Journal of Experimental Psychology:
Learning, Memory, and Cognition
1988, Vol. 14, No. 4, 700–708

Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality

Robert M. Nosofsky
Indiana University

Previously published sets of classification and old–new recognition memory data are reanalyzed within the framework of an exemplar-based generalization model. The key assumption in the model is that, whereas classification decisions are based on the similarity of a probe to exemplars of a target category relative to exemplars of contrast categories, recognition decisions are based on overall summed similarity of a probe to all exemplars. The summed-similarity decision rule is shown to be consistent with a wide variety of recognition memory data obtained in classification learning situations and may provide a unified approach to understanding relations between categorization and recognition.

Recently, there has been an upsurge of interest among categorization researchers in exploring relations between classification learning and old–new recognition memory. This interest has been fueled by the exemplar view of category representation, which holds that people base classification decisions on similarity comparisons with stored exemplars (Hintzman, 1986b; Medin & Schaffer, 1978; Nosofsky, 1986). Recognition data provide a source of converging evidence bearing on the nature of people's category representations. Presumably, if individual exemplars are being stored in memory, the fact ought to be revealed by postacquisition recognition tests.

Indeed, a number of researchers have taken exemplar models to task on grounds of certain dissociations between classification learning and recognition memory, or patterns of recognition data deemed to be inconsistent with the predictions of exemplar-only memory models. In virtually all cases, however, there has been a failure to specify and test an explicit *decision rule* by which exemplar memories are used to make recognition judgments.

A natural decision rule is the one embodied in the memory models of Gillund and Shiffrin (1984) and Hintzman (1986a), namely, that recognition judgments are based on the summed similarity (or activation) of a probe to all stored items. This summed similarity gives a measure of overall familiarity, with higher familiarity values leading to higher recognition probabilities. Medin (1986) recently considered the implications of a summed-similarity decision rule and suggested that it was at least roughly consistent with a set of classification/recognition data collected by Estes (1986b). The purpose of the present article is to follow up on Medin's suggestion and to illustrate more formally, by way of application to other examples in the literature, that a summed-similarity decision rule within the framework of an exemplar storage model may account well for recognition data obtained in classification learning situations and may help explain relations between classification learning and recognition memory. The use of a summed-similarity rule for interpreting typicality judgments is also explored.

General Modeling Approach

The analyses of the categorization and recognition data are conducted within the framework of the context model of classification proposed by Medin and Schaffer (1978) and generalized for application to continuous integral- and separable-dimension stimuli by Nosofsky (1986, 1987). According to the context model, the probability that Stimulus i is classified in Category J, $P(\mathrm{R_J}|\mathrm{S_i})$, is found by summing the similarity of Stimulus i to all Exemplars j belonging to Category J and then dividing by the summed similarity of Stimulus i to all exemplars of all categories,

$$P(\mathrm{R_J}|\mathrm{S_i}) = \frac{\sum_{j\epsilon C_J} N_j s_{ij}}{\sum_{K} (\sum_{k\epsilon C_K} N_k s_{ik})}, \qquad (1)$$

where N_j represents the relative frequency with which Exemplar j was presented during training and where s_{ij} represents the similarity between Exemplars i and j (Estes, 1986a; Nosofsky, 1988). Recognition judgments are assumed to be based on the overall familiarity of a stimulus, F_i, measured by summing the similarity of the stimulus to all exemplars of all categories,

$$F_i = \sum_{K} (\sum_{k\epsilon C_K} N_k s_{ik}). \qquad (2)$$

(Presumably, the subject sets some criterion c such that values of F_i greater than c lead to old responses.) Note that, whereas classification is assumed to be related to relative degree of target-category to contrast-category similarity (Equation 1), recognition is assumed to be related to overall summed sim-

The work reported in this article was supported by Grant BNS 85-19573 from the National Science Foundation to Indiana University.

I would like to thank Steven Clark, Douglas Medin, Janet Metcalfe, and Richard Shiffrin for many helpful discussions and three anonymous reviewers for their criticisms and suggestions regarding an earlier version of this article.

Correspondence concerning this article should be addressed to Robert M. Nosofsky, Department of Psychology, Indiana University, Bloomington, Indiana 47405.

ilarity, independent of category assignment (Equation 2).

To apply these decision rules, a method is needed for computing the s_{ij} similarity values in Equations 1 and 2. The method used throughout this article is based on the multidimensional scaling approach, which assumes that similarity is some decreasing function of distance in a psychological space. Specifically, each exemplar is represented as a point in an n-dimensional space, and the distance between Exemplars i and j is computed using the (Minkowski) power model formula,

$$d_{ij} = \left[\sum_{m=1}^{n} |x_{im} - x_{jm}|^r \right]^{1/r}, \tag{3}$$

where x_{im} is the psychological value of Exemplar i on Dimension m. In accordance with most previous work (e.g., Garner, 1974; Shepard, 1964), a "city-block" metric ($r = 1$ in Equation 3) is used for computing distances among highly analyzable, separable-dimension stimuli, and a Euclidean metric ($r = 2$) for computing distances among integral-dimension stimuli.

