Journal of Experimental Psychology:
Learning, Memory, and Cognition
1984, Vol. 10, No. 2, 234-257

Induction of Category Distributions: A Framework for Classification Learning

Lisbeth S. Fried and Keith J. Holyoak
University of Michigan (Ann Arbor)

We present a framework for classification learning that assumes that learners use presented instances (whether labeled or unlabeled) to infer the density functions of category exemplars over a feature space and that subsequent classification decisions employ a relative likelihood decision rule based on these inferred density functions. A specific model based on this general framework, the *category density model*, was proposed to account for the induction of normally distributed categories either with or without error correction or provision of labeled instances. The model was implemented as a computer simulation. Results of five experiments indicated that people could learn category distributions not only without error correction, but without knowledge of the number of categories or even that there were categories to be learned. These and other findings dictated a more general learning model that integrated distributional representations based on both parametric descriptions and stored instances.

In this article we present a new model of category learning and classification based on the acquisition and use of distributional knowledge. This *category density model*, derived from work by Fried (1979), makes the central assumption that the goal of the category learner is to develop a schematic description of the distributions of category exemplars over a feature space. Highly salient features will tend to be encoded initially, although the learner may actively search for less salient features that are more diagnostic of category membership. We assume that the schematic representation is a parametric encoding of the category distribution over the feature dimensions to which the learner is currently attending. Suppose, for example, that a learner is shown a set of exemplars randomly sampled from a category population that is normally distributed over n independent feature dimensions. The density function for such a category can be sufficiently described by a vector of $2n$ parameters—the mean and variance of the population along each feature dimension. The density model assumes that in this example the effective representation of the category distribution will correspond to this parameter vector. This assumption implies that the presented instances will be treated as a sample that can be used to estimate the distributional properties of an indefinitely large population of potential category exemplars.

A parametric representation also implies that there exists a set of statistics for each feature dimension that is sufficient to describe the learner's conception of the category distribution. The types of category distributions that people can encode parametrically may be

---

Experiment 1A was reported at the meeting of the Psychonomic Society in San Antonio, Texas, November 1978, and Experiment 3 was reported at the meeting of the Mathematical Psychology Society in Santa Barbara, California, August 1981. An earlier version of the present article appeared as Cognitive Science Technical Rep. No. 38, University of Michigan, 1982.

This research was supported by National Science Foundation Grant BNS-7904730 to both authors, and a Rackham Faculty Grant from the University of Michigan and National Institute of Mental Health Grant 1-K02-MH00342-03 to K. Holyoak.

We thank the numerous colleagues and students who helped us clarify our ideas in seminars and discussions. Dorrit Billman, Mary Gick, David H. Krantz, Tracy Sherman, Wilson P. Tanner, Jr., and J. E. Keith Smith provided valuable comments on the earliest draft of this paper. William Estes, Don Homa, Kyunghee Koh, Doug Medin, Edward Smith, and Tom Wallsten commented on subsequent drafts. Jack Abraham, Bill Barr, Holly Brewer, Ellen Junn, Roberta Mehoke, Shannon McDonnell, Allan Salmi, and Jan Stern assisted in testing subjects, and John Patterson provided programming assistance.

Requests for reprints should be sent to Lisbeth S. Fried or Keith J. Holyoak, University of Michigan, Human Performance Center, 330 Packard Road, Ann Arbor, Michigan 48104.

small, although this is an open empirical issue. In the present study we will focus on a specific version of the category density model that accounts for learning of multidimensional normally distributed categories. Normal distributions may have particular ecological importance. Basic-level natural categories seem to consist of a dense central region of typical instances, surrounded by sparser regions of atypical instances (Rosch, 1973, 1978; Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). People may therefore expect new categories to be unimodal and to have roughly symmetrical density functions, which may be well approximated by multidimensional normal distributions.

A second major assumption of the category density model is that classification decisions are based on relative likelihood. This decision rule is related to that of signal detectability theory (Swets, Tanner, & Birdsall, 1961), except that it is based on distributional information acquired during category learning rather than on information assumed known a priori. The subjective probability $\Psi_{i,t}$ that the decision maker considers item X to be a member of category C_i on trial t is assumed to be given by Bayes' theorem; that is,

$$\Psi_{i,t} = p_t(C_i|X) = \frac{p_t(X|C_i)p_t(C_i)}{\sum_{m=1}^{k} p_t(X|C_m)p_t(C_m)}, \quad (1)$$

where $p_t(X|C_i)$ is the subjective conditional probability on trial t of item X given category C_i, $p_t(C_i)$ is the subjective prior probability of C_i as of trial t, and k is the number of alternative categories. The model further assumes that the decision maker's probability of making response C_i on trial t given item X is

$$p_t(C_i|X) = \frac{\beta_i \Psi_{i,t}}{\sum_{m=1}^{k} \beta_m \psi_{m,t}}, \quad (2)$$

where β_i is a constant for each category reflecting factors such as an asymmetrical payoff matrix for different classification responses. Note that when the values of $p_t(C_i)$ and β_i are equal for all categories (as will be assumed in all applications of Equations 1 and 2 in the present article because prior probabilities and payoffs were made equal and symmetric in all experiments), it then follows from Equation 2 that the relative frequencies of the alternative categories as responses to item X will be equal to the subjective relative likelihoods of the item X given the alternative categories. We will therefore refer to Equations 1 and 2 jointly as the *relative likelihood decision rule.* This rule will be used as a heuristic device for measuring distributional learning.

A third assumption of the model, related to Bayesian learning theory (Edwards, Lindman, & Savage, 1963), is that category learning is based on a cyclic process of parameter revision. We assume that people expect feature dimensions of perceptual categories to be normally distributed, and that they enter a category learning task with (perhaps very vague) initial opinions about the central tendency and degree of variability of each category on its salient feature dimensions. People then use presented instances to revise these prior expectations. The revised opinions generate expectancies in terms of which the next observation is evaluated.

The density model includes a mechanism by which normal distributions can be learned by revising parameter vectors in response to each successive instance in a set of training exemplars, with minimal reliance on memory for prior instances (discussion follows). In a typical experiment on classification learning, subjects are told the category to which each training instance belongs. Under such conditions the parameter-revision process for learning normal distributions is straightforward. On each trial the feature values of the current instance are used to update the dimension means and variances for the appropriate category, while the occurrence of the category label is used to update the index of the category's frequency. The new parameter values are then saved; the current instance may be incidentally stored in memory, but it plays no further necessary role in learning or classification.

In naturalistic learning situations, unlike standard experimental paradigms, external error feedback may be delayed, unreliable, or completely absent (Bruner, Goodnow, & Austin, 1956, p. 68). Furthermore, there is experimental evidence that people can sometimes learn to classify instances of probabilistic categories without any error correction or pro-

vision of category labels for instances (Edmonds & Evans, 1966; Fried, 1979), although learning is not always entirely successful under such conditions (Evans & Arnoult, 1967; see also Bersted, Brown, & Evans, 1969; Tracy & Evans, 1967). The possibility of learning without external feedback is also suggested by E. Gibson's theory of perceptual learning (1953, 1969; Gibson & Gibson, 1955). Under certain conditions, parameter revision can be used to learn category distributions even in the absence of error feedback. The task of learning distributions without feedback can be modeled as a problem of decomposing the overall mixture density of the presented instances into its component densities. This decomposition can be accomplished by parameter-revision procedures for mixtures of normal densities if the learner knows the number of categories present in the mixture (Duda & Hart, 1973; Fried, 1979).

Category Density Model and Its Simulation

A version of the category density model for normal distributions was implemented as a FORTRAN program.[1] The program estimates the distributional parameters for categories defined by independent, multidimensional normal distributions (i.e., a mean and variance for each dimension of each category, and a frequency parameter for each category). The program can learn these parameters with or without information about the category membership of training exemplars, and was used to validate the qualitative predictions outlined below.

The learning sequence for the FORTRAN program consists of randomly intermixed n-dimensional stimuli ($n \leq 5$) drawn from k categories ($k \leq 5$). Each category is defined by normal distributions of values over each of the dimensions. The dimensions are statistically independent. The category distributions are specified by providing the program with a mean and variance for each dimension of each category, and a relative frequency for each category. Each stimulus is thus represented by a vector of numbers, with each number representing a value on a dimension. The number of categories and dimensions is specified. Runs used to validate the qualitative predictions tested in the present paper (described later) used two 2-dimensional categories.

The learning process operates either with knowledge of the category membership of the training instances (feedback condition) or without such knowledge (no-feedback condition). In either case learning involves two stages: formation of initial parameter estimates, followed by iterative revision of them. In the feedback condition, the dimension value of the first instance of each category C_i is used as the initial mean $M_{i,j}$ for the ith category's jth dimension. The initial value of the frequency N_i of this category is then 1. The dimension variances $V_{i,j}$ for category C_i are initialized when the second observation for that category is obtained. The initial variance estimate is a weighted average of the sample variance of the two observations and an arbitrary large value that represents the vagueness of the learner's prior opinion about the distribution of category i on dimension j.

In the no-feedback condition, initial parameter estimates are formed after accumulating the first s instances, where s ($\geq k$) is the size of a short-term memory buffer (set at 6 in the runs reported later). The s observations are represented as points in an n-dimensional space, and a clustering algorithm uses the Euclidean distances between the points to divide the observations into k groups. The algorithm used is based on the *centroid* method (Everitt, 1974, pp. 12–14). The distance between two groups is defined as the distance between their centroids, where the centroid is the mean of the coordinate values for the items in a group. Initially each of the s instances is defined as a group with one member. The two closest groups are then merged and replaced by the coordinates of their centroid. This procedure is iterated until k groups remain. The initial frequency N_i of each category is then set equal to the number of instances in a corresponding cluster, and the initial dimensional mean $M_{i,j}$ for each category is set equal to the mean of the clustered instances on each dimension. The initial variance $V_{i,j}$ of each category on each

[1] The simulation model was programmed by Kyunghee Koh, who also helped formulate the implemented procedure for learning without feedback.

dimension is obtained by pooling an arbitrary large value (as in the feedback condition) with the variance of the cluster. (This pooling procedure has the incidental effect of ensuring that even a category with just one initial member will have a nonzero initial variance.)

