect, but

we do believe that our findings are best characterized as a lack of evidence for the recall-to-reject account. This account makes a quite distinctive prediction, that similar items will be increasingly rejected as the recall process has more time to operate. If we assume that recall begins to contribute to item recognition later in processing, then the recall-to-reject account would predict that similar items would be increasingly rejected at longer lags. Instead, we found no evidence that similar words such as *frogs* were increasingly rejected relative to true foils such as *chair*. There was no evidence in our analyses that recalling the presented item, *frog*, led people to reject *frogs*. This finding was consistent over four experiments conducted in three laboratories, using two different kinds of stimulus pairs. Each of these experiments showed significant increases in acceptance of old items late in processing and generally had results associated with high degrees of statistical significance. Thus we would not attribute the lack of evidence for decreasing judgments on similar items to a lack of power in these four experiments.

However, there may indeed be other evidence which seems more favorable to a recall-to-reject account. Although we have argued that false-alarm rates should be interpreted with caution and that discrimination measures such as d_L are more suitable, some may still find Hintzman and Curran's (1992) pattern of false alarms to be suggestive. That is, sometimes false alarms to similar items increased early in processing and decreased late in processing. Even so, a supporter of the recall-to-reject account would still need to explain why a major, distinctive prediction of this account, that people will get better at distinguishing similar items such as *frogs* from completely new items such as *chair* after the recall process begins to contribute to the judgment, was not supported at all in the four experiments we analyzed. Our findings represent a major failure for the recall-to-reject account, regardless of any other evidence which may seem to support this account.

Out of fairness, though, we should point out that our own experiments and those of Hintzman and Curran have a restricted focus, and recall-to-reject processes might be found in other situations which we have not investigated. For example, associative recognition may include a recall-to-reject process because recalling that *A* had been studied with *B*, not with *B′*, would allow one to reject an *A–B′* foil quite readily, even if *B′* had also been studied (e.g., Clark, 1992; Clark, Hori, & Callan, 1993). More concretely, it seems possible that recalling that *quilt* had been studied with *telephone* could facilitate rejection of a test pair like *quilt–soda*, even if *soda* had also been studied in another pair. In a response-signal paradigm, Gronlund and Ratcliff (1989, Experiment 2) asked participants to discriminate pairs of words that had been studied together (i.e., *A–B*) from those that had been studied with other words (i.e., *A–B′*) and from those that were completely new foils (i.e., *X–Y* pairs, where *X* and *Y* are both new words). Although precise d_Ls could not be computed from their published data, false alarms to their rearranged pairs did appear to decrease more rapidly over time than false alarms to completely new foils, as if a recall-to-reject process were operating. Recently, Rotello and Heit (1998) have confirmed that subjects in associative recognition experiments similar to Gronlund and Ratcliff's (1989, Experiment 2) show evidence of using a recall-to-reject process, using the same criteria applied in the analyses in the present paper.

In the Introduction, we noted that our results would be interpreted within a two-process framework in which information from the first process is available early in processing and information from the second process is available late in processing. However, next we begin to interpret the results within a wider variety of possible accounts.

Other Two-Process Accounts

While still assuming that there are two memory processes which contribute to recognition judgments, there are a number of possibilities about their sequence. One plausible version of two-process theory is that the first, familiarity-based, process continues to make new information available even after information from the recall process becomes available. If the first

process were continuing late in processing, then this familiarity process would lead judgments on similar, unpresented items to increase. Indeed, there is some hint of increased acceptance of similar items late in processing in Fig. 4 for our Experiment 1.

In contrast, a recall-to-reject process would act in the opposite direction of the familiarity process, leading people to reject similar items. What would be the implications of *both* a familiarity process and a recall-to-reject process operating simultaneously? Our findings of no significant change in d_L for similar items, late in processing, could be due to a recall-to-reject process canceling out new information from an ongoing familiarity process. The decrease in response due to recall-to-reject could counteract the increase in response due to familiarity. We admit that we cannot rule out this alternative argument, but it does seem like rather weak evidence for a recall-to-reject process in item recognition. Under this explanation, the recall-to-reject process is so ineffectual that it does not actually help people to reject similar, unpresented items.

Assuming a two-process model of recognition in which the familiarity process operates earlier than the recall process, are there other variants of the recall component that might better account for these data? That is, there are many possible second processes that might be applied to recognition judgments, and recall-to-reject is just one of them. Although we are addressing the evidence for each process individually, we would not rule out *a priori* the possibility that more than one second process could be operating.