The distances d_{ij} are transformed to similarity measures using an exponential decay function, namely,

$$s_{ij} = e^{-d_{ij}}, \tag{4}$$

which appears to describe accurately the relation between similarity and psychological distance in classification learning situations (Shepard, 1958, 1986, 1987). As noted by Nosofsky (1984), the combination of a city-block metric and an exponential transformation yields an interdimensional multiplicative-similarity rule of the form proposed by Medin and Schaffer (1978) in their original formulation of the context model. The multiplicative rule has the property that interitem similarity will be high only if the items are similar on all component dimensions. Among other things, this aspect of the exemplar model enables it to be context sensitive. The idea is that a probe will tend to strongly activate only those items in memory for which there is a high degree of match between individual cues as well as contextual information stored in the memory representation (see Medin & Reynolds, 1985, and Medin & Schaffer, 1978, for more extended discussions).

Plan of the Article

The main goal in this article is to illustrate that certain qualitative patterns of recognition/classification data, which previous investigators have interpreted as providing evidence against exemplar-based models, are in fact consistent with the present approach. To maintain focus on the main issues, I apply only baseline versions of the model requiring a minimum of parameter estimation. Limitations in the ability of the baseline models to account for quantitative details of the data sets are noted, and possible extensions are considered. The quantitative tests should be interpreted with caution for two reasons. First, the model is applied to averaged data, and the parameters for individual subjects may be expected to vary. Second, the psychological dimensions along which the stimuli are organized may correspond only roughly to the physical specifications provided by the experimenters.

Undoubtedly, people can avail themselves of a number of alternative strategies in making recognition judgments, and the intent in this article is not to argue that the summed-similarity decision rule is the only one used. Indeed, Gillund and Shiffrin (1984) left open the possibility that, in addition to using overall familiarity as a basis for recognition, one may use search and retrieval strategies (e.g., Atkinson & Juola, 1974; Mandler, 1980; Tulving & Thomson, 1971). However, the global-familiarity rule may be prevalent in classification learning situations, where it is presumably difficult to gain unique access to individual memory representations of similar stimuli.

Application to Examples

Hayes-Roth and Hayes-Roth (1977): Correlations Between Classification and Recognition

Hayes-Roth and Hayes-Roth (1977) collected classification and recognition confidence ratings in a concept learning study that used rule-described categories. Subjects learned to classify descriptions of people into two clubs. The descriptions varied along three relevant dimensions (age, education, and marital status). There were four values per dimension (e.g., married, single, widowed, and divorced for marital status), which Hayes-Roth and Hayes-Roth labeled with the numbers 1–4. The rules governing category membership were summarized by Hayes-Roth and Hayes-Roth (1977) as follows:

> If the number of 1s (2s) exceeds the number of 2s (1s) in an exemplar and there are no 4s, the exemplar is in Club 1 (2); if the number of 1s equals the number of 2s and there are no 4s, the exemplar can be in either club, each with probability .5; if a 4 is present, the individual is in neither club. (The presence of one or more 3s had no implication; i.e., those feature values were irrelevant to Club 1–Club 2 discrimination.) (p. 326)

Hayes-Roth and Hayes-Roth varied the frequency with which individual exemplars were presented during classification training. The actual exemplars, their club assignments, and their frequencies of presentation are summarized in Hayes-Roth and Hayes-Roth (1977, Table 1).

Postacquisition classification and recognition confidence ratings collected by Hayes-Roth and Hayes-Roth (1977) are shown in Table 1. The higher the recognition rating, the more confident a subject was that the stimulus was old; the higher the classification rating, the more confident a subject was that the stimulus belonged to Club 2. A main result of interest was that the category prototypes (111 and 222), which were never presented during training, received the highest classification ratings, whereas certain high-frequency exemplars (e.g., 112) received the highest recognition ratings. In an analysis of Hayes-Roth and Hayes-Roth's data, Anderson, Kline, and Beasley (1979) commented:

> A difficulty for the Medin and Schaffer version of the store-instances-only model was the low correlation found by Hayes-Roth and Hayes-Roth between recognition and classification. They found that the prototypes received the highest classification ratings but the frequently presented nonprototypes received the

Table 1
Mean Z-Transformed Recognition and Classification Ratings for Individual Exemplars, With Context Model Predictions