After initial parameter estimates are formed, each successive observation is then used to revise the estimates. On trial t, the first step is to determine the probability $\Psi_{i,t}$ that category C_i generated the observed instance X. For the feedback condition these probabilities are as follows:

$$\Psi_{i,t} = \begin{cases} 1, & \text{if } X \text{ is labeled as a member of } C_i \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

For the no-feedback condition $\Psi_{i,t}$ is calculated by inserting the parameter estimates as of trial $t-1$ in a version of Equation 1 of the relative likelihood rule. If an item X is represented by a vector of values on n independent feature dimensions, $x_1, x_2, \ldots x_n$, then Equation 1 can be restated as follows:

$$\Psi_{i,t} = p_t(C_i|x_1, x_2, \ldots x_n) = \frac{p_t(C_i) \prod_{j=1}^{n} p_t(x_j|C_i)}{\sum_{m=1}^{k} p_t(C_m) \prod_{j=1}^{n} p_t(x_j|C_m)}, \quad (4)$$

and $p_t(x_j|C_i)$ can be determined by substituting the current estimates of $M_{i,j}$ and $V_{i,j}$ in the equation for the normal distribution; that is,

$$p_t(x_j|C_i) = \frac{1}{(2\pi V_{i,j,t-1})^{1/2}} \exp\left[-\frac{1}{2V_{i,j,t-1}}(x_j - M_{i,j,t-1})^2\right]. \quad (5)$$

In calculating $p_t(C_i)$, the program allows a bias toward an assumption that the k categories are of equal frequency. The bias is given as a weight $w_f (0 \leq w_f \leq 1)$, which can be interpreted as a measure of the learner's confidence after one observation that the categories are equally likely. The result is given by the following:

$$p_t(C_i) = \frac{w_f(1/k) + [T_{t-1}(1 - w_f)](N_{i,t-1}/T_{t-1})}{w_f + T_{t-1}(1 - w_f)}, \quad (6)$$

where T is the total number of observations,

$$T = \sum_{i=1}^{k} N_i. \quad (7)$$

In Equation 6 the observed relative frequency, $N_{i,t-1}/T_{t-1}$, is weighted by $T_{t-1}(1 - w_f)$. Hence if $w_f \neq 0$ or 1, the impact of the observed relative frequency will increase with the number of observations.

After determining the values of $\Psi_{i,t}$, the next step is to revise the parameters for all categories, with the degree of revision for each category weighted in accord with $\Psi_{i,t}$. For the no-feedback condition $\Psi_{i,t}$ may be fractional. For example, suppose X is twice as likely to have been generated by C_1 than by C_2, given the current parameter estimates. Then the parameters will be revised as if two thirds of an observation (with the dimension values of the new item) had accrued to C_1 and one third of an observation had accrued to C_2. The revision procedures are based on standard equations for revising running frequencies, means, and variances (Raiffa & Schlaifer, 1961), generalized to accommodate fractional observations.

The revised N_i is given by the following:

$$N_{i,t} = N_{i,t-1} + \Psi_{i,t}, \quad (8)$$

and the revised $M_{i,j}$ is as follows:

$$M_{i,j,t} = (N_{i,t-1}M_{i,j,t-1} + \Psi_{i,t}x_j)/N_{i,t}. \quad (9)$$

In updating $V_{i,j}$, the program allows a bias toward an assumption that the variances for any given dimension are equal across all categories. $V_{i,j,t}$ is a weighted average of the estimated variance of the individual category C_i on dimension j, $IV_{i,j,t}$, and of the variance pooled over all categories, $PV_{j,t}$, with relative weights determined by a parameter w_v, defined analogously to w_f. The revision of $V_{i,j}$ proceeds by first calculating the individual variance:

$$IV_{i,j,t} = [(N_{i,t-1} - 1)V_{i,j,t-1} + N_{i,t-1}(M_{i,j,t-1} - M_{i,j,t})^2 + \Psi_{i,t}(x_j - M_{i,j,t})^2]/(N_{i,t} - 1), \quad (10)$$

and the pooled variance:

$$PV_{j,t} = \frac{\sum_{i=1}^{k} (N_{i,t} - 1)IV_{i,j,t}}{T_t - k}. \quad (11)$$

The revised variance is then given by the weighted average of $IV_{i,j,t}$ and $PV_{j,t}$,

$$V_{i,j,t} = \frac{w_v PV_{j,t} + T_t(1 - w_v)IV_{i,j,t}}{w_v + T_t(1 - w_v)}. \quad (12)$$

The program can use its current parameter estimates to classify items in accord with the relative likelihood rule. The learning phase proceeds either until some criterion is reached (e.g., 10 correct classifications in a row), or a fixed number of observations have been presented. A transfer phase is then simulated, in which the program uses its final parameter estimates to classify transfer items drawn from very broad distributions around the multidimensional means of the categories presented during the learning phase. An *other* response can optionally be allowed, in which case the program assumes that all instances have a fixed likelihood (a parameter that is specified) of being generated by an *other* category. The program thus treats *other* as a category with a specified uniform distribution, so that the relative likelihood that an item is drawn from the *other* category can be calculated using the relative likelihood rule.

Predictions of the Density Model

The category density model as described in the simulation generates a variety of qualitative predictions. The major predictions tested in the present experiments can be divided into two groups: those that concern learning and those that concern transfer performance.

Predictions concerning learning. The following predictions about learning normally distributed categories can be derived from the density model.

L1. When subjects know the number of categories to be learned, and correctly assume that the distributions are normal, learning is possible even without error feedback or instances labeled with respect to category membership.

L2. Labeled instances will facilitate convergence on accurate estimates of distributional parameters, relative to a no-feedback condition. Feedback enables the learner to use each observation to revise only the correct category distribution, rather than apportioning its value across all categories.

L3. Low-variability categories can be learned with fewer observations than required to learn high-variability categories with the same means.

L4. Learning will be facilitated if the learner knows the number of categories to be learned. Indeed, knowledge of number of categories is an essential prerequisite for the parameter-revision procedure described earlier.

Predictions concerning transfer. The following predictions involve transfer performance after learning has taken place.

T1. Classification performance will be in accord with the relative likelihood rule (e.g., the probability of classifying a novel instance as a member of a category will be directly proportional to its subjective likelihood of being generated by the category distribution).

T2. When a random or *other* alternative is available at transfer, exemplars far from the mean (prototype) of a learned category will more likely be classified as members of that category if the variability of the learned categories is high. In contrast, such exemplars will more often be classified as *other* when the variability of the learned categories is low. This prediction follows from the fact that the likelihood that a category will generate atypical exemplars is greater if the variance of the category is relatively high. In situations in which *other* responses are classified as errors, the percentage correct will be higher for groups trained on high-variability instances.

T3. The above advantage of learning high-variability rather than low-variability categories in classifying exemplars far from the prototype will not be obtained in the absence of an *other* alternative at transfer.

T4. If subjects learn two equally probable categories of unequal variability, they will tend to classify more items into the high-variability category at transfer, including some items that are closer to the mean of the low-variability category but more likely to have been generated by the high-variability one.

Sample of Performance by the Simulation Program

Because the stimuli used in the experiments reported later were complex forms for which the psychological features encoded by subjects were not known, the simulation program cannot generate precise quantitative predictions for our experiments. However, the general model can be applied even if the psychological features are unknown and different subjects encode stimuli in terms of different features, as long as subjects' feature sets can be approximated by normal densities. The qualitative predictions outlined earlier hold regardless of the specific nature or number of features used by subjects. As an illustration of some of our qualitative predictions, we will report the results of some sample runs of the program. These test runs used two 2-dimensional categories, defined by bivariate normal distributions with means (3, 6) and (6, 3), respectively, in arbitrary units. In the simulated low-variability condition the variances of both categories on each dimension were set equal to 1 (so that $d' = 3$ on each dimension), and in the high-variability condition the variances were set equal to 4 ($d' = 1.5$). The arbitrary value used in initiating the variance estimates was 20. In the no-feedback condition parameters were initialized after clustering the first six items. The values of w_f and w_v were set equal to 0.9 and 0.1, respectively. Ten simulated subjects were used in each run.

As a measure of rate of learning, the mean number of trials required to reach a criterion of 10 correct responses in a row was measured. A maximum of 300 learning trials were allowed. The mean number of trials to criterion was 32 for the low-variability, feedback condition; 62 for the low-variability, no-feedback condition; 74 for the high-variability, feedback condition; and 146 for the high-variability, no-feedback condition. These results illustrate Predictions L2 (advantage of labeled instances) and L3 (advantage of low-variability training).

Other runs were used to simulate transfer performance after a fixed number of learning trials (100). The first set of runs, presented in Table 1 (top half), included an *other* alternative at transfer. The subjective likelihood that any item was an *other* was specified to be .002. In fact, all transfer items were drawn from broad distributions around the prototypes of the two categories presented during the learning phase. Table 1 (top half) features both the obtained mean percentage correct and mean percentage *other* responses as a function of the Euclidean distance of transfer items from their generating prototype (as measured in the same arbitrary units). For all learning conditions the percentage correct decreased and the percentage *other* increased with distance from the prototype, exemplifying Prediction T1. The predicted greater percentage correct and lesser percentage *other* responses at high distances for groups trained on high- rather than low-variability exemplars (Prediction T2) was also apparent. Presented in the bottom half of Table 1 are runs in which no *other* alternative was allowed. Here the advantage of high-variability

Table 1
Transfer Performance With and Without Availability of an Other *Category*

	Distance from prototype				
Training condition	1	2	3	4	5
With *other* category					
Low variability					
Feedback					
% correct	.99	.93	.72	.31	.03
% *other*	.01	.04	.15	.45	.72
No feedback					
% correct	.97	.84	.70	.46	.19
% *other*	.01	.04	.12	.29	.54
High variability					
Feedback					
% correct	.88	.82	.71	.61	.49
% *other*	.01	.06	.08	.13	.25
No feedback					
% correct	.82	.76	.68	.58	.49
% *other*	.02	.06	.09	.14	.26
Without *other* category					
Low variability					
Feedback (% correct)	.99	.96	.85	.73	.67
No feedback (% correct)	.98	.88	.82	.71	.67
High variability					
Feedback (% correct)	.89	.86	.77	.72	.69
No feedback (% correct)	.85	.81	.74	.68	.71

Note. Distance from the prototype is measured in arbitrary units.

training in percentage correct for items far from the prototype was eliminated (Prediction T3). As in the runs that included the *other* alternative, low-variability training yielded higher percentage correct for items close to the prototype, because the subjective likelihood of such items is relatively high when the estimated category variance is relatively low. In the runs in Table 1 the feedback conditions tended to yield higher percentage correct than the no-feedback conditions, indicating that the latter had not achieved asymptotic learning after 100 training trials.

In general, no-feedback learning by the simulation program is more variable than learning with feedback, largely because the former is more sensitive to the accuracy of early parameter estimates. We have run the program with s set at 2, thus simply using the first two items as the initial estimates of the means of the two categories. We have also run the program setting w_f and w_v to 0, thus removing any biases toward the assumptions of equal category frequencies and equal category variances. In both cases learning still takes place, although somewhat more slowly than with the parameter values used in the runs presented earlier. However, if both changes are made (i.e., $s = 2$ and $w_f = w_v = 0$), virtually no learning takes place in the high-variability condition, although some learning is still possible in the low-variability condition.

Comparison With Previous Models

With respect to its representational assumptions, the category density model is most similar to prototype models (Posner & Keele, 1968; Reed, 1972). Like prototype models, the density model assumes that a true induction process takes place: The learner goes beyond the sampled instances to infer category-level information. Furthermore, both types of models assume that this category-level information is represented parametrically. But whereas a simple prototype represents the central tendencies of the category instances on their feature dimensions, parameters can also be used to represent the variability of a distribution, as discussed earlier, and perhaps other distributional properties as well (e.g., skewness). Like a prototype, however, representations postulated by the density model can be characterized as schemata (Attneave, 1957; Oldfield, 1954). Indeed, for the special case of multidimensional normal distributions (the focus of the present article), the density model is equivalent to a model that assumes the learner abstracts the prototype plus variance for each category.

