Journal of Experimental Psychology:
Learning, Memory, and Cognition
1988, Vol. 14, No. 1, 3-11

0278-7393/88/$00.75

A New Method for Investigating Prototype Learning

Jerome R. Busemeyer and In Jae Myung
Purdue University

Past researchers studied prototype learning by asking subjects to categorize exemplars constructed from different prototypes. This procedure is less than ideal because learning must be inferred from the percentage of correct categorizations pooled across many trials or subjects or both. An alternative procedure is proposed in which subjects are asked to reproduce their estimate of the prototype on each trial, thereby providing trial-by-trial information about changes in the estimated prototype. This procedure provides straightforward tests of three basic properties implied by several prototype learning models: additivity across exemplars, noninterference among features, and time invariance of serial position effects. An experiment is reported and the results provide reasonably good support for the properties of additivity and noninterference, but clear violations of time invariance were observed. The implications of the results for distributed-memory models and multiple-trace models of prototype learning are discussed.

It seems quite easy to produce an image of an ideal circle despite the fact that our experience is based on thousands of different imperfect examples. This natural ability to abstract and reproduce a single image from a myriad of examples is often referred to as prototype learning.

The purpose of this article is to describe a paradigm for investigating prototype learning. Subjects are shown a sequence of exemplars generated from one or more prototypes. After observing each exemplar, they are asked to reproduce (graphically or numerically) their current estimate of each prototype. Obviously, this procedure is limited to stimuli that can be easily reproduced by the subject.

It may be useful to compare the prototype production task with the categorization task introduced by Posner and Keele (1968, 1970). Exactly the same exemplars can be used during training in both tasks. In the categorization task, subjects are presented with an exemplar and then are asked to produce a category label. In the prototype production task, subjects are presented a category label and then are asked to produce a prototype estimate.

The *storage* of exemplar information may be quite similar in the two tasks (e.g., a multiple-trace memory system, or a composite distributed-memory system). However, the *use* of this stored information is quite different: The prototype production task requires some sort of abstraction procedure (e.g., form an average of the traces associated with a category), whereas the categorization task requires some sort of classification procedure (e.g., choose the category associated with a trace that is most similar to the probe).

There are several reasons for investigating the prototype production task. First, it is a naturally occurring task. Prototypic drawings of organs, bones, and cell structures frequently appear in physiological and medical textbooks. A second example is the use of prototypic symptom patterns to describe the behavior of patients suffering from different types of psychoses. Another example is stereotypic personality trait descriptions of minority groups or working classes.

Second, the prototype production task provides a direct view of the trial-by-trial evolution of a prototype separately for each subject. In the categorization task, learning must be inferred from the percentage of correct decisions pooled across subjects or blocks of trials. It is not even clear that category decisions are based on prototypes, and instead they may be based solely on memory for past exemplars (see Busemeyer, Dewey, & Medin, 1984, for a recent discussion of the problem of distinguishing exemplar and prototype models of classification).

Third, the prototype production task provides direct tests of some basic properties common to parallel distributed memory models (Knapp & Anderson, 1984; McClelland & Rumelhart, 1985), holographic memory models (Metcalfe-Eich, 1982), and multiple-trace memory models (Hintzman, 1986). Three basic properties common to these models will be empirically tested: additivity, noninterference, and time invariance. Because it is easier to understand these basic properties with a concrete example in mind, they will be described after presenting the following prototype production experiment.

Method

Procedure

The stimuli were constructed from random dot patterns similar to that used by past researchers (e.g., Homa & Cultice, 1984). Obviously, dot stimuli are less representative of natural stimuli than are wings of insects or personality traits. Also, random placement of dots may discourage subjects from using the complex features that they normally use with natural stimuli. However, the learning principles being tested are assumed to hold across stimulus sets, including dot stimuli. Also, the simple features that random dot stimuli tend to encourage permit more rigorous tests of these principles. Because one purpose of the present experiment was to test some basic properties of memory models, highly artificial dot stimuli were used in the attempt to obtain a simple feature representation.

We thank Douglas Medin, Richard Heath, and the reviewers for many helpful comments on an earlier draft.

Correspondence concerning this article should be addressed to Jerome R. Busemeyer, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

Subjects were asked to imagine that each dot pattern was a telescopic photograph of a star pattern containing four stars located far out in space. They were also told that because atmospheric disturbances distorted each photo, five photographs were obtained from each star pattern. The subjects' task was to identify the true star locations on the basis of the sample of five distorted photos. They were instructed to observe each photo sequentially and to provide a drawing of the estimated star locations after each photo. The subjects were also told that different star patterns were photographed by randomly changing the position of the telescope and focusing on a different region of space.

A typical photo is shown in Figure 1. (The actual photos were a little more than twice as large). Each photo contained four dots located within a bounded 15×15 cm^2 plane. Horizontal and vertical axes were drawn, and for each axis, tic marks with numerical labels were placed at one unit intervals ranging from -15 to $+15$ units. Each unit equaled 0.5 cm. Subjects drew their estimates of the prototype on a plane exactly like that used for the photos except the dots were initially absent.

Each subject received one practice and eight experimental sessions. A total of 24 star patterns were presented each session, producing a total of $5 \times 24 = 120$ photos per session. Each photo was preceded by a unique category label (a random number) that was common to all photos generated from the same star pattern. Also, the end of a sequence of five photos for a given star pattern was clearly distinguished by asking subjects to rate task difficulty after finishing every fifth photo.

The 120 photos per session were printed in a booklet. The booklet contained two pages for each photo. The photo was printed (by a computer) on the top page, and the estimated star pattern was drawn (by the subject) on the bottom page. Each photo contained four integers (1,2,3,4) positioned in the plane; each number represented a distorted photographic image of a corresponding star in the true star pattern. For each photo, subjects first connected the four points printed in the top page in numeric order. They then positioned four new numbers in a plane on the bottom page representing the estimated locations of the four stars and they connected these four points in numeric order. After completing each pair of pages for a given photo, both pages were turned over and permanently hidden from view.