Test exemplar	Recognition[a] rating	Summed similarity	Classification[b] rating	Predicted category 2 response probability
112[c]	3.27	14.891	−2.43	.220
121[c]	3.85	14.891	−2.46	.220
211[c]	3.09	14.891	−2.46	.220
113	−0.06	5.032	−2.57	.231
131	0.88	5.032	−2.44	.231
311	0.18	5.032	−2.44	.231
133	−3.45	3.723	−2.09	.315
313	−2.29	3.723	−2.09	.315
331	−2.10	3.723	−2.22	.315
221[c]	1.73	14.891	2.12	.780
212[c]	1.07	14.891	2.32	.780
122[c]	2.17	14.891	2.22	.780
223	−0.91	5.032	2.08	.769
232	0.01	5.032	1.97	.769
322	0.10	5.032	2.11	.769
233	−1.69	3.723	1.94	.685
323	−2.22	3.723	1.78	.685
332	−1.74	3.723	1.95	.685
132[c]	1.13	13.860	0.00	.500
321[c]	1.58	13.860	0.02	.500
213[c]	1.30	13,860	−0.09	.500
231	−0.61	4.258	0.03	.500
123	−1.23	4.258	−0.09	.500
312	−0.95	4.258	0.10	.500
111	0.49	5.474	−2.82	.135
222	1.50	5.474	2.39	.865
333	−4.19	1.546	1.78	.500
444	−0.92	1.242	1.32	.500

Note. Adapted from "Concept Learning and the Recognition and Classification of Exemplars" by B. Hayes-Roth and F. Hayes-Roth, *Journal of Verbal Learning and Verbal Behavior,* 1977, *16,* Table 2, p. 329. Copyright 1977 by Academic Press. Adapted by permission.
[a] Original scale: −5 = *new/most confident*, . . ., +5 = *old/most confident.* [b] Original scale: −5 = *Club 1/most confident*, . . ., +5 = *Club 2/most confident.* [c] High-frequency exemplars.

> highest recognition ratings. This suggests that information is acquired both about the instances and about their more abstract characteristics. (p. 314)

Context model analysis. Contrary to Anderson et al.'s (1979) assertion, however, a single-parameter version of the context model can account simultaneously for the classification and recognition data reported by Hayes-Roth and Hayes-Roth (1977). I assume for simplicity that the distance between Exemplars i and j mismatching on m dimensions is given by $d_{ij} = mD$. A computer search was conducted to find a single value of D that yielded good ordinal predictions for both the classification and recognition ratings. With $D = 2.12$, the Spearman rank-order correlation between the Category 2 response probabilities predicted by the context model and the observed Category 2 confidence ratings is .95 and the Spearman rank-order correlation between summed-similarity values (F_i in Equation 2) and observed recognition ratings is .94. The predicted values are shown with the observed values in Table 1. The model predicts correctly that the prototypes have the most extreme classification ratings and that the high-frequency exemplars have the highest recognition ratings. Although the overall summed similarity for the prototypes is not as large as for the high-frequency exemplars, the similarity of the prototypes to members of their own category relative to members of the contrast category is larger than for the high-frequency exemplars. Thus, the exemplar model accounts for the dissociation between classification and recognition ratings observed in Hayes-Roth and Hayes-Roth's study.

Limitations and extensions. The analyses of Hayes-Roth and Hayes-Roth's (1977) data assumed only an ordinal relation between predicted classification probabilities and observed ratings and between summed-similarity and recognition ratings. More quantitative analyses would require, for example, the specification of precise functional relations between classification probabilities and ratings, an issue outside the focus of the present article. An evident shortcoming of the single-parameter model, however, is that Exemplar 444 received a much higher recognition rating than predicted. In the category structure used by Hayes-Roth and Hayes-Roth, Feature 4 on any dimension was a sufficient feature in the sense that it signaled membership in neither club regardless of the values on the other dimensions. Possibly, a value-specific form of selective attention may have arisen in which subjects stored in memory only partial representations of exemplars containing Feature 4. One way to describe such a process in terms of the model would be to relax the assumption of the constant mismatch-parameter D. The role of selective attention in modifying similarity relations among exemplars has been discussed extensively in previous work (Medin & Edelson, 1988; Medin & Schaffer, 1978; Nosofsky, 1986).

Omohundro (1981): Recognition, Classification, and Category Size.

Omohundro (1981, Experiment 1) conducted a classification learning study using dot patterns (with connecting lines) that were constructed by distorting prototypes. Subjects learned to classify the distortions into three categories of size 4, 8, and 12. Following training, subjects were given a series of forced-choice recognition memory tests. On each trial subjects were presented with an old exemplar and two foils. The foils were equidistant from the old exemplar, but one was a low distortion of the category prototype and the other was a high distortion. Omohundro hypothesized that if a category prototype had been stored in memory during learning, subjects should tend to err on the low-distortion foils rather than on the high-distortion foils. Furthermore, if subjects become increasingly likely to store a prototype as category size increases (Homa, Sterling, & Trepel, 1981), ability to discriminate between the old exemplars and low-distortion foils should decrease with category size. Both predictions were confirmed.