However, in terms of its decision rule for classification, the density model is more similar to feature frequency (e.g., Hayes-Roth & Hayes-Roth, 1977) and instance models (Medin & Schaffer, 1978) than to simple prototype models.[2] Unlike the closest prototype decision rule, the relative likelihood rule is sensitive to category variability and other factors that influence the degree of overlap among exemplars of alternative categories.

Unlike other classification models, the category density model provides an explicit mechanism by which categories can, under some conditions, be learned without any external, instance-specific feedback. Regardless of whether instances are being averaged to form prototypes, used to tabulate feature frequencies, or simply stored in memory, other models have tacitly assumed that error feedback is critical in category learning, since the learner must know the category to which an instance belongs in order to use it to modify the appropriate category representation. The

---

[2] Medin and Schaffer (1978) pointed out that the distance and cue validity decision rules proposed in the classification literature (Hayes-Roth & Hayes-Roth, 1977; Reed, 1972) are *independent cue models*, that is, rules that assume the information entering into category judgments is based on an additive combination of the information derived from the component feature dimensions. It follows from the nature of probability that the relative likelihood rule is not an independent cue model; rather, as Equation 4 makes clear, it implies a multiplicative combination of dimensional information. In this respect the relative likelihood rule resembles the context model proposed by Medin and Schaffer. But whereas the latter model assumes that dimensional information is combined to calculate a measure of instance-to-instance similarity, the relative likelihood rule assumes that such information is combined to calculate the conditional probability of an instance given a particular category.

Wallsten (1976) presented evidence indicating that the impact of a dimension value on subjects' decisions depends on the dimension's salience as well as on its diagnosticity. Salience could be represented by different weights associated with each dimension.

generalization procedures that have been proposed to reduce the storage requirements of feature frequency models (Anderson, Kline, & Beasley, 1979; Patterson, 1979) depend upon provision of error correction.

Experiments 1A and 1B

Experiment 1A focused on tests of transfer predictions T1–T3. Prediction T2 is in fact supported by Posner and Keele's (1968) finding that greater variability of training exemplars produced slower initial learning, but more accurate transfer performance in a classification task. The bulk of the transfer errors in the Posner and Keele study (Experiment 2) were made by the low-variability group, and involved the erroneous classification of the highly distorted exemplars of meaningful prototypes into a category based on a random dot configuration. The random-prototype category might have been viewed by subjects as a flat, rectangular distribution with a very wide range of acceptability on the feature dimensions.

Recall that in terms of the category density model, if category prototypes are kept constant, an increase in category variability will result in reduced discriminability between categories (measured in d'), making learning more difficult (Prediction L3). In a subsequent transfer task, however, subjects trained on two high-variability categories will view highly distorted exemplars as relatively likely to have been generated by the category, and so will classify them correctly. In contrast, those trained on two low-variability categories will not view highly distorted exemplars as likely to have been drawn from the category. Consequently, if a random alternative category is available, they will tend to classify those items as *random*. Thus when this alternative category is available, some highly distorted items are predicted to be classified differently depending only on the variability of the training items. This prediction of the category density model is not accounted for by a distance to prototype decision rule, because the items are the same distance from the prototypes in both groups. The above prediction was supported in previous research when feedback was provided during training (Fried, 1979), but has not been previously tested when learning takes place without trial-by-trial error correction. The predicted difference between the two variability groups depends on the presence of an alternative random category, because without such an alternative category the relative likelihood rule predicts no advantage for a group that learned relatively high-variability categories (Prediction T3). This latter prediction was investigated in Experiment 1B, in which the random alternative or *other* category was removed.

Method

Stimuli. The choice of stimuli was guided by several criteria. We wanted stimuli: (a) that would allow an essentially infinite population of category exemplars; (b) for which objective measures of both distance between any two items and of the likelihood of any item given any category could be calculated; (c) for which category variability could be systematically manipulated; (d) with a relatively realistic degree of perceptual complexity; and (e) that could be generated and displayed under computer control. These criteria were met by visual grid patterns of the sort depicted in Figure 1. The categories to be learned in Experiments 1A and 1B consisted of two sets of such visual patterns, each composed of instances derived from a standard pattern by means of a probabilistic distortion rule. All patterns consisted of light and dark cells in a 10×10 grid displayed on a computer-controlled TV screen. The two standard patterns, shown in Figure 1, were created using a modification of the method for generating figures specified by Attneave and Arnoult (1956; see Fried, 1979). The standards were adjusted so that 50 cells in each were dark and 50 were light. In addition, 50 cells overlapped between the two standards. Distortions were generated on-line by changing each cell of the standard from light to dark or vice versa with some specified distortion probability, p. Increasing the distortion probability in the range .00 to .50 increases the variability of the distribution of instances, defined in terms of number of cells changed from the standard. These distributions were binomial approximations to the normal, with the mean number of cells changed equal to $100p$ and variance equal to $100p(1-p)$, where p is the distortion probability and 100 is the number of cells in each pattern. A total of 2^{100} patterns were possible. The likelihood of any particular pattern was $p^N(1-p)^{100-N}$, where N is the number of cells distinguishing the pattern from the generating standard. The categories were thus distinguishable only in their likelihood of generating each of the 2^{100} possible patterns. The standard itself was the most likely individual pattern of each category, but since its probability was nonetheless vanishingly small, $(1-p)^{100}$, it never actually was presented during learning trials in our experiments. As in Fried (1979), distortion probabilities of .07 and .15 were used for the low- and high-variability learning conditions, respectively, in all experiments to be reported. Figure 1 illustrates .07 and .15 distortions of each of the standard

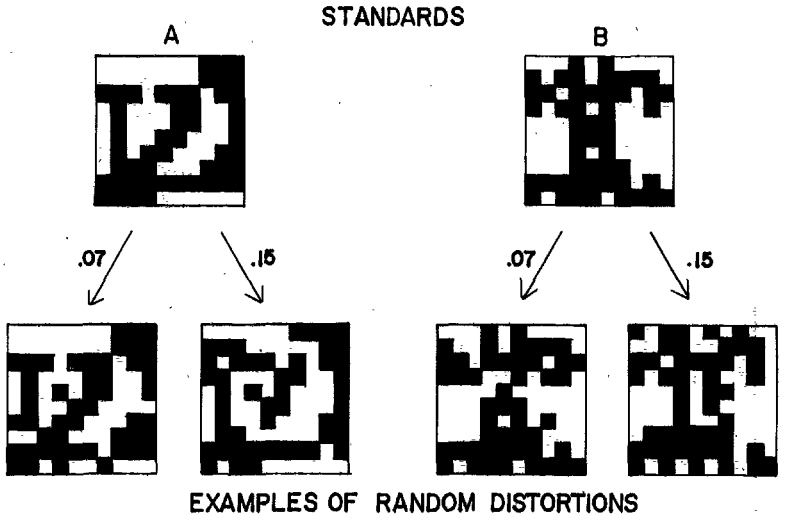

Figure 1. The two standards used in Experiment 1, and examples of .07 and .15 distortions of each.

patterns. In these illustrative examples the number of changed cells is set equal to exactly $100p$.[3]

Design and procedure (Experiment 1A). Subjects were randomly assigned to one of four conditions, defined by the 2 × 2 (Variability × Feedback) factorial combination of low versus high variability of training exemplars (.07 and .15 distortion probabilities, respectively) and presence versus absence of item-specific error feedback. All subjects were told that they would see a mixture of geometric patterns designed by two artists, named Smith and Wilson, and that they would have to distinguish the work of Smith from that of Wilson. The two standards shown in Figure 1 were used for all subjects, but each subject saw a different random sample of distortions, and no subject saw the standard.

The patterns were displayed on a TV screen controlled by an IBM 1800 computer. During the learning phase all subjects classified a series of patterns into two categories by pressing one of two response keys. A maximum of 7 s was allowed to make each response. Subjects receiving instance-specific error feedback were told whether or not they were correct immediately after each response, thus effectively labeling the instances with respect to category membership. Subjects not receiving instance-specific error feedback did not receive such information. However, all the subjects were told the number of correct and incorrect responses they had made for each block of 10 trials. All subjects in Experiment 1A thus received general information about whether their classification accuracy was improving. However, the nonspecific feedback subjects were never told the category to which any particular instance belonged.

Subjects received a bonus of 1 cent for each correct answer and were fined 1 cent for each error. In addition, their pay decreased 1 cent for every 10 trials they required to learn the categories. This learning phase continued until subjects responded correctly 10 times in a row, or reached a maximum of 200 trials. The response key assigned to a particular category by subjects in the nonspecific feedback condition was necessarily arbitrary. Their responses were scored as correct in the manner that maximized their score over all learning trials.

After completion of the learning phase, all subjects received an additional 100 transfer trials, without error correction. Subjects were told that the patterns would include new Wilsons and new Smiths, but also an unspecified number of patterns designed by other people. In fact, there were no true *others;* all the patterns presented during the transfer phase were actually derived with equal frequencies from the two original standards. Equal proportions of the transfer items were created at each of four distortion probabilities: .10, .20, .30, and .40. The transfer patterns therefore included instances at higher levels of distortion than those that were presented during learning, even in the high-variability learning conditions. On each trial subjects pressed one of three response keys to classify the pattern as a Wilson, a Smith, or an *other.* Subjects in the nonspecific feedback condition had to maintain the same response-key assignments as they had established during the learning phase. A maximum of 5 s was allowed to make a response. Subjects received a 1 cent bonus for each correct classification, lost 1 cent for classifying a Wilson as a Smith or vice versa, and neither won nor lost money for *other* responses.

Forty-five University of Michigan undergraduates served as paid subjects.

Design and procedure (Experiment 1B). The design and procedure used in Experiment 1B were identical to

[3] We assume that on average there is a monotonic relationship between number of cells changed (an objective city-block measure of distance from the standard) and psychological distance, for the range from 0 to 50 changed cells. A distortion probability of .50 (expected number of changed cells equal to 50) yields patterns statistically unrelated to the generating standard.

those used in Experiment 1A, except for two changes. First, the summary information provided to subjects in Experiment 1A after every 10 learning trials was eliminated. Subjects in the resulting no-feedback condition, unlike those in the nonspecific feedback condition of Experiment 1A, therefore received no information about the degree to which their classification accuracy was improving. The feedback condition in Experiment 1B received the same item-specific feedback as did the comparable condition of Experiment 1A. Second, subjects were told that all transfer patterns were either Smiths or Wilsons, and were required to classify each pattern into one of those two categories; that is, no third *other* alternative was available at transfer.

Forty-five University of Michigan undergraduates served as paid subjects.

Results and Discussion (Experiment 1A)

Learning phase. Of the 45 subjects tested, 8 had not reached criterion within the maximum 200 trials. As expected (Prediction L3), all of these subjects were in the high-variability condition: 2 subjects who received specific error feedback and 6 who received nonspecific feedback. The mean number of learning trials for all subjects was 39 for the low-variability, specific feedback condition; 51 for the low-variability, nonspecific feedback condition; 108 for the high-variability, specific feedback condition; and 141 for the high-variability, nonspecific feedback condition. The learning-trials measure proved to be highly variable (MS_E = 2,692), reducing statistical power. Nevertheless, as in previous research (Fried, 1979; Posner & Keele, 1968), subjects in the high-variability conditions required significantly more learning trials to reach criterion, $F(1, 41) = 25.9$, $p < .001$, in accord with Prediction L3. The nonspecific feedback conditions tended to require more learning trials than the specific feedback conditions; however, this trend was not significant, $F(1, 41) = 2.09$, $p < .20$. The fact that learning was possible without specific feedback provides support for Prediction L1.