We consider two specific possibilities, the first of which we call the *recall-to-accept* account. In this account, the second process has a positive influence, in accepting old test items, rather than a negative influence, in rejecting new test items. When a subject studies *frog* and is tested on *frog*, a matching memory trace is recalled, thus increasing the probability that *frog* will be called "old." In contrast, if *frogs* is tested, normally no matching trace would be recalled, so *frogs* judgments would not obtain an increment in the positive response rate from the second process. The critical difference between this recall-to-accept account and the recall-to-reject account is that recall-to-accept would facilitate the correct acceptance of old words like *frog*, whereas recall-to-reject would facilitate the correct rejection of similar foils such as *frogs*. Hence, in terms of the predictions for discrimination performance, the recall-to-accept account of recognition would say that discrimination between old items and completely new foils would continue to increase late in process because recall-to-accept facilitates acceptance of studied items. This prediction is consistent with the data from all four of the experiments we have described.

Finally, we consider a different kind of second process, which would be an alternative to both recall-to-accept and recall-to-reject. In this account, which we refer to as *exhaustive search*, the second memory process searches through memory for an exact match, and the subject rejects the test item if no match is recalled. Thus, the subject would make a positive judgment on an old word such as *frog* and reject similar words such as *frogs* and completely new words such as *chair*. The critical difference between this account and the recall-to-reject account is that the recall-to-reject account depends on recalling the presented item, *frog*, in order to reject the similar item, *frogs*. In contrast, the exhaustive search process would simply reject *frogs* when no *frogs* memory traces are recalled, regardless of whether *frog* is recalled. We emphasize that this exhaustive search account is not the process described by Clark (1992), Clark and Gronlund (1996), Gronlund and Ratcliff (1989), or Hintzman and Curran (1994). These other accounts clearly describe a recall-to-reject process. However, the exhaustive search account is compatible with descriptions of a metamemory process in which the failure to retrieve a memory trace for a subjectively memorable event is diagnostic of its nonoccurrence (e.g., Brown, Lewis, & Monk, 1977; Strack & Bless, 1994; cf., Rotello, in press).

Our data do not allow us to discriminate between these two alternatives at present. Indeed, our main positive finding, that people

increasingly favor old items such as *frog* over completely new items such as *chair* late in processing, is compatible with both exhaustive search and recall-to-accept. That is, the result could be due to increased responses to *frog* (according to recall-to-accept) or to decreased responses to *chair* (according to exhaustive search).

One-Process Accounts

Our purpose so far has been to make a contribution to two-process theories of memory, by specifying variants in greater detail and deriving their predictions. We have not been trying to distinguish two-process accounts of memory, as a class, from one-process accounts. However, there is value in developing better one-process models as well. We turn not to a memory model but to a model of categorization (Lamberts, 1995) which describes how classification decisions are made at different time lags. The critical idea behind Lamberts's model is that the decision whether to place some object into a particular category depends on the object's similarity to the category representation, but similarity depends on different features at different points in processing (see also Goldstone, 1994; Goldstone & Medin, 1994). Features that are especially salient will affect classification early in the time course of judgment, whereas less salient features will become available later.

It is natural to suggest an analogous, one-process account for recognition memory. That is, recognition would depend on a single, familiarity-based process, but comparative information about different features would become available at different times. (Similarly, Ratcliff and McKoon, 1989, suggested that individual feature information would be available early, but information about feature relations would be available later.) Thus some test item might initially seem familiar on the basis of its salient features, but later on when more detailed information is available, the item might be a worse match to what is in memory. For example, consider the *frog–frogs* stimuli used by Hintzman and Curran (1994). It is plausible that the presence or absence of the final "s" would be less salient than other features. After all, the "s" is the last letter of the word and it does not change the meaning much (see also Murrell & Morton, 1974). Early in the time course of recognition, *frogs* might seem quite similar to a memorized item, *frog*, if information about the final "s" is not yet available. Later in the time course of recognition, with due consideration of the final "s," a *frogs* test item would not seem as familiar as a *frog* test item. The result would be that people would not be able to distinguish between *frog* and *frogs* at early test deadlines, but judgments on these two items would diverge at later deadlines. (See Brockdorff, 1998, for an implementation of these ideas.)

For our own Experiments 1 and 2, it is possible to apply the same sort of one-process account. Previous research on reading has suggested that consonants tend to be more salient than vowels (see Berent & Perfetti, 1995, for a review), so *PRUMIR* and *PRAMIR* might well seem similar early but not late in the time course of judgment. Thus for our own experiments as well as Hintzman and Curran (1994), the different pattern of results early and late in processing could be attributed to a single, familiarity-based process, but with different information or features being available at different times.