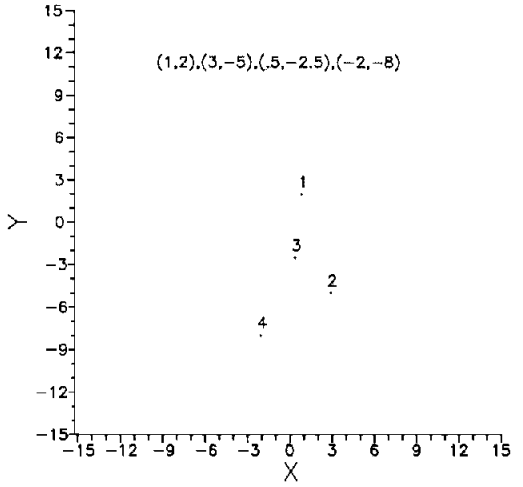

Figure 1. An example photo of a star pattern. Each number represents a distorted image of one of the true stars in the pattern.

Each session lasted approximately 45 minutes, and each subject was run individually in a quiet room. Subjects were told that their pay at the end of the experiment depended on their performance, where performance was measured by the sum of squared deviations from the estimated and true location of each star. The photos were scored by projecting each dot drawn by the subject onto each axis, and recording the position in 0.5-cm units.

Stimulus Design

Each photo can be represented as an eight-element vector, denoted $\mathbf{E}$. The first pair of elements, denoted X_1 and Y_1, represent the horizontal and vertical coordinates of the first point in the plane, the second pair of elements, denoted X_2 and Y_2, represent the horizontal and vertical coordinates of the second point, and so on. For example, the row vector in Figure 1 represents the four points shown in the figure. Each photo can be decomposed into two parts, $\mathbf{E}(t) = \mathbf{P} + \mathbf{d}(t)$, where $\mathbf{P}$ is the true star pattern (the prototype), and $\mathbf{d}(t)$ is the disturbance for trial t. The trial number, t, represents the number of photos that the subject has observed from the same star pattern (t ranges from 1 to 5).

The main design of the experiment was based on the manipulations of the two elements, X_1 and Y_2. The manipulation of X_1 was used to provide a test of the additivity and time invariance properties. The manipulation of Y_2 was used to test the noninterference property. (These properties are described in the Results section).

The location of X_1 was assigned to either -5 or $+5$ units for each trial. (One unit equals 0.5 cm.) For example, X_1 was assigned the values $(-5, -5, +5, -5, +5)$ across the five trials for one of the star patterns. All 32 possible -5, $+5$ sequences of length five were used in the experiment, one sequence for each star pattern. The location for Y_2 was fixed at either -5, 0, or $+5$ units for all five trials of a given star pattern, but this value varied across star patterns. The three possible values for Y_2 were crossed with the 32 sequences for X_1 to produce a 3×2^5 factorial design with 96 conditions. Each subject observed all 96 conditions in a different random order.

The remaining elements $(Y_1, X_2, X_3, Y_3, X_4, Y_4)$ were generated as follows. First, the true star position for each element was randomly sampled from a normal distribution with zero mean and a standard deviation of five (0.5 cm) units. The five disturbances for each element were randomly sampled from a normal distribution with zero mean and a standard deviation of two. For the practice stimuli, all eight of the elements were constructed from the same procedure used to construct the elements X_3 or Y_3.

Subjects

Subjects were 8 graduate psychology students (6 female and 2 male) who were paid \$4 per session for their voluntary participation in the experiment.

Results

Overview

Before we present the four different types of analyses, we introduce some descriptive statistics, followed by tests of additivity, noninterference, and time invariance. A linear system model of prototype learning was fit separately to the data from each subject.

Note that this paradigm quickly generates a massive data base. Each subject generated a vector of eight responses,

denoted $\mathbf{R} = (RX_1, RY_1, \ldots, RX_4, RY_4)^T$ corresponding to the vector of eight stimulus coordinates $\mathbf{E} = (X_1, Y_1, \ldots, X_4, Y_4)^T$ for each of 480 photos. However, all of the analyses reported below were based on only the first four coordinates to reduce the amount of redundancy in the data analyses. Because a large number of significance tests were performed, all significance tests were conducted at the .01 level to reduce the overall Type I error rate.[1]

Descriptive Statistics

The frequency distribution provides a qualitative test of averaging. If subjects averaged the stimulus coordinates, then the response distributions (pooled across all 480 stimuli for a given subject) should be unimodal, symmetric, and centered at zero for all measures except RY_2. Alternatively, if subjects simply retrieved a single trace of an exemplar and based their estimate on this single trace (e.g., the mode), then the response distribution for RX_1 should be bimodal with large frequencies centered at −5 and +5. In fact, the frequency distributions for RX_1, RY_1, and RX_2 were unimodal, symmetric, and centered at zero for all subjects. The only exception was the distribution of RY_2, which was trimodal with heavy concentrations at −5, 0, and +5.

The variance of the responses provide a simple test of some standard models. The optimal prototype estimate is the arithmetic average of all t exemplars. Because the variance of a sample mean equals the variance of the scores divided by the sample size (t), the optimal model predicts that the variance of the responses decreases at a rate equal to $1/t$.

A proportional change model is often used for prediction. This model assumes that the new prototype estimate is a weighted average of the previous prototype estimate and the new exemplar. In this case, each new exemplar receives the same weight independent of sample size. This model predicts that the variance initially increases as the sample size (t) increases, but eventually the variance levels off at some asymptote (see Chatfield, 1975, p. 45).

The observed variances (pooled across response coordinates) initially decreased but then leveled off as t increased. The observed variances were 19.6, 16.8, 15.4, 15.4, 15.4 for t = 1 to 5, and the optimal model predicts 19.6, 9.8, 6.5, 4.9, 3.92 for t = 1 to 5. Thus, the observed rate of decrease in the variances is inconsistent with both the proportional change and optimal model. Later, more evidence will be reported that indicates the use of a complex weighted averaging scheme.

One final observation is that the correlations among all responses were all less than .01 in magnitude. This would be expected if there was no direct influence of one response coordinate on another.