However, these patterns are also predicted by exemplar-only memory models that assume that recognition judgments are based on overall summed similarity. A rough geometric analogy to Omohundro's (1981) conditions is shown in Figure 1. There are four exemplars positioned on the solid circle that have been generated from a central prototype. For each exemplar, there is an equidistant low-distortion foil and high-distortion foil. The low-distortion foils lie on the inner circle and the high-distortion foils on the outer circle. Although the foils are equidistant from their parent exemplar, it can be seen that the low-distortion foils are more similar overall to the remaining exemplars of the category than are the high-distortion foils. Thus, summed similarity would be greater for the low-distortion foils, which may explain Omohundro's main result.

Context model analysis. I conducted computer simulations to corroborate this interpretation. Three category prototypes were generated by selecting random numbers in the interval (0, *b*) for each of 10 dimensions. Exemplars were generated by adding a random number in the interval (−*w*, *w*) to each dimension of each prototype vector. (Note that whereas *b* primarily determines between-category dissimilarity, *w* primarily determines within-category dissimilarity.) There were 4, 8, and 12 exemplars generated for Categories A, B, and C, respectively. A low-distortion foil of each exemplar was generated by adjusting each dimension value a magnitude *adj* in the direction of the prototype value. A high-distortion foil was generated by adjusting each dimension value a magnitude *adj* in the opposite direction. Similarity between exemplars was computed using a Euclidean metric and exponential decay similarity function. The following decision rule was used to predict the probability that probe i would be selected from among probes i, j, and k as the old exemplar:

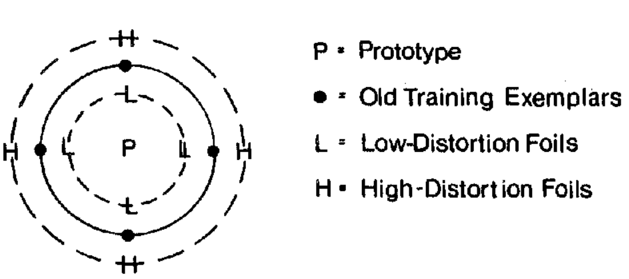

Figure 1. Geometric analogy to Omohundro's (1981) experimental conditions.

$$P\,(\mathrm{i}|\mathrm{i},\ \mathrm{j},\ \mathrm{k},) = \frac{F_\mathrm{i}}{F_\mathrm{i} + F_\mathrm{j} + F_\mathrm{k}}, \qquad (5)$$

where F_i is computed using the summed-similarity rule (Equation 2).

Computer simulations were conducted to find the values of *b*, *w*, and *adj* that minimized the chi-square fit between predicted and observed recognition responses. To introduce additional constraints, the model was also used to simultaneously predict correct classification responses for the old exemplars as a function of category size. The predicted and observed recognition and classification probabilities are reported in Table 2. The best fitting parameters were $b = 2.8$, $w = 1.2$, and $adj = .4$. The resulting chi-square of 3.65 ($df = 6$, $N = 864$) is remarkably small and not sufficient to reject even the present baseline model ($p > .50$).

The exemplar model predicts correctly that low-distortion foils will be called "old" substantially more than high-distortion foils and also that old–new discrimination will decrease as category size increases. Apparently, low-distortion foils tend to be at least as similar to other exemplars as individual exemplars are to one another (e.g., see Figure 1). With increases in category size, the relative contribution that an exemplar's self-similarity makes to overall summed similarity tends to be diluted, so old–new discrimination falters. The exemplar model also predicts the commonly reported finding of increased classification accuracy with increases in category size (as noted previously by Busemeyer, Dewey, & Medin, 1984, and Hintzman, 1986b). Most impressive, the model predicts the effects for classification and recognition simultaneously in quantitative detail.

Limitations and extensions. The model fits shown in Table 2 are for the delayed condition reported by Omohundro (1981), in which there was a 1-week interval between initial classification learning and subsequent testing. Omohundro also reported data from an immediate testing condition. The pattern of data was the same as in the delayed condition, but the magnitude of the category size effect for recognition was more extreme. Subjects recognized old exemplars from the Size 4 category, with probability approximately .7, and from the Size 8 and Size 12 categories, with probability approximately .4. To begin to predict an effect of this magnitude in terms of the summed-similarity decision rule required boosting the value of the between-category dissimilarity parameter *b*. The large value of *b*, however, then led to predictions of nearly perfect classification, which was not observed by Omohundro. Possibly, subjects were able to make use of a search and retrieval strategy in recognizing exemplars from the Size

Table 2
Predicted and Observed Recognition and Classification Probabilities as a Function of Category Size in Omohundro's (1981) Experiment 1 (Delayed Condition)

	Category size		
Probe	4 ($N = 72$)	8 ($N = 144$)	12 ($N = 216$)
	Old recognition probability		
Old			
Observed	.58	.46	.43
Predicted	.51	.47	.45
Low foil			
Observed	.28	.36	.39
Predicted	.32	.36	.38
High foil			
Observed	.14	.18	.18
Predicted	.17	.17	.16
	Correct classification probability		
Old			
Observed	.85	.87	.90
Predicted	.81	.86	.90

Note. N = number of observations on which each probability is based (number of probes per category [4, 8, or 12] times number of subjects [18]). Observed values in each cell were estimated from Omohundro's (1981) Figure 2. Predicted values were based on 1,000 simulated experiments.