Transfer phase. Of the 8 subjects who had not reached criterion within 200 trials, 3 (all in the nonspecific feedback condition) responded with accuracy levels significantly above chance during the transfer task. Since we were interested in transfer performance after at least some learning had taken place, data from the other 5 subjects were excluded from transfer analyses. The remaining 40 subjects included 10 in each of the four conditions.

The relative likelihood rule predicts that if subjects had learned the mean (or generating standard) of each category, the percentage of patterns called *other* would increase as a function of distance from the standards. The rule also predicts that if subjects had learned category variability, those in the high-variability conditions would classify fewer patterns far from the standard as *other* (Prediction T2). Furthermore, this pattern should obtain regardless of whether specific error feedback is given (Prediction T1). Presented in Figure 2 is the percentage of patterns called *other* as a function of the number of cells by which the distorted pattern differed from the standard used to generate it (averaging over blocks of 10 cells).[4] Percentage of patterns called *other* increased with increasing distance from the standard for all groups, $F(4, 144) = 23.0$, $p < .001$.[5] Subjects in the low-variability conditions tended to make more *other* responses overall than did those in the high-variability conditions, $F(1, 36) = 3.76$, $p < .10$. More importantly, this difference became greater as the distance from the transfer pattern to the standard increased, $t(144) = 2.60$, $p < .02$, by a bilinear trend test. Furthermore, lack of item-specific error feedback did not affect the overall pattern of results, $F(1, 36) = 1.78$, $p > .25$, and produced no significant interactions.

The relative likelihood rule predicts that the percentage called *other* should be a decreasing function of relative likelihood; that is, the greater the relative likelihood the greater the probability that the item will be classified into the appropriate category, and the lower the probability that the pattern will be put into an erroneous category, such as *other.* Presented in Figure 3 is the percentage of patterns called *other* as a function of the natural logarithm of the likelihood ratio in favor of the correct category, $p(X|S_C)/p(X|S_A)$, where S_C and S_A are the correct and alternative standards, respectively. The data points plotted in Figure 3 were obtained by averaging over blocks of approximately 20 log units of likelihood ratio. Likelihood ratio reaches higher levels for the

[4] About 3% of the transfer patterns were actually closer to the alternative standard than to the standard used to generate them. However, the pattern of results was unchanged when the closer standard was scored as correct.

[5] Throughout this article, all analyses of variance on proportions were performed after applying an arc sine transformation.

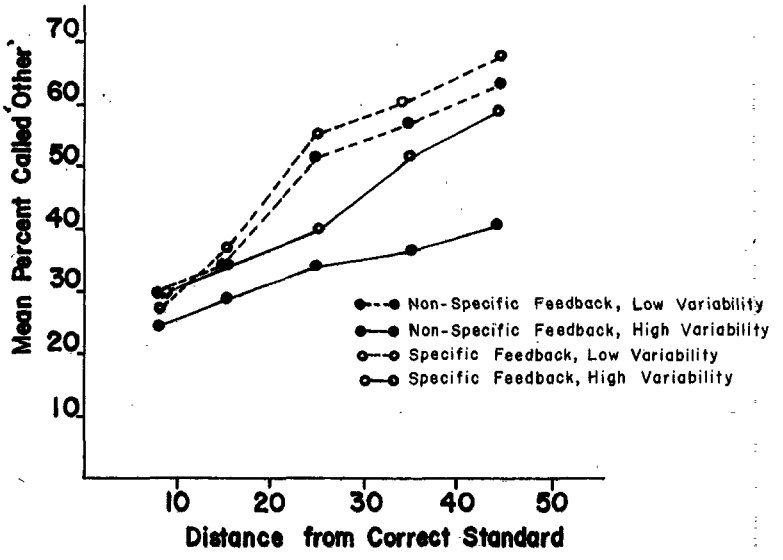

Figure 2. Percentage of transfer patterns called *other* as a function of distance from the correct standard (Experiment 1A).

low-variability conditions, since low-variability distributions make it more likely that patterns close to the standard will be generated, and also make it less likely that such instances could have been derived from the alternative standard. This analysis shows that the per-

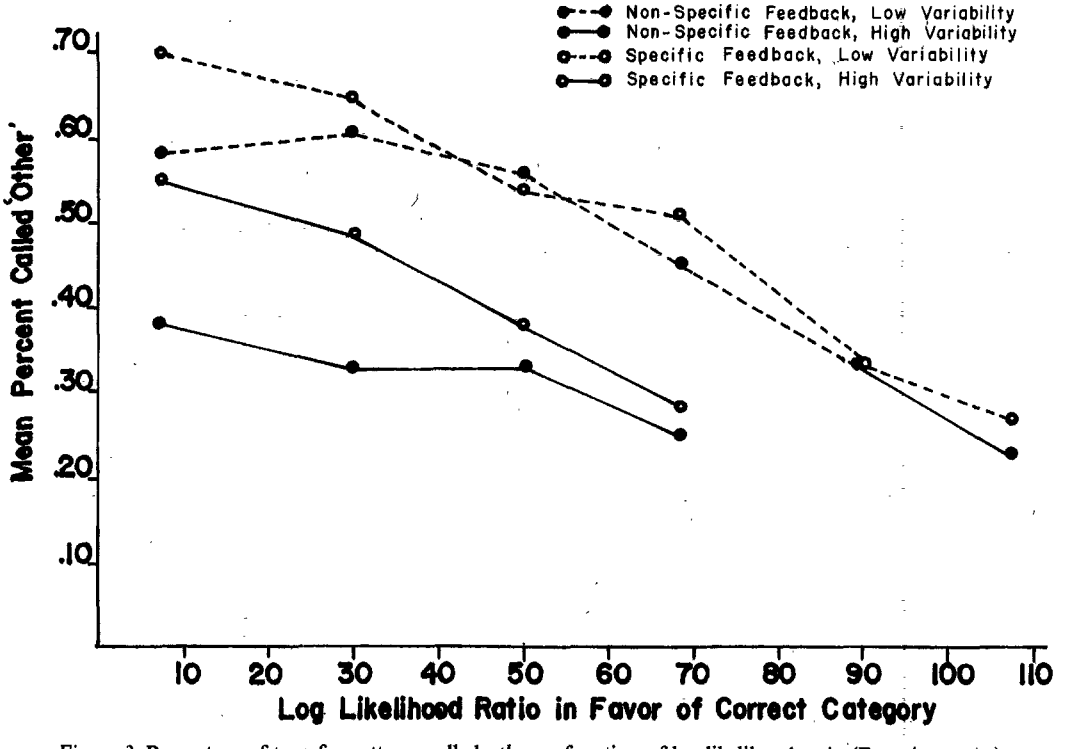

Figure 3. Percentage of transfer patterns called *other* as function of log likelihood ratio (Experiment 1A).

centage called *other* decreased monotonically across the four levels of relative likelihood at which all conditions can be compared, $F(3, 144) = 13.5$, $p < .001$. The relative likelihood rule predicts that among patterns equated on likelihood ratio with respect to the two learned categories, low-variability training will produce a greater proportion of *other* responses (since in this case the *other* category will often seem more likely to have generated the item than either of the two low-variability categories). This prediction was supported, $F(1, 36) = 16.1$, $p < .001$. The effect of specificity of error feedback did not approach significance. (Since for equal-variance categories log likelihood is highly correlated with distance from the prototype, in subsequent cases we will report only one measure.)

Because the high-variability conditions produced fewer *other* responses, did they then produce more correct responses? Presented in Figure 4 is the percentage correct for all four groups as a function of likelihood ratio. Accuracy increased with increasing likelihood ratio, $F(3, 108) = 25.8$, $p < .001$. Provision of specific error feedback during training did not produce an advantage in percentage correct ($F < 1$). However, variability of the training instances significantly affected transfer accuracy. The high-variability group was significantly more accurate for patterns equated on likelihood ratio, $F(1, 36) = 16.5$, $p < .001$. It is apparent from inspection of Figure 4 that for any given level of likelihood ratio the low-variability group was more likely to call a pattern *other* (and thus have it counted as an error), whereas the high-variability group was more likely to classify it correctly. This pattern supports Prediction T2.

Results and Discussion (Experiment 1B)

Learning phase. Five of the 45 subjects failed to reach the criterion of 10 correct trials in a row within the maximum allotment of 200 trials. All of these were in the high-variability conditions, with 3 in the no-feedback condition, and 2 in the feedback condition.

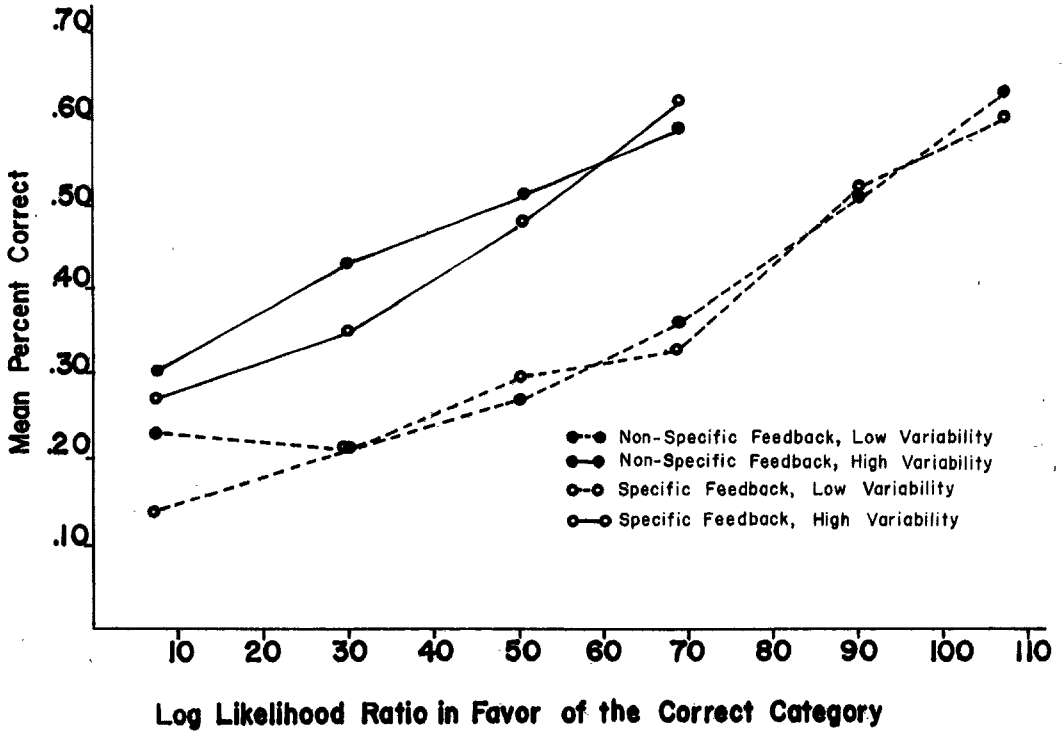

Figure 4. Percentage of transfer patterns classified correctly as a function of log likelihood ratio (Experiment 1A).