Definitions of the Three Basic Properties

First, it is necessary to distinguish between the subject's and the experimenter's representation of each exemplar. We chose to represent each exemplar by the eight-element column vector **E** defined in terms of rectangular coordinates. The subject's representation of the same exemplar will be symbolized by the column vector, **f**, that contains at least eight and possibly more elements or features. It is assumed that **f** is related to **E** by an affine transformation, $\mathbf{f} = \mathbf{TE} + \mathbf{B}$, where **T** is a $(n \times 8)$ matrix of constants, and **B** is a $(n \times 1)$ matrix of constants. The map from **E** to **f** is assumed to be one to one so that each feature vector **f** corresponds to only one feature vector **E** and vice versa.

For example, subjects may contract, translate, or rotate the rectangular coordinates **E**. However, these are only special cases, and more complex affine transformations are possible. For example, subjects also may encode the differences between each pair of points. These differences would also be related to **E** by an affine transformation.

It is also necessary to distinguish between the image retrieved by the subject, denoted **F**, and the response vector **R**. We chose to record **R** in terms of rectangular coordinates. The retrieved image is assumed to be related to the recorded response by the same affine transformation that relates the experimenter's and subject's representations of the exemplars.

The three properties—additivity, noninterference, and time invariance—can be understood by referring to a linear system model of prototype learning. This model states that the image retrieved immediately before trial $(t + 1)$ is a weighted sum of the features of the t previously experienced exemplars generated from the same prototype.

$$\mathbf{F}(t + 1) = \Sigma\, W(t - k)\mathbf{f}(k), \qquad (1)$$

where the summation extends across the serial positions $k = 1, \ldots, t$ for trials t = 1 to 5. Serial position k refers to the kth position within a sequence of t trials. For example, if $t = 4$ photos of the same star pattern have been presented so far, then $k = 2$ refers to the second of these four photos.

The weight, denoted $W(t - k)$, is a scalar that multiplies all of the features of the exemplar presented at serial position k. Note that the weight depends only on the lag = $(t - k)$, which is the difference between the current trial number and the serial position that an exemplar was presented. For example, if $t = 4$ exemplars have been presented so far, then the lag for the first exemplar ($k = 1$) is $(4 - 1) = 3$. The lag for the fourth (current) exemplar is zero.

The property of *additivity* refers to the assumption that the weighted features of each exemplar are added together. The property of *noninterference* refers to the assumption that the value of the jth coordinate of the retrieved prototype, $F_j(t + 1)$, should be influenced only by the value of the jth coordinate of each exemplar. The *time invariance* property refers to the assumption that the magnitude of the effect of each exemplar depends only on the lag. The Appendix shows that if the experimenter's and subject's features are related by an affine transformation, then empirical tests of the validity of these three basic properties are not influenced by the choice of affine transformation.

Both the optimal and the proportional change model imply additivity and noninterference. The proportional change

[1] Reducing the significance level to $\alpha = .05$ would not change the main conclusions, but it would produce a large increase in the Type I error rate.

model also implies time invariance (it is a special case of Equation 1), but the optimal model does not because the weights of the optimal model depend on the sample size, t.

Tests of Additivity

Additivity across exemplars was assessed by analyzing the interaction effects of the manipulations of X_1 on RX_1 (cf. Anderson, 1964). Before describing the results in general, it may be useful to consider the example shown in Table 1, which illustrates a test of an interaction among Photos 1, 4, and 5 on the response following all five photos. The first three columns indicate the values of stimulus coordinate X_1 for Photos 1, 4, and 5. The fourth column shows the mean values of the RX_1 responses after all five photos, and the last column shows differences between adjacent rows. Violations of additivity are indicated by differences among the differences in the last column.

Five repeated measures analyses of variance (ANOVAS) were performed to test the significance of the main and interaction effects among the photos. The first ANOVA was a one-way analysis performed on the responses following the first photo, the second ANOVA was a two-way 2^2 analysis performed on the responses following the first two photos, the third was a three-way 2^3 analysis performed on the responses following the first three photos, the fourth was a four-way 2^4 analysis performed on the responses following the first four photos, and the fifth was a five-way 2^5 analysis performed on the responses following all five photos. All main effects were large and significant. Only one of the 42 possible interactions (the three-way interaction among Photos 1, 4, and 5 from the five-way analysis seen in Table 1) was significant, $F(1, 7) = 25.05$, $MS_e = 0.5$.

As can be seen in Table 1, when Photos 1 and 4 were both negative, then the observed effect of a negative coordinate value for the fifth photo was smaller than the effect predicted by additivity. A similar result occurred when Photos 1 and 4 were both positive: The effect of a positive coordinate value for the fifth photo was smaller than the effect predicted by additivity. It seems as if subjects were not willing to draw a point at the extreme negative or positive sides of the page.

Table 1
Means for the Interaction Effect of Stimulus Coordinate X_1 for Photos 1, 4, and 5 on Response Coordinate RX_1

Condition				
Photo 1	Photo 4	Photo 5	RX_1	Difference
−5	−5	−5	−2.92	3.00
−5	−5	+5	0.08	
−5	+5	−5	−1.58	3.86
−5	+5	+5	2.28	
+5	−5	−5	−1.81	3.43
+5	−5	+5	1.61	
+5	+5	−5	0.02	3.27
+5	+5	+5	3.29	

Tests of Interference

Two six-way, $2^5 \times 3$, repeated-measures ANOVAS were performed to test the main and interaction effects of stimulus coordinates X_1 and Y_2 on the responses following all five photos for coordinates RY_1 and RX_2. The only significant finding from a total of 126 independent tests was the interaction effect of the first photo for coordinate X_1, the third photo for coordinate X_1 and the value of coordinate Y_2 on the response coordinate RY_1 after observing all five photos, $F(2, 14) = 9.9$, $MS_e = 10.7$. This interaction effect was complex and difficult to interpret. Given that one significant finding out of 126 tests is about the number expected by chance, it seems likely that a Type I error had occurred.