4 category in the immediate condition. With the addition of a single retrieval parameter, the exemplar model is able to achieve accurate quantitative predictions of recognition and classification in Omohundro's immediate condition.

Metcalfe and Fisher (1986): Classification and Recognition Contingencies

Metcalfe and Fisher (1986) conducted classification learning and recognition memory experiments using dot patterns constructed from prototypes. The central hypothesis advanced in their study was that classification and recognition judgments may be mediated by separate memory systems, an implicit or semantic system for classification, and an explicit or episodic system for recognition. They reported experimental and statistical dissociations between classification and recognition that they argued were problematic for single-system exemplar memory models.

In an initial study phase (Experiment 2), subjects were presented with three lists of dot patterns (A, B, and C), with each list containing six patterns that were small distortions of a category prototype. Following study, subjects were tested in a transfer phase that included presentation of the category prototype, old small distortions, new small distortions, and large distortions. The main experimental manipulation was that one group of subjects was given classification learning instructions prior to study, whereas a second group was given recognition memory instructions. Subjects given classification instructions performed better on classification transfer tests than subjects given recognition instructions; however, prior instructions had no effect on recognition performance. These results seem consistent with exemplar memory models. For subjects given recognition instructions, there is no reason to store associated category labels with the presented exemplars, so it is not surprising that subjects performed relatively poorly on the subsequent classification test. By contrast, if subjects store individual exemplars in memory during classification learning, then it is not surprising that recognition performance was just as good for the group that was given classification instructions as for the group that was given recognition instructions.

The main focus of Metcalfe and Fisher's (1986) research was to study contingency relations between classification and recognition. They argued:

> Models that propose that classification is based on memory for specific instances suggest that there should be a relation between recognition of items as old and classification of the items, since both judgments presumably use the same information. . . . Subjects should be better at classifying items they think are old than those they think are new, because on average the "believe old" items are more likely to be in or like items in memory than the "believe new" items. (Metcalfe & Fisher, 1986, p. 164)

To test this idea, Metcalfe and Fisher (1986, Experiment 2) calculated difference scores between the probability of correct classifications given old recognition responses and probability of correct classifications given new recognition responses, $P(\text{correct}|\text{respond old}) - P(\text{correct}|\text{respond new})$, separately for old and new exemplars. Although these difference scores turned out to be slightly positive, they generally did not reach statistical significance, so Metcalfe and Fisher argued that the results were problematic for single-system exemplar memory models.

Medin (1986) recently questioned Metcalfe and Fisher's (1986) claim about the predictions of exemplar models and suggested that any of a wide variety of contingency relations may be observed, depending on experimental conditions. For example, consider a probe that is highly similar to exemplars from two categories. Because its overall summed similarity would be large, there would be a high probability of calling the probe "old"; however, because both categories would be competing for the probe, classification accuracy would be relatively low. Likewise, consider a probe that is located in an isolated portion of the stimulus space but is still far more similar (in a relative sense) to the exemplars of the target category than to the exemplars of the contrast category. In this case, classification accuracy would be high, but recognition probability would be low.

Context model analysis. Although conditions can be arranged to reverse the classification/recognition contingency hypothesized by Metcalfe and Fisher (1986), the question remains as to whether or not such a contingency would be expected under their particular experimental conditions. To explore this issue, I conducted computer simulations intended to mimic their experiment. Three category prototypes were generated by selecting random numbers in the interval $(0, b)$ for each of 10 dimensions. Small distortions were generated by adding random numbers in the interval $(-w, w)$ to each individual dimension of each prototype vector. Large distortions were generated by adding random numbers in the interval $(-L, L)$ to each dimension of each prototype vector.

Similarity between exemplars was computed using a Euclidean distance metric and exponential decay similarity function. For simplicity, it was assumed that recognition probability for a probe was linearly related to its familiarity, $P(R_i) = aF_i + c$.

As a preliminary step, simulations were conducted to find values of the model parameters that yielded good quantitative fits to the overall levels of recognition and classification performance observed in Metcalfe and Fisher's (1986) study. Next, 10,000 simulated experiments were conducted. On each simulation, theoretical classification and recognition probabilities were computed separately for an old probe and a new probe, and classification and recognition responses were simulated in accordance with these probabilities. In particular, two random numbers (r_1 and r_2) in the interval (0, 1) were selected. Let $P(C)$ and $P(R)$ denote the theoretical predictions of correct classification and of old recognition responses, respectively. A correct classification response was selected if $r_1 < P(C)$, and an old recognition response was selected if $r_2 < P(R)$.