As in Experiment 1A, the learning trials measure was highly variable (MS_E = 3,532). Once again subjects in the high-variability conditions required more learning trials than did those in the low-variability conditions (M = 91 trials and M = 43 trials, respectively), $F(1, 41) = 7.17$, $p < .025$. Although the trend favored the subjects who received error feedback over those who did not (65 trials vs. 75 trials), this difference did not approach significance ($F < 1$). Weak statistical power suggests caution in accepting the null hypothesis; however, it is clear that error feedback was not a necessary condition for category learning in the present task. It can also be concluded that the nonspecific feedback in Experiment 1A was not instrumental in producing learning in that condition.

Transfer phase. The 5 subjects who failed to reach the learning criterion were excluded from analyses of transfer performance. Presented in Figure 5 is the percentage correct classification at transfer as a function of log likelihood ratio in favor of the correct category. As in Experiment 1A, percentage correct increased as a function of likelihood ratio, $F(3, 108) = 11.1$, $p < .001$. Also as in Experiment 1A, overall percentage correct was not influenced by error feedback ($F < 1$). However, level of feedback did interact significantly with variability of the training set, $F(1, 36) = 8.20$, $p < .01$. As is apparent in Figure 5, this interaction was mainly due to the especially accurate performance of the high-variability feedback condition at its two highest levels of likelihood ratio. Error feedback did not have a significant effect for the low-variability condition, $F(1, 18) = 1.41$, $p > .25$.

As predicted by the relative likelihood rule (Prediction T3), and in sharp contrast to Experiment 1A, high-variability training instances did not improve transfer accuracy when an *other* category was not available. When only patterns equated on likelihood ratio were considered (thus excluding patterns at the two highest levels of likelihood ratio, experienced only by the low-variability groups), variability had no significant effect, $F(1, 36) =$

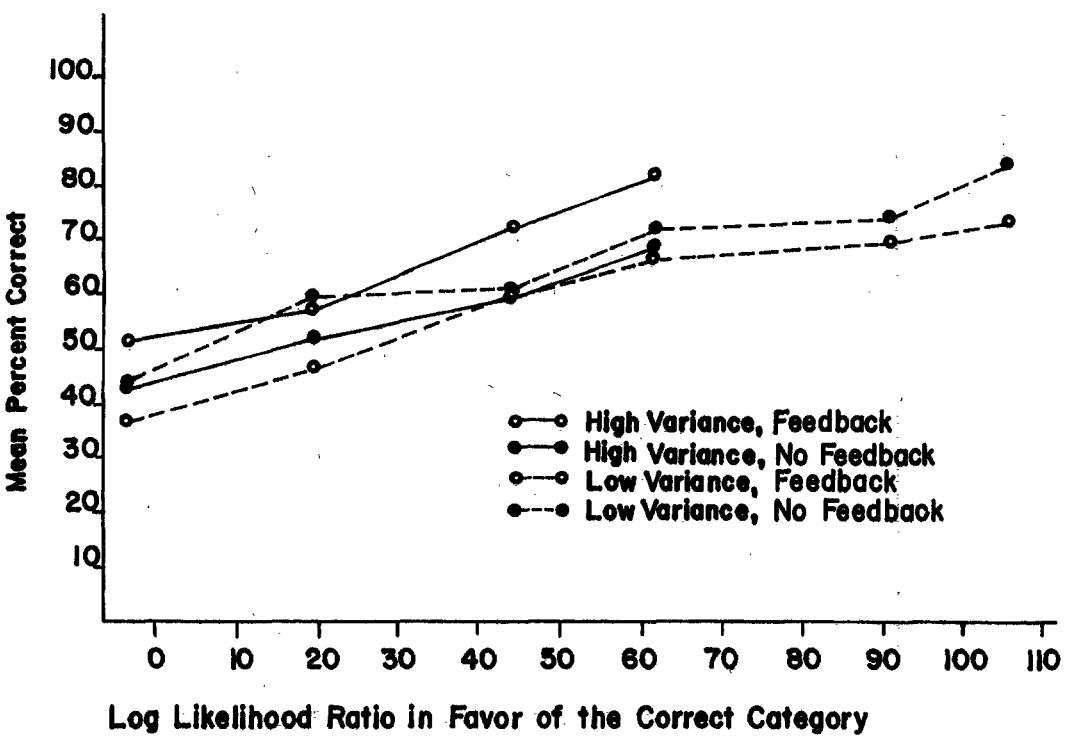

Figure 5. Percentage of transfer patterns classified correctly as a function of log likelihood ratio (Experiment 1B).

1.68, $p > .20$, as the relative likelihood rule predicts. The results of Experiment 1B indicate that high-variability training does not always produce more accurate transfer performance for exemplars far from the standard, nor does exposure to a larger set of training items (because in both Experiments 1A and 1B, subjects in the high-variability condition received more instances before reaching the learning criterion). Rather, the existence of a random or *other* alternative at transfer (as in Experiment 1A) is critical to producing an advantage of high-variability training.

Experiment 2

In Experiments 1A and 1B variability was manipulated across different groups of subjects. The relative likelihood rule can also be tested by training a single group of subjects with two equally probable categories of unequal variability. If subjects learn the category distributions, the rule predicts that they will tend to classify more patterns into the high-variability category, even though exemplars of the two categories are equally likely a priori (Prediction T4). In particular, some patterns that are physically more similar to the standard of the low-variability category will be more likely to be generated by the high-variability category. If subjects learn the distributions and follow the relative likelihood decision rule, they should tend to classify such patterns into the high-variability category. In contrast, if subjects employ a closest prototype rule, based on some monotonic function of physical distance, they will tend to classify such patterns into the low-variability category.

Method

Two new standard patterns were used in Experiment 2. These were constructed in the same manner as the standards used in Experiments 1A and 1B except that the new standards were the same in only 40 (rather than 50) of the 100 cells. During the learning phase the instances of one standard were derived by a .07 distortion probability (the low-variability category), and the instances of the other standard were derived by a .15 distortion probability (the high-variability category). Assignment of the two standards to distortion level was counterbalanced across subjects.

One other major change was introduced in the learning phase of Experiment 2. The results of Experiment 1 and earlier studies indicate that high-variability categories are harder to learn than low-variability categories. If subjects simply had to discriminate instances of a low- versus a high-variability category, they could do so by learning only the low-variability category, and then assigning all remaining instances to the high-variability category. Subjects would thus never need to acquire a clear conception of the high-variability category. Under these conditions the high-variability category would presumably be treated as an *other* alternative during transfer. As a result, high-level distortions of the low-variability category would tend to be classified as members of the high-variability category, but not for the theoretically relevant reason.

It was therefore important to ensure that subjects would actually learn the distribution of the high-variability category during the learning phase, rather than treat it as a vague *other* category. Accordingly, the training set consisted of equal numbers of .07 distortions of the standard for the low-variability category, .15 distortions of the standard for the high-variability category, and .50 distortions of both standards. Instances created by a .50 distortion probability are truly random (i.e., they are statistically independent of the generating standard), and thus constituted a true *other* category. To reach the learning criterion of 10 successive correct trials, subjects therefore had to learn not only to discriminate instances of the low- versus high-variability categories but also to discriminate instances of the high-variability category from *others*.

Subjects were required to make a decision for each pattern within 7 s. Half the subjects received error correction on each trial and half never received error correction. For those trained without error feedback the assignment of category label (*Smith* or *Wilson*) to the two standard categories was arbitrary; however, the category to be labeled *other* was nonarbitrary.

At the beginning of the transfer phase, subjects were told that they would see new works by Wilson and Smith, not necessarily in equal numbers, and that no works by other people would be included. They had to classify each pattern as either a Wilson or a Smith, and thus were forced to discriminate solely between the high-variability and the low-variability categories. Chance accuracy was therefore 50%. The transfer set consisted of a total of 100 patterns, half derived from each standard, with equal numbers generated at distortion probabilities of .15, .25, .30, and .35. As in Experiments 1A and 1B, the assignments of category labels to instances in the transfer phase had to be the same as those established during learning (i.e., for subjects in the no-feedback condition, assignments were not arbitrary at transfer).

Twenty-five University of Michigan students served as paid subjects.

Results and Discussion

Learning phase. The mean number of learning trials was virtually identical for subjects who received feedback and those who did not (although medians favored the feedback subjects, 119 vs. 149). However, consistent with Prediction L2, 5 subjects in the no-feedback condition (vs. none in the feedback condition) failed to reach the learning criterion within 200 trials. In all cases the nonlearners had

difficulty discriminating the high-variability and *other* categories.

Transfer phase. Only data for the 20 subjects who reached the learning criterion were analyzed. An overall picture of transfer performance is provided by Figure 6, in which appears the percentage of items classified into the high-variability category as a function of the log likelihood ratio favoring that category. Percentage classified into the high-variability category was an increasing function of likelihood ratio in favor of that category, $F(5, 90) = 23.0$, $p < .001$. Error feedback did not significantly influence the pattern of results; however, the large percentage of nonfeedback subjects who failed to learn cautions against accepting the null hypothesis. The results also did not differ as a function of which standard was assigned to the high-variability category.

The main concern in Experiment 2 was to determine whether subjects base their classification decisions on likelihood or distance. The relative likelihood rule predicts that more patterns will be classified into the high-variability than the low-variability category, since subjects will have learned a broader density function for the former category. In contrast, a strict distance-to-prototype rule predicts that an equal proportion of instances will be classified into each category, because the distributions of instances around their standards were actually identical for the two categories during the transfer phase. The prediction of the relative likelihood rule was supported, as 61% of the transfer items were placed in the high-variability category. This figure was significantly higher than the 50% predicted by the distance rule, $t(19) = 3.28$, $p < .01$.

A separate analysis was performed for just those items that were closer to the standard of the low-variability category (in terms of changed cells), but more likely to be generated by the high-variability category. These items provide a particularly strong test of whether subjects used a decision rule based on distance or likelihood. If classifications were based on any monotonic function of physical distance that is constant over category variability, these patterns would tend to be placed in the low-variability category; but if classifications were based on any monotonic function of likelihood, these items would tend to be placed in the high-variability category. The prediction of the relative likelihood rule was confirmed; 64% of these critical items, which were physically closer to the low-variance standard, were actually classified into the high-variability category. This percentage was significantly higher than 50%, $t(19) = 2.67$, $p < .02$.