Serial Position Effects

The weights shown in Equation 1 imply that serial position effects are solely a function of the lag, $(t - k)$. In other words, the main effect of the most recent photo (lag 0) does not depend on the number of photos presented, and the main effect of the second most recent photo (lag 1) does not depend on the number of photos presented, etc. The following analysis directly tests this assumption by using a method developed by Weiss and Anderson (1969).

Separate estimates of the main effects of stimulus coordinate X_1 on response coordinate RX_1 were obtained following each photo, producing five sets of serial position effects. The first set, containing only one main effect equal to 8.08, is the effect of the first photo when only one photo was presented. The second set (containing two main effects) was estimated from the main effects of Photos 1 and 2 after observing only two photos. The fifth set (containing five main effects) was estimated from the main effects of Photos 1 through 5 after observing all five photos. Each main effect was estimated from 384 observations.

The last four sets of serial position effects are shown in Figure 2 plotted as a function of the lag. Consider, for example, the curve with the parameter $t = 4$. The point at lag 0 shows the main effect of the fourth photo when four photos were presented. The point at lag 3 shows the main effect of the first photo when four photos were presented.

Before discussing the details, it might be helpful to point out several predictions. According to the optimal model, the serial position curves should be flat, and the level of each flat line should decrease the sample size (t) increases. According to the linear system model (including the proportional change model), the serial position curves should all line up on top of each other forming a single curve.

Figure 2 shows that both the optimal and linear system models are incorrect. The strong recency effect violates the equal weight assumption of the optimal model. The systematic reduction of the recency effect with increased sample size (t) violates the time invariance property of the linear system model.

Finally, note that when five photos were presented, a primacy effect occurs when comparing lags 4 and 3 (corresponding to Serial Positions 1 and 2, respectively). A primacy effect

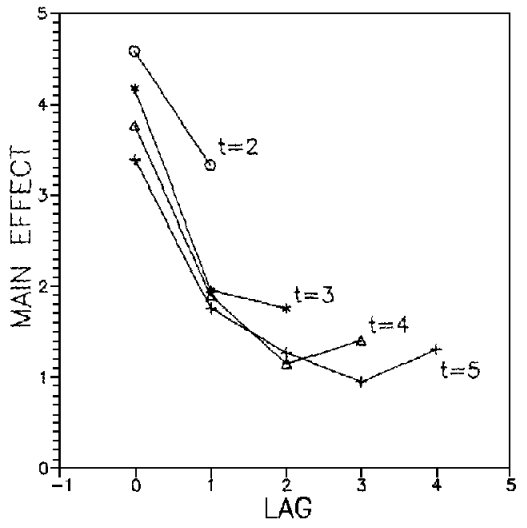

Figure 2. Serial position effects plotted as a function of lag after observing $t = 2$, 3, 4, or 5 photos. For example, consider the curve labeled $t = 4$. The point plotted at lag = 2 is the main effect of second photo on the response following four photos. Each point is an average of 384 observations.

Table 2
Serial Position Weights (W_k) and Percentage of Variance (R^2) Predicted by the Linear System Model (Equation 1) Fit to the Response Coordinates RY_1 and RX_2 for Each Subject

Subject	W_0	W_1	W_2	W_3	W_4	R^2
1	.50	.27	.11	.02	.04	.90
SE	.02	.02	.02	.02	.03	
2	.42	.23	.06	.01	−.02	.79
SE	.02	.02	.02	.03	.03	
3	.33	.12	.02	.03	−.04	.65
SE	.02	.03	.02	.03	.03	
4	.64	.20	.06	−.03	.03	.86
SE	.02	.03	.03	.03	.03	
5	.49	.33	.07	−.02	.04	.84
SE	.03	.03	.03	.03	.03	
6	.51	.26	.11	.02	−.01	.87
SE	.02	.02	.02	.02	.03	
7	.59	.31	.07	.00	.03	.94
SE	.02	.02	.02	.02	.02	
8	.48	.27	.07	.01	.02	.88
SE	.02	.02	.02	.02	.02	
M	.50	.25	.07	.0	.01	.84

Note. $N = 768$ data points per subject.

also occurs after observing four photos. However, when only 2 or 3 photos were presented, a recency effect occurs at Serial Positions 1 and 2. Thus, the rank ordering of main effects according to serial position varies as the number of photos presented increases. This rules out attention decrement models which assume that the weights depend solely on serial position (see Busemeyer, 1987).

The serial position curves shown in Figure 2 were also calculated separately for each subject. All subjects showed violations of time invariance in the direction indicated by Figure 2. However, there were striking individual differences: 5 subjects produced patterns very similar to Figure 2, 3 subjects produced patterns with strong recency effects and no primacy effects, and 1 subject produced weak primacy effects.

Quantitative Analyses

The linear system model (Equation 1) was fit separately to each subject's data. The model was applied to the responses following Photos 2, 3, 4, and 5 for all 96 star patterns and two response coordinates, RY_1 and RX_2, producing a total of $4 \times 96 \times 2 = 768$ observations per subject. Five weight parameters (W_{t-k}, for $(t - k) = 0$, 1, 2, 3, and 4) plus a separate intercept for each response was estimated using multiple regression analysis. The estimated parameters and percentage of variance predicted by the model (R^2) for each subject is shown in Table 2. Also shown are the estimated standard errors. The R^2 values are fairly high (an average of 84% of the variance was predicted).

According to this analysis, all subjects produced strong recency effects. The average serial position weights shown at the bottom of Table 2 can be approximated by the exponential function 0.5^{t-k}.

So far, all of the preceding analyses have been based on the assumption that the subjective features (**f**) are related to the rectangular coordinates (**E**) by an affine transformation. It is possible that some other coordinate system was used. For example, subjects may have encoded each point in terms of its angle and length (polar coordinates), which is not an affine transformation of the rectangular coordinates. To test this hypothesis, the position of the first point of each photo was predicted using Equation 1 with either rectangular coordinates or polar coordinates. Both models contained exactly the same parameters.[2] The percentage of variance predicted by the rectangular coordinate system was greater for all subjects. On the average, the rectangular coordinate model produced a 10% increase in predicted variance over the polar coordinate model.