The resulting contingency matrixes revealed no relation between correct classification and old recognition responses. The values of $P(\text{correct}|\text{respond old}) - P(\text{correct}|\text{respond new})$ were −.004 for old probes and −.013 for new probes. Apparently, exemplar memory models do not necessarily predict strong positive contingencies between classification and recognition in the usual prototype-distortion paradigms.

Limitations and extensions. The overall levels of predicted and observed classification and recognition probabilities are shown in Table 3. The best fitting parameters were $b = 1.45$, $w = 0.50$, $L = 1.05$, $a = .16$, and $c = 0$, with a resulting chi-square of 13.8 ($df = 3$, $N = 4{,}116$, $p < .01$). The predictions are in the ballpark of the observed data, but the model is rejected quantitatively. Its main shortcoming is that it fails to predict the magnitude of the prototype enhancement effect. (Note, however, that the exemplar model predicts correctly that the nonpresented prototypes have the highest recognition probabilities among the four types of probes.) One possible line of extension is to introduce assumptions about memory distortion into the modeling (e.g., Hintzman, 1986b). Presumably, people's memory representations of the exemplars are not veridical, but subject to effects of random noise. Although any given exemplar is highly similar to itself, the prototype tends to be at least fairly similar to many items. With increases in memory noise, the redundancy afforded the prototype should give it an advantage. Another problem that needs investigation is to explain why Metcalfe and Fisher (1986) actually observed a statistically significant positive contingency between correct classification and correct recognition for old probes in their recognition instructions condition. A number of hidden variables could produce such a contingency (e.g., subject and trial selection effects), but why is it observed in some experimental conditions and not in others?

Bourne (1982): Classification and Typicality

The final reanalysis in this article considers relations between classification and typicality judgments rather than old–new recognition. However, the theme parallels the one established previously for recognition. I reanalyze an illustrative data set published by Bourne (1982) and suggest that under his experimental conditions, the degree to which a probe was judged as being typical of a target category may have been based on the summed similarity of the probe to all exemplars of the target category. Classification, on the other hand, was based on relative degree of target-category to contrast-category similarity.

Table 3
Predicted and Observed Recognition and Classification Probabilities for Metcalfe and Fisher's (1986) Experiment 2 (Classification Instructions Condition)

Probe	"Old" recognition	Correct classification	*N*
Prototype			
Observed	.73	.65	147
Predicted	.66	.60	
Old distortion			
Observed	.63	.62	882
Predicted	.62	.64	
New small distortion			
Observed	.49	.57	882
Predicted	.51	.56	
New large distortion			
Observed	.28	.39	147
Predicted	.27	.49	

Note. N = number of observations on which each probability is based (number of subjects [49] times number of probes per category [1 or 6] times number of categories [3]). Observed values in each cell were estimated from Metcalfe and Fisher's (1986) Figure 2. Predicted values were based on 1,000 simulated experiments.

In Bourne's (1982) study, subjects learned to classify stimuli as either positive or negative instances of a logically defined concept. The stimuli were geometric forms varying along four dimensions (shape, color, size, and number) with three values per dimension. The concept was defined in terms of single values on two relevant dimensions, and subjects were informed which dimensions were relevant. With x denoting the critical value on Dimension 1 and y the critical value on Dimension 2, then $x\bar{y}$ and $\bar{x}y$ (stimuli with exactly one critical value) were always positive instances of the concept, $\bar{x}\bar{y}$ (stimuli with neither critical value) were always negative instances of the concept, and xy (stimuli with both critical values) were sometimes positive and sometimes negative. Across five conditions, the probabilities with which xys were assigned as positive instances were .0, .25, .50, .75, and 1.00. Given some additional experimental constraints described by Bourne (1982, p. 5), the basic category structures across the five conditions can be schematized as shown in Table 4. In the table, a value of 1 on either dimension is a critical value, whereas values of 2 or 3 denote the other values on the relevant dimensions.

Following concept-identification training, subjects were given various postacquisition tests. The tests I consider here are the pairwise typicality comparisons and the speeded classifications. In the typicality comparisons, subjects were given pairs of stimuli and were asked to choose the better example of the concept in each pair. The data obtained by Bourne (1982) in this test are reported in Table 5. Each row of the

Table 4

Relative Presentation Frequencies for Exemplars Used in Bourne's (1982) Concept-Learning Experiment in Each Condition (p)

Exemplar	Positive concept	Negative concept
12	2	0
13	2	0
21	2	0
31	2	0
22	0	1
23	0	1
32	0	1
33	0	1
11	$4p$	$4 - 4p$