Conclusion

We conclude on a positive note. Building and testing models of memory is a difficult enterprise, because even well-established, successful models eventually face some unexplainable results. Further complicating the picture is the fact that it can be quite difficult to distinguish between broad classes of models, such as two-process versus one-process models. Still, sometimes quite consistent results appear which prove very useful in suggesting and constraining new models of memory. For example, one such regularity concerns the ratio of variances for recognition judgments on new items and old items; this ratio is typically lower than what is predicted by global memory models (Ratcliff et al., 1992). This finding has led to innovative developments in memory models (e.g., McClelland & Chappell, 1995; Shiffrin & Steyvers, 1997).

In the present paper, we have found initial

evidence for another regularity. For two different kinds of stimuli, we have found that, over the time course of judgment, people do not seem to get any better at distinguishing similar but unpresented items from new items. Previous results (Hintzman & Curran, 1994) provide an extremely clear pattern here. For example, the line representing judgments on similar items in Fig. 3 is almost perfectly flat. Our own results replicate Hintzman and Curran's in terms of showing increases for old items but no improvement on rejecting similar foils. Of course, more empirical work is needed before any statement about a regularity with greater generality can be made. Yet the present findings have so far been quite useful in showing that the recall-to-reject account of recognition memory is inappropriate for explaining these item recognition experiments.

REFERENCES

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). New York: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary development in mathematical psychology: Learning, memory, and thinking* (pp. 243–293). New York: Freeman.

Berent, I., & Perfetti, C. A. (1995). A rose is a REEZ: The two-cycles model of phonology assembly in reading English. *Psychological Review,* **102,** 146–184.

Borowiak, D. S. (1989). *Model discrimination for nonlinear regression models.* New York: Dekker.

Brockdorff, N. (1998). *A feature-sampling theory of classification and recognition.* Ph.D. thesis, University of Birmingham, UK.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology,* **29,** 461–473.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory and Cognition,* **20,** 231–243.

Clark, S. E., Hori, A., & Callan, D. E. (1993). Forced-choice associative recognition: Implications for global-memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **19,** 871–881.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of memory: How the models match the data. *Psychonomic Bulletin & Review,* **3,** 37–60.

Dosher, B. A. (1984). Discriminating preexperimental (semantic) information from learned (episodic) associations: A speed-accuracy study. *Cognitive Psychology,* **16,** 519–555.

Dosher, B. A., & Rosedale, G. (1991). Judgments of semantic and episodic relatedness: Common time-course and failure of segregation. *Journal of Memory and Language,* **30,** 125–160.

Estes, W. K. (1994). *Classification and cognition.* New York: Oxford Univ. Press.

Feenan, K., & Snodgrass, J. G. (1990). The effect of context on discrimination and bias in recognition memory for pictures and words. *Memory and Cognition,* **18,** 515–527.

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

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **20,** 3–27.

Goldstone, R. L., & Medin, D. L. (1994). Time-course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **20,** 29–50.

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

Heit, E. (1993). Modeling the effects of expectations on recognition memory. *Psychological Science,* **4,** 244–252.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **24,** 712–731.

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

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **20,** 275–289.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language,* **33,** 1–18.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **18,** 667–680.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language,* **30,** 513–541.

Jones, C. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **19,** 799–812.

Kortge, C. A. (1990). Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Meeting of the Cognitive Science Society* (pp. 764–771). Hillsdale, NJ: Erlbaum.

Lamberts, K. (1994). Towards a similarity-based account of compatibility effects. *Psychological Research,* **56,** 136–143.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General,* **124,** 161–180.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review,* **87,** 252–271.

McClelland, J. L., & Chappell, M. (1995). *Familiarity breeds differentiation: A Bayesian approach to the effects of experience in recognition memory*. Technical Report PDP.CNS.95.2, Carnegie Mellon Univ.

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

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review,* **95,** 183–237.

Mulligan, N., & Hirshman, E. (1995). Speed-accuracy trade-offs and the dual process model of recognition memory. *Journal of Memory and Language,* **34,** 1–18.

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

Murrell, G. A., & Morton, J. (1974). Word recognition and morphemic structure. *Journal of Experimental Psychology,* **102,** 963–968.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **14,** 700–708.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology,* **43,** 205–234.

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

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

Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology,* **21,** 139–155.

Rotello, C. M. (in press). Metacognition and memory for nonoccurrence. *Memory.*

Rotello, C. M., & Heit, E. (1998). Associative Recognition: Evidence for recall-to-reject processing. [manuscript submitted for publication]

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving efficiently from memory. *Psychonomic Bulletin and Review,* **4,** 145–166.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General,* **117,** 34–50.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language,* **33,** 203–217.

Wickelgren, W. A., & Corbett, A. T. (1977). Associative interference and retrieval dynamics in yes-no recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory,* **3,** 189–202.

(Received August 12, 1998)
(Revision received October 23, 1998)