Discussion

The present experiment was designed to empirically test three basic properties of prototype learning models by using the prototype production paradigm. The first property was the additive effects of a sequence of exemplars on the evolving prototype, and a small but statistically reliable deviation from additivity was obtained. The second property was noninterference across exemplar features, and there was not much evidence for interference effects. The third property was the time-invariance property of serial position effects, and the

[2] The rectangular coordinates were translated by adding 15 to both coordinates before transforming to polar coordinates. This translation was selected to improve the fit of the polar coordinate model. This additive factor actually provided the polar coordinate model one extra parameter.

results clearly violate this property. The implications for various memory models are elaborated next. First, it is shown that for the present task, the predictions generated from these models are the same as those generated from the linear system model (Equation 1). Thus, they all imply additivity, noninterference, and time invariance. Afterward, explanations for the violations of additivity and time invariance are considered.

Multiple Trace Models

Hintzman's (1986) schema abstraction model (MINERVA 2) can be applied in a fairly direct manner to the prototype production task. Each pairing of exemplar with category label produces a separate memory trace. Each memory trace is represented by a vector of features. A subset of the elements within each vector represents the features of the exemplar presented on trial t, denoted $\mathbf{f}(t)$. The remaining subset of elements represents the features of the category label associated with each exemplar, denoted $\mathbf{g}(t)$. Note that the category-label features, $\mathbf{g}(t)$, vary across different prototypes, but they are constant across all exemplars generated from the same prototype.

When subjects are asked to reproduce the prototype, the category label operates as a retrieval cue for activating all of the traces in memory that have a nonzero lag. The degree of activation is a function of inner product, $\mathbf{g}(t+1)^T\mathbf{g}(k)$, between the current category-label features preceding trial $(t+1)$, and the trace of the category-label features from serial position k. (The inner product $\mathbf{X}^T\mathbf{Y} = \Sigma X_j Y_j$ is a measure of the similarity between $\mathbf{X}$ and $\mathbf{Y}$.)

According to Hintzman (1986, see his Equation 4) the retrieved prototype image is calculated by Equation 1 of the present article, and the weight for lag $(t-k)$ equals the inner product raised to the third power, $W(t-k) = [\mathbf{g}(t+1)^T\mathbf{g}(k)]^3$. The inner product remains constant for all category labels paired with exemplars generated from a common prototype, and it is approximately zero otherwise. Therefore, the weight is solely a function of the lag. Note that MINERVA 2 retrieves a prototype estimate even though such an abstraction was never stored.

Holographic Memory Models

Metcalfe-Eich's (1982) holographic model of prototype learning can also be applied to the prototype production task. According to this model, the association of an exemplar with a category label is represented by the convolution of the features of the category label with the features of the exemplar, producing the association vector $[\mathbf{g}(t)*\mathbf{f}(t)]$. The associations produced by a sequence of t pairings are summed to form a composite trace $\mathbf{M}(t) = \Sigma\,[\mathbf{g}(k)*\mathbf{f}(k)]$, where the summation ranges from $k = 1$ to t. When a category label is presented and subjects are asked to reproduce the prototype estimate, the category label operates as a retrieval cue. The retrieved image of the prototype estimate is obtained by correlating the category-label features with the composite memory trace, symbolized as $\mathbf{F}(t+1) = \mathbf{g}(t+1)\#\mathbf{M}(t)$. (See the Appendix for definitions of the convolution and correlation operations.)

Given the usual assumptions regarding the construction of feature vectors (see Metcalf-Eich, 1982, p. 632), it can be shown (see Appendix) that $\mathbf{F}(t+1) = \mathbf{g}(t+1)\#\mathbf{M}(t)$ reduces to Equation 1 plus noise. According to this model, the weight for each lag is equal to the inner product between the category-label features, $W(t-k) = \mathbf{g}(t+1)^T\mathbf{g}(k)$. This inner product is constant for category labels paired with exemplars generated from a common prototype, and approximately zero otherwise. Therefore, the weight is solely a function of the lag.

Distributed Memory Models

Several researchers have suggested that distributed memory provides a natural explanation for prototype learning (Heath & Fulham, 1985; Knapp & Anderson, 1984; McClelland & Rumelhart, 1985). A modified version of the Knapp and Anderson model is described because of its relative simplicity. However, the basic properties derived from this particular model are fairly general.

The values of the features used to encode the category label presented on trial t are defined as a (column) vector denoted $\mathbf{g}(t)$. The values of the features used to encode the exemplar presented on trial t are defined as a (column) vector denoted $\mathbf{f}(t)$. The connection strength on trial t between ith category-label feature $g_i(t)$ and the jth exemplar feature $f_j(t)$ is denoted $a_{ij}(t)$. In the prototype production task, the category-label features are presented as input and prototype features are retrieved as output. The value of the prototype feature j retrieved before trial $t+1$, denoted $F_j(t+1)$, is obtained from the sum of products

$$F_j(t+1) = \Sigma\, g_i(t+1)a_{ij}(t),$$

where the sum ranges across the category-label feature index i.

The connection strengths, $a_{ij}(t)$, are assumed to be updated according to either a Hebb rule or a delta rule. According to the Hebb rule, the change in connection strength is determined by the product of the input and output activation levels. One version of Hebb rule can be stated as follows:

$$a_{ij}(t) = (1-c)a_{ij}(t-1) + (c)g_i(t)f_j(t),$$

where c is a learning rate parameter between zero and one.

According to the delta rule, the change in connection strength depends on the discrepancy between the observed and predicted feature values. The delta rule (cf. Stone, 1986) can be stated as follows

$$a_{ij}(t) = a_{ij}(t-1) + (c)g_i(t)[f_j(t) - F_j(t)],$$

where $c > 0$ is the learning rate parameter.