Note. Conditions are denoted by p, the probability with which "11" exemplars were assigned as members of the positive concept. Across five conditions, the values of p were .00, .25, .50, .75, and 1.00. So, for example, in the $p = .25$ condition, "11" exemplars occurred with relative frequency equal to 1 in the positive concept and with relative frequency equal to 3 in the negative concept.

table gives the probability with which the first member in each pair was chosen as a better example than the second member. For example, xy was chosen as a better example than $x\bar{y}$ (or $\bar{x}y$) with probability .27 in the $p = .0$ condition, .41 in the $p = .25$ condition, and so forth. There is an important crossover effect in row 1 of the typicality pair-comparison matrix. For values of $p < .5$, $x\bar{y}$ and $\bar{x}y$ are judged more typical than xy, but the reverse is observed for values of $p \geq .5$. Bourne reported that in the final two blocks of concept acquisition, $x\bar{y}$ and $\bar{x}y$ were classified as members of the positive concept with nearly 100% accuracy in all conditions, whereas in the $p = .50$ and $p = .75$ conditions, xy were classified as members of the positive concept with probabilities of .59 and .79, respectively. Note, therefore, that there were conditions in which xy were judged as more typical of the positive concept than were $x\bar{y}$ and $\bar{x}y$ yet were classified by subjects into the positive concept with lower probability than were $x\bar{y}$ and $\bar{x}y$. The actual probabilities with which xy were classified as members of the positive concept in a post-acquisition speeded classification test are shown in Table 6.

Table 5

Predicted and Observed Probabilities With Which the First Member of Each Pair Was Chosen as the Better Example of the Concept in Bourne's (1982) Experiment

	Condition					
Pair	$p = .00$	$p = .25$	$p = .50$	$p = .75$	$p = 1.00$	N
$xy - \bar{x}y(x\bar{y})$						
Observed	.27	.41	.53	.71	.89	
Predicted	.21	.38	.58	.77	.88	192
$xy - \bar{x}\bar{y}$						
Observed	.74	.87	.97	1.00	1.00	
Predicted	.65	.84	.94	.98	.99	96
$x\bar{y} - \bar{x}y$						
Observed	.51	.55	.51	.46	.48	
Predicted	.50	.50	.50	.50	.50	96
$x\bar{y}(\bar{x}y) - \bar{x}\bar{y}$						
Observed	.93	.95	.99	.96	.89	
Predicted	.87	.90	.92	.93	.94	192

Note. N = number of observations on which each probability is based (number of subjects per condition [24] times number of probe-pair tests per subject [4 or 8]).

Context model analysis. I interpret the typicality and classification data in terms of the context model as follows. The *typicality strength* for a given probe is found by summing the similarity of the probe to all exemplars of the target category. Taking 11, 12, and 22 as representative instances of xy, $x\bar{y}$, and $\bar{x}\bar{y}$, respectively, then the typicality strength for each probe in each p condition is given by

$$T_p(11) = 8exp(-D) + 4p \tag{6}$$

$$\mathrm{T}_p(12) = 2 + 2\exp(-D') + 4\exp(-2D) + 4p\exp(-D) \tag{7}$$

$$\mathrm{T}_p(22) = 4\exp(-D) + 4\exp(-D - D') + 4p\exp(-2D). \tag{8}$$

These equations assume that mismatches between exemplars on critical features (i.e., 1–2 and 1–3) contribute distance D, and mismatches on noncritical features (i.e., 2–3) contribute distance D'. The probability that probe a is chosen as a better example of the concept than probe b in each p condition is then given by the logistic transformation,

$$P_p(a, b) = \frac{1}{1 + \exp\{-c[\mathrm{T}_p(a) - \mathrm{T}_p(b)]\}}, \tag{9}$$

where c is a freely estimated scale parameter. The minimum chi-square parameters were $D = 1.344$, $D' = 0.702$, and $c = 1.14$. The predicted probabilities are shown with the observed probabilities in Table 5. Although the model is rejected quantitatively, $\chi^2(17, N = 2{,}880) = 60.6$, $p < .01$, by conventional criteria it performs remarkably well. The exemplar model accounts for 96.1% of the variance in the pair-comparison matrix and captures some important qualitative trends, most notably, the crossover effect in row 1. The fit of the exemplar model (Equation 1) to Bourne's (1982) classification data is shown in Table 6, and it is impressive, $\chi^2(3, N = 480) = 1.27$, $p > .50$. (The chi-square fit was calculated using the data from only the $p = .00$ to $p = .75$ conditions, because expected and observed frequencies for incorrect responses in the $p = 1.0$ condition were less than five. There is a corresponding loss of one degree of freedom from the data.) The fit to the classification data required estimation of only a single distance parameter ($D = 2.868$). In summary, the typicality and classification data reported by Bourne are well characterized by assuming that typicality judgments are governed by summed similarity of a probe to all exemplars of the target category and that classification judgments are governed by relative degree of target-category to contrast-category similarity.