Experiment 3

The purpose of Experiment 3 was to provide a more extensive investigation of the role of

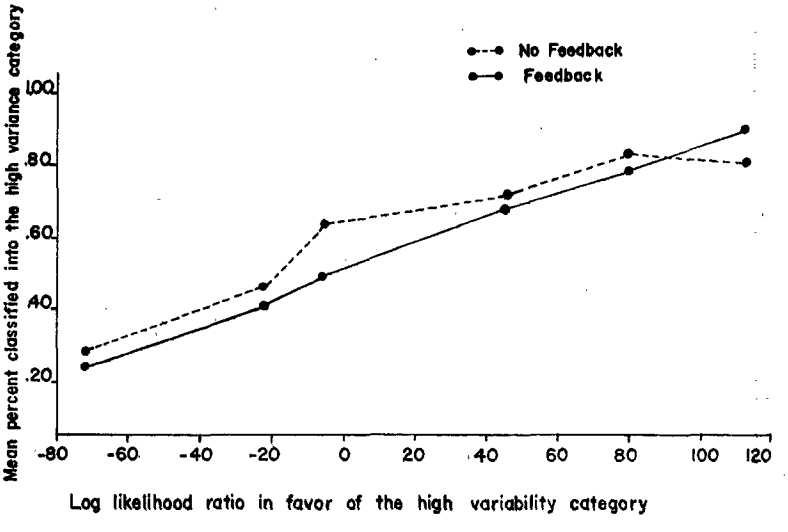

Figure 6. Percentage of transfer patterns classified into the high-variability category (Experiment 2).

labeled instances and other types of supplementary information (Tracy & Evans, 1967) in facilitating category learning, and in particular to determine whether knowledge of the number of categories facilitates learning (Prediction L4). Different groups of subjects received one of four levels of supplementary information. Subjects in the labeled instances condition received a category label (*Wilson* or *Smith*) with each exemplar. Subjects in the number known condition received neither category labels nor error correction, but were told (correctly) that there were two categories to be learned. The first of these two conditions represented complete information, whereas the second was similar to the no-feedback conditions of Experiments 1B and 2 in the amount of information available. Subjects in two additional conditions received still less information about the learning task. Those in the number unknown condition were told they were to try to learn the categories represented in the training set, but were not told how many categories would be present. Finally, those in the observation only condition were not told that their task was to learn categories, nor that categories were present; they were simply told to "pay the utmost attention" to each pattern in the set (cf. Reber & Allen, 1978). Prior to a subsequent transfer task, all subjects were informed that they had seen exemplars drawn from exactly two categories. The category density model predicts that the labeled instances condition will yield superior learning to the number known condition (Prediction L2), and that the latter condition will yield superior learning to the remaining two conditions (Prediction L4). In fact, because the parameter-revision procedure can only operate if the number of categories is known, the model predicts that no learning will take place in the number unknown and observation only conditions (unless subjects in these conditions happen to guess the correct number of categories).

Method

Apparatus and patterns. The patterns were presented on a Hazeltine text terminal controlled by a PDP 11/34 computer. Each pattern was composed of a 10 × 10 grid, in which each cell consisted of two horizontally aligned character spaces. If the cell was defined as black, both character spaces were blank; if it was defined as white, each space was occupied by the rectangular ASCII character 127 rubout. Two standard patterns were generated for each subject. The first standard was created by randomly making each cell black or white with an equal probability. The second standard was derived from the first by switching 50 randomly selected cells from black to white or vice versa. Different standards were randomly generated for each subject within a given condition, whereas across learning conditions subjects were yoked with respect to the standards, training exemplars, and transfer set. As in previous experiments, exemplars were generated by probabilistic distortions of the standards.

Design and procedure. Subjects were randomly assigned to one of eight experimental conditions, defined by the 2 × 4 (Variability × Condition) factorial combination of low- versus high-variability of training exemplars (.07 vs. .15 distortion probabilities), and the four instructional conditions outlined earlier. Following initial instructions, all the experimental subjects participated in a training phase in which they viewed 200 exemplars. These consisted of a random mixture of 100 instances derived from each standard. A major methodological change introduced in Experiment 3 was thus to present subjects with a fixed number of training exemplars, rather than to allow subjects to reach a learning criterion. Using a fixed number of learning trials avoids several methodological problems inherent in a criterion procedure. A criterion measure creates a confounding between learning difficulty and number of training trials. In addition, the exclusion of those subjects who failed to reach the criterion can bias analyses of transfer performance in favor of conditions that are more difficult to learn. Although using a fixed number of learning trials avoids the above problems, it does have a disadvantage of its own. Differences in learning rate among instructional conditions may not be observed if all conditions achieve asymptote within the allotted number of learning trials. This is most likely to occur for groups who receive low-variability training exemplars.

Each pattern was presented for just 2 s. The subject then pressed a key to initiate presentation of the next pattern. Subjects did not make overt classification responses during learning, but simply observed the exemplars. Except for the labeled instances condition, for which a category label, *Wilson* or *Smith*, was written beneath each pattern, all subjects had the same type of observation experience.

After completing the training phase, subjects in the number unknown and observation only conditions were debriefed to find out whether they had noticed that the patterns were drawn from two categories. Subjects in the observation only condition were asked a series of increasingly directive questions: (a) Did you notice anything interesting about the patterns? (b) Did you notice any similarities among them? (c) Did you notice that the patterns fell into different groups of categories? and finally, (d) How many categories do you think the patterns were divided into? The latter two questions were also asked of subjects in the number unknown condition. All subjects were then informed that the number of categories was two.

The subjects then received 100 transfer trials. In the manner of Experiment 1A, subjects were told that the patterns would include new works by Wilson and Smith, as well as an unspecified number of works by other people. The transfer patterns actually consisted of 50 exemplars

from each category, 10 at each of five exact distances from the standard (in terms of number of changed cells): 5, 15, 25, 35, and 45. Each pattern was presented for a maximum of 7 s, and subjects pressed one of three keys to classify the pattern as a Wilson, a Smith, or an *other*. No error correction was given. As in previous experiments, subjects in all except the labeled instances condition were free to interchange the keys for Wilson and Smith; their key assignments were determined afterward by the usual consistency test.

In addition to the eight experimental conditions described so far, an additional transfer control condition was included.[6] Subjects in this condition did not receive any learning trials. The transfer control condition was included to provide a base-rate estimate of the amount of learning that could occur without feedback solely during the transfer phase of the experiment.

One hundred undergraduates served as paid subjects. Ten were assigned to each of the eight experimental conditions, and 20 to the transfer control condition.

Results

Knowledge of category number. A preliminary assessment of the difficulty of the learning task in Experiment 3 is provided by the responses to the questions asked subjects in the number unknown and observation only conditions. The first two questions directed to the observation only subjects (Did you notice anything interesting about the patterns? Any similarities among them?) failed to elicit any clear statements regarding the presence of categories. When asked whether they had noticed that the patterns were divided into categories, 8 subjects in the observation only condition said yes, 11 said no, and 1 did not respond (without notable differences between those who had viewed low- vs. high-variability distributions). Finally, subjects in both the observation only and number unknown conditions were directly asked to estimate the number of categories. (Subjects in the observation only condition were first told that the patterns were indeed drawn from different categories.) The distributions of estimates did not differ across either instructional conditions or levels of training variability, so we will report the aggregate results for all 40 subjects: The median estimate was 4, with a range of 2 to 15. Previous studies have also reported overestimation of the number of categories (Bersted et al., 1969; Hartley & Homa, 1981). Only 4 subjects said there had been two categories; all of these 4 had received observation only instructions, and 3 had seen high-variability distributions—the learning situation one might well suppose had the least a priori likelihood of enabling the number of categories to be learned. Given the diversity of the estimates, it seems quite likely that the few subjects who gave the correct answer did so by fortuitous guessing. It is clear that virtually (and perhaps literally) none of the subjects learned the number of categories used to generate the patterns, even with exposure to 200 examples.

Transfer performance. Presented in Figure 7 is the percentage of patterns classified correctly as a function of distance from the standard, plotted separately for each instructional condition. The results for the conditions that received low-variability distributions during the learning phase appear in panel A, whereas the results for the conditions that received high-variability distributions appear in panel B. For purposes of comparison, the results for the transfer control condition are plotted in both panels. An analysis of variance was performed on these data, with the transfer control subjects divided into two arbitrary groups to create a balanced design. Percentage correct declined significantly with increasing distance from the standard for each of the eight experimental groups ($p < .01$) but not for the transfer control condition. Since subjects in the latter condition did not receive a training phase, any category learning would have had to take place over the course of the transfer trials. In fact, even this control condition produced a significant effect of distance from the standard when only data for the second half of the transfer trials were considered, $F(1, 76) = 11.4$, $p < .001$, indicating that learning did take place during the transfer phase. Percentage correct declined from .47 for the items closest to the standard to .36 for those farthest from it.

When the response *other* is available, percentage correct should show a greater decline with increasing distance for the low- than for the high-variability groups (as in Experiment 1A). As can be seen by comparing panels A and B in Figure 7, this prediction was confirmed, $F(4, 360) = 8.28$, $p < .001$. Collapsing over the four experimental conditions within each variability level, this interaction took the

[6] This control condition was suggested by Michael Flannagan.

form of a crossover: For those patterns closest to the standard, the percentage correct was higher for the low-variability condition (77% vs. 63%), whereas for those patterns furthest from the standard this difference reversed (20% vs. 32%). This pattern confirms the comparable result (Prediction T2) obtained in Experiment 1A.

The results of primary interest concern the effects of the different types of supplementary information on learning. Because subjects in the various instructional conditions within each variability level received exactly the same exemplars during learning, differences among the groups within each level of variability must reflect differential learning of the distributions. Superior learning should be evidenced by higher percentage correct for those patterns most likely to belong to the training distribution (i.e., those relatively close to the standard). (From the learner's point of view, patterns relatively far from the standard ought to be classified into the *other* category.) For the low-variability groups, an analysis was performed on the percentage correct data for patterns 5 cells from the prototype. The labeled instances and number known conditions did not differ ($t < 1$) but were superior to the number unknown and observation only conditions ($p < .01$). The latter two groups did not differ from each other ($t < 1$) but were superior to the transfer control condition ($p < .01$). A comparable analysis was performed for the high-variability groups, examining percentage correct for patterns 5 and 15 cells from the standard (the patterns most consistent with the high-variability training distributions). In this analysis the labeled instances group ex-