It can be shown (see Appendix) that for both the Hebb and the delta learning rule, the distributed memory model reduces to Equation 1. According to the Hebb rule,

$$W(t-k) = c(1-c)^{t-k}\mathbf{g}(t+1)^T\mathbf{g}(k),$$

that is the inner product multiplied by a recency weight. The inner product is constant for all category labels paired with exemplars generated from the same prototype. Therefore, the weight is solely a function of the lag.

According to the delta rule,

$$W(t - k) = (c)\mathbf{g}(t + 1)^T\mathbf{Q}(t, k)\mathbf{g}(k).$$

($\mathbf{X}^T$ symbolizes the transposition of the vector $\mathbf{X}$.) For lag zero, $\mathbf{Q}(t, t) = \mathbf{I}$ (an identity matrix), and for nonzero lags, $\mathbf{Q}(t, k)$ is defined as the product of $(t - k)$ matrices.

$$\mathbf{Q}(t, k) = [\mathbf{I} - (c)\mathbf{g}(t)\mathbf{g}(t)^T] \ldots [\mathbf{I} - (c)\mathbf{g}(k + 1)\mathbf{g}(k + 1)^T].$$

Because the category-label features are constant for all category labels paired with exemplars generated from the same prototype, the matrix $\mathbf{Q}(t, k)$, and consequently the weight $W(t - k)$, is solely a function of the lag.

Explanations for Nonadditivity

All three memory models imply that the exemplars are combined according to an additive rule, but the interaction shown in Table 1 violates additivity. However, it may be worth considering alternative explanations that allow one to retain the linear system model, for two reasons. Linear systems are mathematically more tractable than nonlinear systems. In addition, the observed violations of additivity were few in number (only one) and small in magnitude. The following two explanations allow one to retain the linear system model.

Incorrect features. One possible explanation for the deviations from additivity is that the wrong features were used in the analyses.[3] Perhaps the additive property may be retained if a different coordinate system was used that was nonlinearly related to the rectangular coordinate system. For example, suppose that the memory system is additive, but subjects used polar coordinates to encode each exemplar (i.e., encode each dot in terms of a length and angle). Then the analyses in the Results section would yield both nonadditive and interference effects since those analyses were based on a rectangular coordinate system. However, it is unlikely that subjects were using polar coordinates, because if they were, interaction and interference effects would be much more extensive than those obtained in the present study. Also, rectangular coordinates provided a better fit to Equation 1 than polar coordinates.

Response bias. A more plausible explanation is that the memory system accumulates information in an additive fashion, and the small interaction may be due to the way that subjects squeeze the image at the extremes to fit their drawings into the bounded plane outlined on the computer page. This nonlinear response transformation can produce interaction effects analogous to floor or ceiling effects obtained with percent correct measures of performance. Conjoint measurement techniques may be used to evaluate this hypothesis (cf. Krantz, 1973).

Explanations for Serial Position Effects

The time-invariance property of memory systems is important because it greatly reduces the number of parameters needed to describe the learning process. All of the accounts of previous research by Hintzman (1986), Knapp and Anderson (1984), and Metcalf-Eich (1982) were based on this assumption. However, the analysis of serial position effects shown in Figure 2 clearly indicates that the system is time variable. Two explanations for violations of time invariance follow.

Contextual cues. In previous research, it has been assumed that the category-label features operate as the primary retrieval cue. Alternatively, one could assume that context (e.g., extraneous thoughts) experienced during learning become associated with the exemplars and this context also operates as a retrieval cue. This context fluctuates across trials (perhaps at different rates for different features). To be more explicit, suppose that the retrieval cue can be decomposed into two parts $\mathbf{g}(t) = \mathbf{c} + \mathbf{e}(t)$, where $\mathbf{c}$ represents the common category-label features associated with a prototype and $\mathbf{e}(t)$ is a noise vector that changes across trials. Thus, the inner product, $\mathbf{g}(t + 1)^T\mathbf{g}(k)$, is no longer constant across exemplars generated from the same prototype.

This context hypothesis is similar to that proposed by Glenberg, Bradley, Kraus, and Renzaglia (1983). However, normally it is assumed that the similarity of the context at different trials depends only on the lag: That is, $\mathbf{e}(t)$ is a weakly stationary stochastic process with autocorrelations that depend only on lag (cf. Chatfield, 1975). Thus, the context hypothesis helps explain recency effects, but it does not explain violations of time invariance.

Variable learning rate. An alternative hypothesis is that the learning rate (or attention) parameter of the distributive memory model is a function of serial position. For example, setting $c(t) = c/m(t)$, where $m(t)$ is an increasing function of t, would produce an averaging mechanism that has properties similar to both the optimal model and the proportional change model. Setting $c = 1$ and $m(t) = t$ yields an arithmetic average, and setting $m(t) = 1$ produces a proportional change model. Of course, it is also possible to include a time variable learning rate parameter in the holographic memory model (see Lewandowsky & Murdock, 1986) and the multiple-trace model.

Output interference. Weiss and Anderson (1969) found that serial position curves change depending on whether an estimate is required after each stimulus or only once at the end of the sequence. This suggests that subjects may average (according to Equation 1) both the previous exemplars and their previous drawings when estimating the prototype (perhaps with different weights). According to the theory described in the Appendix, the inclusion of previous drawings into the estimate would influence the weight parameters for each lag, but this would not influence the tests of the validity of the three basic properties. In particular, the estimate would still have the property of time invariance (see Appendix). Unfortunately, one cannot test time invariance when a response is required only once at the end of a sequence.

Relation to serial position effects in recall. It seems reasonable to suspect that the serial position effects obtained with a

[3] Knapp and Anderson (1984) used a different set of features to represent their dot stimuli. The difference is due to a procedural change between their experiment and the present experiment. In Knapp and Anderson's experiment, the dots were not labeled and there was no way to form a correspondence among dots across exemplars. In the present study, the dots were labeled so that subjects could form a correspondence among dots with identical labels across exemplars.

prototype production task can be explained by serial position effects observed with free recall. However, note that recency effects obtained from serial recall are not influenced by list length (Murdock, 1962), whereas the recency effects obtained in the present study decrease with list length. More important, a series of studies that directly compared abstraction and recall (see Dreben, Fiske, & Hastie, 1979, for a review), found no correlation between serial position effects obtained from abstraction tasks and serial position effects obtained from free-recall tasks.