Limitations and extensions. The main quantitative shortcoming of the model as applied to the typicality data is that it underestimates preference for $x\bar{y}$ versus $\bar{x}\bar{y}$ in the $p = .50$ condition. I offer no speculation about the basis of subjects' extreme preference for $x\bar{y}$ in this condition. Perhaps more significant, estimates of the distance parameter D are discrepant across the typicality and classification conditions. On the other hand, stimulus conditions were not invariant, so parameter invariance might not be expected.

Table 6
Predicted and Observed Probabilities With Which xy Instances Were Called Positive in Bourne's (1982) Speeded Classification Test

Classification probability	Condition $p = .00$	$p = .25$	$p = .50$	$p = .75$	$p = 1.00$	N
Observed	.10	.31	.57	.81	.97	120
Predicted	.10	.33	.55	.77	1.00	

Note. N = number of observations on which each probability is based (number of subjects per condition [24] times number of *xy* tests [5]).

Summary and Discussion

The main purpose of this article was to illustrate an exemplar-based approach to interpreting old–new recognition memory data obtained in classification learning situations. The key assumption of the approach is that, whereas classification is determined by relative degree of target-category to contrast-category similarity, recognition may be determined by overall summed similarity of a probe to all exemplars stored in memory. Thus, classification and recognition may often be based on common representational substrates, but different decision rules may underlie performance in each task. The model was shown to be consistent with the following qualitative patterns reported in the literature: (a) low correlations between recognition and classification, (b) lack of positive contingencies between correct classification and old recognition responses, (c) high false-alarm rates for nonpresented prototypes and foils that are low distortions of a prototype, (d) faltering old–new discrimination with increases in category size, and (e) dissociations between classification and typicality judgments. These demonstrations are important because previous investigators have interpreted the patterns as providing evidence against exemplar-only memory models. Quantitative tests reported in this article revealed some striking success in using the model to account for detailed relations between classification and recognition but also some limitations that point out directions for future work and extensions.

A question that may arise concerns the utility of the summed-similarity exemplar approach vis-à-vis central-tendency prototype models (as formalized, e.g., by Reed, 1972). In the prototype model, classification of a probe is based on its similarity to the central tendency of the category exemplars. The central tendency is computed, of course, by summing information over the category exemplars. However, the summed-similarity exemplar model is not simply a disguised prototype model. For example, when exemplar information is summed to form a prototype, information is lost concerning correlated values along individual component dimensions (e.g., Ashby & Gott, 1988; Medin & Schaffer, 1978; Nosofsky, 1986). This point is clearly illustrated in the Hayes-Roth and Hayes-Roth (1977) data set considered earlier in this article. People had higher recognition confidence for high-frequency exemplars such as 112, 121, and 211 than for the nonpresented prototype (111), a trend that was correctly predicted by the exemplar model. But had subjects stored only a prototype, recognition confidence should have been highest for the prototype. Because the relation between similarity and psychological distance is highly nonlinear (Medin & Schaffer, 1978; Nosofsky, 1984; Shepard, 1958, 1987), computing the summed similarity of a probe to individual exemplars can lead to dramatically different predictions of classification and recognition than computing the similarity between a probe and the category central tendency.

References

Anderson, J. R., Kline, P. J., & Beasley, C. M. (1979). A general learning theory and its application to schema abstraction. In G. H. Bower (Ed.), *The psychology of learning and motivation.* New York: Academic Press.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

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 developments in mathematical psychology, Vol. 1: Learning, memory, and thinking* (pp. 243–293). San Francisco: Freeman.

Bourne, L. (1982). Typicality effects in logically defined categories. *Memory & Cognition, 10,* 3–9.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 638–648.

Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 115,* 155–174.

Garner, W. R. (1974). *The processing of information and structure.* New York: Wiley.

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

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16,* 321–328.

Hintzman, D. L. (1986a). *Judgments of frequency and recognition memory in a multiple-trace memory model* (Tech. Rep. No. 86-11). Eugene: University of Oregon.

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

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418–439.

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

Medin, D. L. (1986). Comment on "Memory storage and retrieval processes in category learning." *Journal of Experimental Psychology: General, 115,* 373–381.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

Medin, D. L., & Reynolds, T. J. (1985). Cue-context interactions in discrimination, categorization, and memory. In P. D. Balsam & A. Tomie (Eds.), *Context and learning* (pp. 323–356). Hillsdale, NJ: Erlbaum.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Metcalfe, J., & Fisher, R. P. (1986). The relation between recognition memory and classification learning. *Memory & Cognition, 14,* 164–173.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Omohundro, J. (1981). Recognition vs. classification of ill-defined category exemplars. *Memory & Cognition, 9,* 324–331.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review, 65,* 242–256.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General, 115,* 58–61.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory. *Journal of Experimental Psychology, 87,* 352–373.

Received January 27, 1986
Revision received October 7, 1987
Accepted October 21, 1987 ■