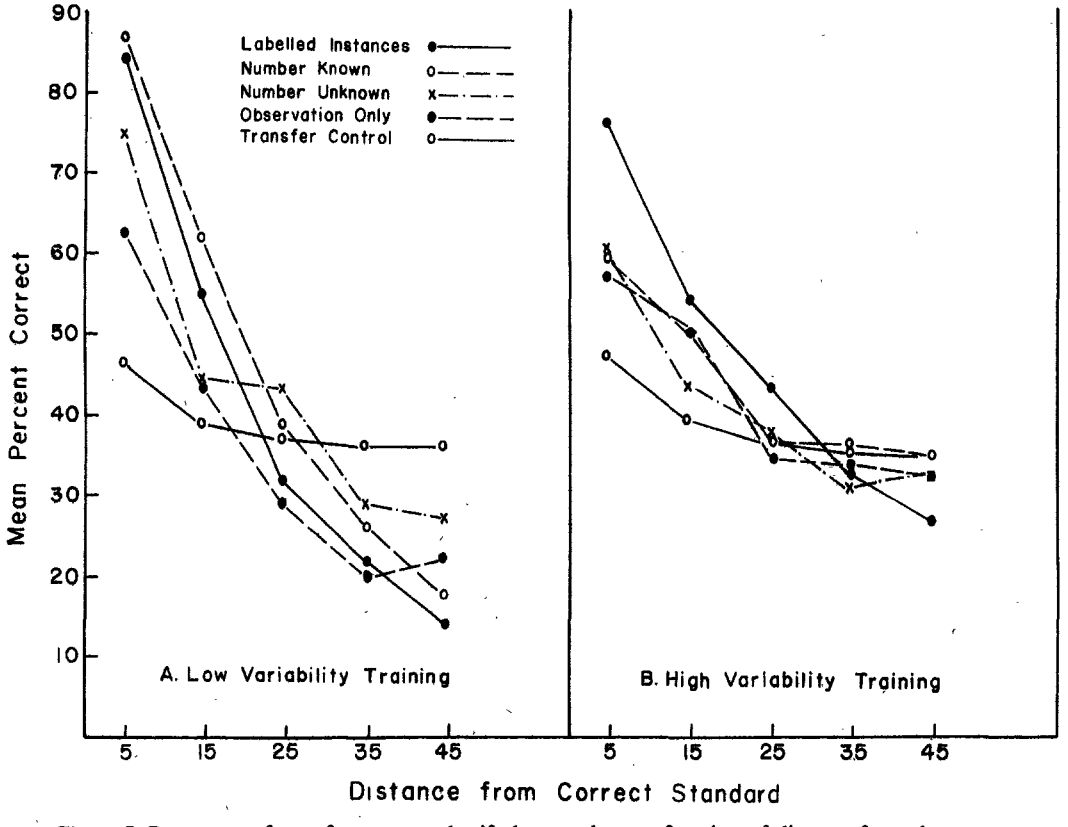

Figure 7. Percentage of transfer patterns classified correctly as a function of distance from the correct standard for the various instructional conditions (Experiment 3). (Data for low-variability conditions appear in panel A, and data for high-variability conditions appear in panel B; data for the transfer control condition are plotted in both panels.)

ceeded the other three experimental conditions ($p < .01$). The latter did not differ among themselves, but were collectively superior to the transfer control condition ($p < .05$).

A possibility to be considered is that the above differences in percentage correct only reflect differences among the propensities of different groups to use the *other* category, since *other* responses were always scored as incorrect. However, if only such response biases were operating, groups ranking relatively high in overall percentage correct would tend to rank relatively low in percentage correct of those patterns classified into a non*other* category. No such trade-off was apparent; rather, groups ranked relatively high in overall percentage correct tended to also be ranked relatively high in percentage correct of those patterns not classified *other*. For the low-variability groups conditional percentage correct was .79, .79, .73, .71, and .60 for the labeled instances, number known, number unknown, observation only, and control conditions, respectively. Comparable figures for the high-variability groups were .75, .63, .66, .61, and .60.

Discussion

The results just presented provide a mixture of support and difficulty for the category density model. First, consider the conditions that received low-variability training exemplars. Subjects in the labeled instances and number known conditions were most accurate in classifying the patterns closest to the standard. These are the only two learning conditions that had the prerequisite information for use of a strategy of parameter revision. The lack of difference between the labeled instances and number known conditions at the lower level of variability suggests that learning had approached asymptote in these two conditions after the 200 training trials. This result is thus consistent with the category density model. The advantage displayed by the number known condition relative to the number unknown and observation only conditions also supports the model (Prediction L4), because knowledge of category number is critical to the postulated parameter-revision process. However, the substantial, albeit lesser, degree of learning by subjects in the latter two conditions cannot be explained by a parameter-revision process. These subjects were not told the number of categories in advance, nor did they determine the number of categories during the learning trials; accordingly, they lacked the prerequisite information for parameter revision. The results thus implicate a second type of learning mechanism that requires even less supplementary information than does parameter revision.

The data for the high-variability condition also present a mixed picture for the category density model. The predicted advantage of receiving labeled instances (Prediction L2) was confirmed for categories with a high degree of overlap. However, as in the low-variability condition, substantial learning also took place in the number unknown and observation only conditions. Unlike the result obtained for the comparable low-variability groups, the number known condition was not superior to the two conditions that lacked knowledge of the number of categories (a failure of Prediction L4). The results thus suggest that subjects had available some other learning mechanism that operates as effectively for high-variability categories as does parameter revision without feedback.

Experiment 4

The relative efficacy of the various instructional conditions should be independent of whether or not an *other* alternative is provided at transfer. In the absence of an *other* category, superior learning should again be evidenced by higher percentage correct for patterns at the distortion levels most likely to have been observed during training. Without an *other* category, accuracy will necessarily approach an asymptote at chance level as distortion level is increased. Differences in percentage correct across instructional conditions should therefore be progressively attenuated as transfer patterns become more remote from the standards. As a result, the function relating percentage correct to distance from the correct standard should have a steeper slope for an instruction condition that produces superior learning (when variability of the training exemplars is held constant). Experiment 4 was performed to examine the effects of alternative levels of supplementary information on transfer performance in the absence of an *other*

alternative. Rather than repeating the full design of Experiment 3, only two comparisons of particular theoretical import were made. First, the number known and number unknown conditions with low-variability training were compared. We wished to replicate the advantage of the number known group (Prediction L4), observed for the comparable conditions in Experiment 3. Second, the labeled instances and number known conditions with high-variability training (for which subjects are not likely to reach asymptote after 200 learning trials) were also compared. Obtaining the predicted advantage of providing labeled instances (Prediction L2) would extend the comparable result observed in Experiment 3.

Method

The method of Experiment 4 was identical to that of Experiment 3, except that only the four groups mentioned above were included in the design, and no *other* alternative was provided during the transfer phase (as in Experiment 1B). Thirteen subjects served in each of the two low-variability conditions, and 8 served in each of the high-variability conditions.

Results and Discussion

Presented in Figure 8 is the mean percentage correct for the four conditions as a function of distance from the standard. All conditions yielded response functions with significant negative slopes ($p < .01$) approaching asymptote at the chance expectation of 50% for patterns maximally dissimilar to the standard. Both of the critical comparisons between conditions were in accord with the predictions of the density model. For the two low-variability groups (panel A in Figure 8), the slope of the distance function was significantly steeper for the number known than for the number unknown condition, $t(96) = 2.72$, $p < .01$ (Prediction L4). Tests of the simple main effects

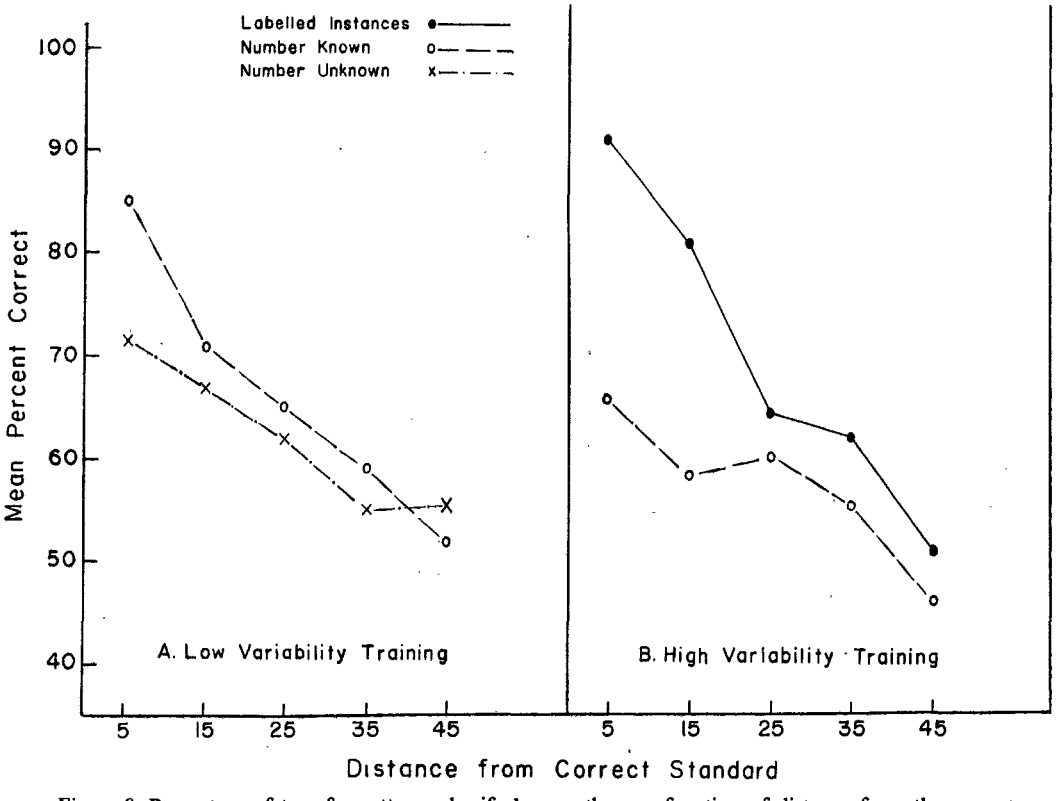

Figure 8. Percentage of transfer patterns classified correctly as a function of distance from the correct standard for the various instructional conditions (Experiment 4). (Data for low-variability conditions appear in panel A, and data for high-variability conditions appear in panel B.)

of instructional condition revealed that the number known condition produced significantly higher percentage correct for patterns five cells from the standard ($p < .01$) whereas the two groups did not differ significantly for patterns at higher distortion levels. As in Experiment 3, subjects in the number unknown group could not report the true number of categories (2) at the end of the learning phase. The median estimate was 4, with a range of 1 to 15. The superior performance exhibited by the number known condition for the patterns most likely to be generated by the training distribution can be attributed to the use of a parameter-revision strategy, which knowledge of category number makes possible. These results replicate and extend the parallel findings of Experiment 3.

Presented in panel B of Figure 8 are the results for the two high-variability conditions that were tested. The slope of the distance function was higher for the labeled instances than for the number known condition, $t(56) = 4.96$, $p < .001$, and percentage correct was significantly higher for the former condition at the two lowest distortion levels ($p < .01$, Prediction L2). These results extend the comparable findings obtained in Experiment 3.

General Discussion

Summary and Implications

The present results provide a broad initial basis of support for the category density model and its assumptions about distribution learning, as well as suggesting ways in which the model requires revision. The first two experiments yielded a number of results predicted by the proposed relative likelihood decision rule. These include the following: (a) the superior transfer performance that resulted from training on high-variability instances when an *other* category was available (Experiment 1A); (b) the absence of any clear advantage for the high-variability condition when the *other* category was removed (Experiment B); and (c) the tendency to classify items into the more likely high-variance category, even for patterns physically closer to the low-variance standard (Experiment 2). These results indicate that learners use exemplars to induce the distribution functions of categories, and then classify novel instances according to a relative likelihood rule based on these induced density functions.

Other results obtained in Experiments 1 and 2, plus the more detailed exploration of the learning process in Experiments 3 and 4, have implications for possible mechanisms of distribution learning. Major findings included the following: (a) Category distributions can generally be learned regardless of whether or not the learner receives labeled instances or error correction (Experiments 1A–3); (b) labeled instances, which obviate the need for a decomposition process, facilitate category learning under conditions of high distributional overlap (Experiments 3–4); (c) without labeled instances, prior knowledge of the number of categories (a prerequisite for paramenter revision) facilitates acquisition of category knowledge when the distributions do not overlap excessively (Experiments 3–4); and (d) category structure can still be learned when the learner does not receive error correction, information about category number, or even instructions to learn categories (Experiment 3).