Relation to research on intuitive statistics. The prototype production task can be considered a multidimensional version of the unidimensional mean estimation task (see Busemeyer, 1987, for a review). The results of the previous research with univariate mean (Weiss & Anderson, 1969) and relative frequency (Shanteau, 1970) estimation tasks are very similar to the present results: Only small violations of additivity are found, but large violations of time invariance are observed. In particular, recency effects decrease as the number of observations increase, and striking individual differences in the shape of the serial position curve are found.

Conclusion

The results of the present study indicate that linear, time-variable, memory systems provide a fairly good description of the evolution of prototypes. (This conclusion is based on the belief that the small deviation from additivity observed in this study was due to a response bias.) Previous memory models of prototype learning have assumed time invariance, and therefore, they fail to account for the fact that the magnitude of a recency effect decreases as the number of exemplars increases. These conclusions are limited, of course, to the use of random dot stimuli, and further research is needed to see how well these models perform with more complex natural stimuli. Nevertheless, it is important to establish first that these models work well with simple artificial stimuli.

A more powerful test of memory models of prototype learning can be realized by combining both the categorization paradigm and the prototype production paradigm into a single study. The same memory system (either a multiple-trace system or composite distributed system) could be applied to the responses obtained from both tasks. This would provide converging operations for empirical tests of competing memory models.

References

Anderson, N. H. (1964). A note on weighted sum and linear operator models. *Psychonomic Science, 1*, 189–190.

Busemeyer, J. R. (in press). Intuitive statistical estimation. In N. H. Anderson (Ed.) *Contributions to information integration theory.* New York: Academic Press.

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.

Chatfield, C. (1975). *The analysis of time series: Theory and practice.* London: Chapman and Hall.

Dreben, E. K., Fiske, S. T., and Hastie, R. (1979). The independence of evaluation and item information: Impression and recall order effects in behavior-based impression formation. *Journal of Personality and Social Psychology, 37*, 1758–1768.

Glenberg, A. M., Bradley, M. M., Kraus, T. A., & Renzaglia, G. J. (1983). Studies of the long term recency effect: Support for a contextually guided retrieval hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6*, 355–369.

Heath, R. A., & Fulham, R. (1985). Applications of system identification and adaptive filtering techniques in human information processing. In G. d'Ydewalle (Ed.), *Cognition, information processing, and motivation: Vol. 3. XXIII International congress of psychology* (pp. 117–147). Amsterdam: North Holland.

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

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 83–94.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 616–637.

Krantz, D. H. (1973). Measurement-free tests of linearity in biological systems. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-3*, 266–271.

Lewandowsky, S., & Murdock, B. B. (1986). *Memory for serial order.* Unpublished manuscript.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159–188.

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

Murdock, B. B., Jr. (1962). Serial position effects in free recall. *Journal of Experimental Psychology, 64*, 482–488.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304–308.

Schönemann, P. H. (1987). Some algebraic relations between involutions, convolutions, and correlations with applications to holographic memories. *Biological Cybernetics, 56*, 367–374.

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology, 85*, 181–191.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition: Volume 1. Foundations* (pp. 444–459). Cambridge: MIT Press.

Weiss, D. J., & Anderson, N. H. (1969). Subjective averaging of length with serial presentation. *Journal of Experimental Psychology, 82*, 52–63.

Appendix

Mathematical Models

Effects of Affine Transformations

Recall that **E** and **R** are representations of the exemplar and the prototype estimate, respectively, in terms of coordinates selected by the experimenter, but **f** and **F** are the corresponding representations selected by the subject. It is assumed that $\mathbf{f} = \mathbf{TE} + \mathbf{B}$ and $\mathbf{F} = \mathbf{TR} + \mathbf{B}$, where the n by 8 matrix **T** has a rank equal to 8. The latter condition implies that $\mathbf{H} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$ exists, and note that $\mathbf{H}(\mathbf{TE}) = \mathbf{E}$. ($\mathbf{X}^T$ represents the transpose of the matrix **X**, and $\mathbf{X}^{-1}$ is the inverse of **X**). Thus $\mathbf{R} = \mathbf{H}(\mathbf{F} - \mathbf{B}) = \mathbf{HF} - \mathbf{HB}$, and multiplying both sides of Equation 1 by **H** yields

$$\mathbf{R}(t + 1) + \mathbf{HB} = \mathbf{HF}(t + 1) = \Sigma\ W(t - k)\cdot\mathbf{Hf}(k),$$

where the sum extends from $k = 1$ to t. Substituting $\mathbf{TE} + \mathbf{B}$ for **f** yields $\mathbf{R}(t + 1) = \Sigma\ W(t - k)\cdot\mathbf{E}(k) - [I - \Sigma\ W(t - k)]\mathbf{HB}$. The last term subtracts out when the differences are computed for the tests of additivity, noninterference, and time invariance.

Holographic Memory Model

Recall that $\mathbf{g}(t)$ is a vector of category-label features and the symbol $g_j(t)$ is the jth element of this vector. Also, $\mathbf{f}(t)$ is a vector of exemplar features and the symbol $f_j(t)$ is the jth element of this vector. For the holographic model, it is normally assumed that both vectors are infinite sequences with a finite number of nonzero elements. The convolution $\mathbf{h}(t) = [\mathbf{g}(t)*\mathbf{f}(t)]$ produces a new vector, and the jth element of the new vector is defined by the sum

$$h_j(t) = \Sigma\ f_i(t)\cdot g_{(j-i)}(t), \tag{A1}$$

where the summation extends across the index i. The composite memory after t pairings is the sum $\mathbf{M}(t) = \Sigma\ [\mathbf{g}(k)*\mathbf{f}(k)]$, where the summation extends across serial positions $k = \mathrm{i}, \ldots, t$. The correlation between $\mathbf{g}(t + 1)$ and $\mathbf{M}(t)$ produces a new vector

$$\mathbf{r}(t + 1) = \mathbf{g}(t + 1)\ \#\mathbf{M}(t) = \Sigma\ [\mathbf{g}(t + 1)\ \#\mathbf{h}(k)], \tag{A2}$$

where the summation extends across serial positions $k = 1, \ldots, t$. The mth-element of the vector $[\mathbf{g}(t + 1)\ \#\mathbf{h}(k)]$ is defined as

$$[\mathbf{g}(t + 1)\ \#\mathbf{h}(k)]_m = \Sigma\ g_{(j-m)}(t + 1)\cdot h_j(k), \tag{A3}$$

where the summation extends across the index j. By substituting the definition of $h_j(k)$ given by Equation A1 into Equation A3, and algebraically rearranging terms, it is possible to show that

$$\mathbf{g}(t + 1)\ \#[\mathbf{g}(k)*\mathbf{f}(k)] = [\mathbf{g}(t + 1)\ \#\mathbf{g}(k)]*\mathbf{f}(k).$$

(For a complete table of identities, including the one shown above, see Schönemann, 1987).

There is a special vector, δ, that has the following property $[\delta*\mathbf{f}(t)] = \mathbf{f}(t)$. Metcalf-Eich (1982, p. 632) assumes that all feature vectors have the property that $\mathbf{g}(t)\ \#\mathbf{g}(t)$ is approximately equal to δ. For the simulations on prototype learning, Metcalf-Eich assumed that the category-label features were constant across exemplars generated from the same prototype. Therefore, $[\mathbf{g}(t + 1)\ \#\mathbf{g}(k)]*\mathbf{f}(k)$ is approximately equal to $\delta*\mathbf{f}(k) = \mathbf{f}(k)$. Substituting this last result into Equation A2 yields Equation 1 (with weights equal to unity).

Distributed Memory Model

Each exemplar is represented by a column vector of features, denoted $\mathbf{f}(t)$, and the category label is represented by a column vector of features, $\mathbf{g}(t)$. Both of these vectors are assumed to have a finite number of elements. It is convenient to represent the connections between the category-label features and the exemplar features by a matrix $\mathbf{A}(t)$, where $a_{ij}(t)$ is the cell in row i column j and this cell represents the strength of the connection between category-label feature i and exemplar feature j. On the basis of these definitions, the retrieved prototype estimate is obtained by the matrix product

$$\mathbf{F}(t + 1) = \mathbf{A}(t)^T\mathbf{g}(t + 1). \tag{A4}$$

The Hebb rule can be stated in matrix form as follows (assuming $\mathbf{A}(0) = 0$):

$$\mathbf{A}(t) = (1 - c)\cdot\mathbf{A}(t - 1) + c\cdot\mathbf{g}(t)\mathbf{f}(t)^T.$$

The solution to the difference equation yields

$$\mathbf{A}(t) = \Sigma c\cdot(1 - c)^{t-k}\cdot\mathbf{g}(k)\mathbf{f}(k)^T, \tag{A5}$$

where the summation extends from $k = 1$ to t. Substituting the right-hand side of Equation A5 into Equation A4 yields Equation 1 with $W(t - k) = c\cdot(1 - c)^{t-k}\cdot\mathbf{g}(t + 1)^T\mathbf{g}(k)$. If $\mathbf{g}(t)$ is assumed to be constant for all exemplars generated from the same prototype, then the inner product $\mathbf{g}(t + 1)^T\mathbf{g}(k)$ is a constant.

The delta rule can be written in matrix form as follows (assuming $\mathbf{A}(0) = 0$):

$$\mathbf{A}(t) = \mathbf{A}(t - 1) + c\cdot\mathbf{g}(t)[\mathbf{f}(t) - \mathbf{F}(t)]^T.$$

The solution to the difference equation yields

$$\mathbf{A}(t) = \Sigma c\cdot\mathbf{Q}(t, k)\mathbf{g}(k)\mathbf{f}(k)^T, \tag{A6}$$

where the summation ranges from $k = 1$ to t. For $k = t$, $\mathbf{Q}(t, t) = \mathbf{I}$, the identity matrix. For $t > k$,

$$\mathbf{Q}(t, k) = \prod_{i=1}^{t-k} [\mathbf{I} - c\cdot\mathbf{g}(t - i + 1)\mathbf{g}(t - i + 1)^T].$$

Substituting the right-hand side of Equation A6 into Equation A4 yields Equation 1 with $W(t - k) = c\cdot\mathbf{g}(t + 1)^T\mathbf{Q}(t, k)\mathbf{g}(k)$. If $\mathbf{g}(t)$ is assumed to remain constant across exemplars generated from the same prototype, then $\mathbf{Q}(t, k) = \mathbf{Q}^{t-k}$, a constant matrix raised to the power $(t - k)$.

Output Interference

Suppose subjects average both the previous exemplars and the previous prototype estimates according to Equation 1 with perhaps different weights assigned to the exemplars and the previous estimates. More specifically, for $t > 1$,

$$\mathbf{F}(t + 1) = W(0)\mathbf{f}(t) + \Sigma\ W(t - k)\cdot\mathbf{f}(k) + V(t - k)\cdot\mathbf{F}(k + 1), \tag{A7}$$

where the summation extends across $k = 1$ to $t - 1$. The weights $W(t - k)$ and $V(t - k)$ are solely a function of the lag. The first estimate, denoted $\mathbf{F}(2)$, is simply equal to $\mathbf{F}(2) = W(0)\cdot\mathbf{f}(1)$. Given this first estimate, Equation A7 can be solved iteratively and expressed as

$$\mathbf{F}(t + 1) = \Sigma\ W^*(t - k)\cdot\mathbf{f}(k),$$

where the summation extends across $k = 1$ to t. The new weights can be defined recursively as $W^*(0) = W(0)$, and for $j > 0$

$$W^*(j) = \Sigma\ V(i)W^*(j - i) + W(j),$$

where the summation extends across $i = 1$ to j.

Received July 31, 1986
Revision received March 16, 1987
Accepted March 23, 1987 ■