Toward a General Model of Distribution Learning

This article has focused on a specific version of the category density model that can account for the induction of normally distributed categories when subjects know (or assume) the form of the distributions and the number of categories. It is clear that this specific model is too restrictive to account for all aspects of human capacity to learn category distributions. As we argued at the outset, normal distributions may well be an ecologically important special case; nonetheless, there is experimental evidence that people can sometimes learn markedly nonnormal distributions (Neumann, 1977). Furthermore, the results of Experiment 3 demonstrated that people can learn distributions to some degree not only without knowledge of the number of categories, but without knowledge that they are in a category-learning task at all. Perhaps a learning model entirely different from the category density model is required. However, we will instead suggest how the model can be revised and extended. The present findings, as well as other considerations discussed later, lead us to sketch a more general learning model which, while speculative, yields testable predictions.

The parameter-revision process described earlier allows a schematic representation of distributional knowledge to be updated without exemplars necessarily being stored in memory. The model is thus quite contrary in spirit to instance-storage models such as that proposed by Medin and Schaffer (1978). However, a more general model can be devised by integrating these two approaches. In this dual-representation, dual-process model, category distributions can be represented either by statistical parameters or by memory traces of instances. The latter type of representation permits a mechanism for learning category distributions without knowledge of the number of categories or the form of their density functions. This strategy, which corresponds closely to a proposal made by Evans (1967), involves storing traces of presented instances until separable clusters emerge. These memory traces need not be highly veridical, as long as they collectively approximate the mixture density of the sample. Each observed instance could be encoded as a point in a feature space, with an associated gradient of generalization. As further instances are encoded, a multidimensional frequency histogram will gradually be built up, providing a nonparametric representation of the mixture density.

An instance-clustering process might enable a person to learn categories without error correction, even if the learner does not initially know either the number of categories or the form of their distributions. If the underlying category distributions are sufficiently discriminable, separable clusters will eventually emerge, corresponding to peaks and valleys in the histogram for the mixture. Various clustering algorithms could be used to model the decomposition process (Duda & Hart, 1973; Everitt, 1974). As is the case for parameter revision, feedback would presumably facilitate decomposition.

The two learning mechanisms we have outlined—parameter revision and instance clustering—could be integrated by elaborating the learning procedure embodied in the computer simulation described in the beginning of this article. In the simulation, a clustering algorithm is used on the first few instances to derive initial parameter estimates, which are then updated using parameter revision. More generally, people may initially store and cluster instances in order to get a general idea of what the categories being presented are like. They may then summarize the information extracted from the stored instances as a parametric description, which subsequently can be fine-tuned using parameter revision. Initial instance clustering can be used to form initial conceptions of the category distribution, or to check a priori assumptions; parameter estimation and revision can be implemented at any point during the learning process once the learner is sufficiently confident about the number and form of the distributions.

Even after a parameter-revision strategy is invoked, continued incidental instance storage may play a role in detecting violations of basic assumptions about the form of the category distributions. For example, suppose the learner assumes the categories to be learned are normally distributed over a feature space, when in fact the distributions are markedly nonnormal (e.g., *V* shaped). A strict parameter-revision process would never be able to correct this misconception. Presumably, however, current parameter estimates would yield expectations about the frequencies of possible instances of the categories. If the learner found that supposedly rare instances were appearing too frequently, this would cast doubt on the assumed form of the density functions. A rational learner would then temporarily abandon parameter revision and shift to instance clustering.

In addition, there are very likely circumstances in which stored instances provide the only possible memory representation for a category distribution. Possible examples include the following situations:

1. The distributions of the underlying categories may be so irregular that no simple parametric description of them exists. For example, the category of *things stored in my attic* may have no simpler description than a list of all the items.

2. The learning set presented to subjects may be so small and/or variable that the categories appear to be collections of unrelated objects, as in Situation 1 just described. Numerous studies in the classification literature may exemplify this type of situation (e.g., Peterson, Meagher, Chait, & Gillie, 1973).

3. The learning task may emphasize rote memorization of instances and disguise the underlying category structure present in the learning set (Brooks, 1978).

These three situations are cases in which no simple schematic description of the category is available. This raises a question that is basic to the distributional framework: What is the range of category distributions that people can encode as parameter vectors? The present paper has emphasized normal distributions, which we suggested may be psychologically natural for continuous dimensions; however, other types of parametric representations may also prove important. For example, a binary feature dimension can be described by a Bernoulli process with its single parameter, p. Presumably there are limits on the types of distributions that people can veridically represent parametrically, and on those that may be learned by instance storage. Some research on the acquisition of nonnormal distributions has been done in studies of decision making (Pitz, Leung, Hamilos, & Terpening, 1976) and of category learning (Flannagan, Fried, & Holyoak, 1981; Neumann, 1977), but more work on this issue is clearly called for.

A second question that needs to be explored concerns the learning and representation of correlations between features within a category (e.g., members of small-bird species are more likely to sing than members of large-bird species). Medin and Schaffer (1978) have shown that people are sensitive to within-category feature correlations for artificial categories. Such correlational information could be represented parametrically by the equivalent of a variance–covariance matrix. Alternatively, a superordinate category composed of several distinct subcategories (i.e., separate or overlapping instance clusters) could be represented as the disjunction of the distributions of the subcategories; the features within each subcategory might be independent. Individual stored instances can be viewed as the limiting case of a category representation based on the disjunction of multiple distributions.[7]

Further Directions

We have already touched on a number of directions in which the category density model may guide research. There is some evidence that prior expectations can influence the induction of category distributions (Neumann, 1977), but the issue has just begun to be investigated systematically by manipulating the form of the distributions of category exemplars over known feature dimensions (Flannagan et al., 1981). In addition, the possibility that the learning process may undergo qualitative changes (e.g., a shift from instance clustering to parameter revision) calls for further studies that vary degree of category acquisition (Homa, Sterling, & Trepel, 1981). The types of categories studied also need to be extended. The category density model is not inherently restricted to categories defined solely by perceptual features as in the present study and most similar research. In principle people could learn category distributions over semantic or functional dimensions as well as perceptual ones.

Finally, it should be emphasized that the relationship between instance and category knowledge has broad import for cognitive theory. For example, Gick and Holyoak (1983) have investigated the induction of a "problem schema" from experience with multiple analogous problems. The category density model assumes that the ideal outcome of category learning is a representation of the dimensions of variation among category exemplars, together with a parametric description of category distributions over these dimensions. Since in its broader sense category learning is clearly

---

[7] The relative likelihood rule can be applied even if categories are represented solely by stored instances. Each stored instance will establish a *microdistribution*, equivalent to a generalization gradient around the point in a feature space defined by the instance. The relative likelihood rule then predicts that the subjective probability of a particular novel item given a particular category will be proportional to the sum of the subjective probabilities of the item given the microdistributions associated with each of the stored instances of that category (assuming the stored instances to be equally likely).

It should be noted that in this special case, in which distributions are represented solely by stored instances, the relative likelihood rule is isomorphic to the decision rule specified by Medin and Schaffer (1978, p. 211, Assumptions 2, 4, and 5). Their rule operates on a similarity parameter for each feature, which is defined to range in value between 0 and 1. The corresponding parameter of the relative likelihood rule, expressed in terms of features (Equation 4), is interpreted as the subjective likelihood of observing the feature value of the item to be classified given the feature value of the stored instance to which it is being compared. Because this likelihood ranges between 0 and 1 and is assumed to decrease monotonically with increasing distance between the two feature values, it has the same properties as the corresponding similarity parameter specified by Medin and Schaffer's rule. The two rules operate on these basic corresponding parameters in an entirely parallel fashion.

involved in various complex domains, such as problem solving and story understanding, a general theory of the learning process could serve to highlight commonalities among a wide range of cognitive activities.

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* (Vol. 13, pp. 277–318). New York: Academic Press.

Attneave, F. (1957). Transfer of experience with a class-schema to identification learning of patterns and shapes. *Journal of Experimental Psychology, 54,* 81–88.

Attneave, F., & Arnoult, M. D. (1956). The quantitative study of shape and pattern perception. *Psychological Bulletin, 53,* 452–471.

Bersted, C. T., Brown, B. R., & Evans, S. H. (1969). Free sorting with stimuli clustered in a multidimensional space. *Perception & Psychophysics, 6,* 409–414.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 170–207). New York: Wiley.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley.

Edmonds, E. M., & Evans, S. H. (1966). Schema learning without a prototype. *Psychonomic Science, 5,* 247–248.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Evans, S. H. (1967). A brief statement of schema theory. *Psychonomic Science, 8,* 87–88.

Evans, S. H., & Arnoult, M. D. (1967). Schematic concept formation: Demonstration in a free sorting task. *Psychonomic Science, 9,* 221–222.

Everitt, B. (1974). *Cluster analysis.* London: Heinemann Educational Books.

Flannagan, M., Fried, L. S., & Holyoak, K. J. (1981, November). *Perceptual category learning and distributional structure.* Paper presented at the 22nd meeting of the Psychonomic Society, Philadelphia, PA.

Fried, L. S. (1979). *Perceptual learning and classification with ill-defined categories* (Tech. Rep. No. MMPP 79-6). Ann Arbor: University of Michigan, Michigan Mathematical Psychology Program.

Gibson, E. J. (1953). Improvement in perceptual judgments as a function of controlled practice or training. *Psychological Review, 50,* 401–431.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Appleton.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62,* 32–41.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Hartley, J., & Homa, D. (1981). Abstraction of stylistic concepts. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 33–46.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16,* 321–338.

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.

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

Neumann, P. G. (1977). Visual prototype information with discontinuous representation of dimensions of variability. *Memory & Cognition, 5,* 187–197.

Oldfield, R. C. (1954). Memory mechanisms and the theory of schemata. *British Journal of Psychology, 45,* 14–23.

Patterson, J. F. (1979). *Hypothesis testing in a prototype abstraction task.* Unpublished doctoral dissertation, University of Michigan.

Peterson, M. J., Meagher, R. B., Chait, H., & Gillie, S. (1973). The abstraction and generalization of dot patterns. *Cognitive Psychology, 4,* 378–398.

Pitz, G. F., Leung, L. S., Hamilos, C., & Terpening, W. (1976). The use of probabilistic information in making predictions. *Organizational Behavior and Human Performance, 17,* 1–18.

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

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Cambridge, MA: Graduate School of Business Administration of Harvard University.

Reber, A. S., & Allen, R. (1978). Analogic and abstraction strategies in synthetic grammar learning: A functionalist interpretation. *Cognition, 6,* 189–221.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 383–407.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 28–46). New York: Wiley.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies on the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340.

Tracy, J. F., & Evans, S. H. (1967). Supplementary information in schematic concept formation. *Psychonomic Science, 9,* 313–314.

Wallsten, T. S. (1976). Using conjoint-measurement models to investigate a theory about probabilistic information processing. *Journal of Mathematical Psychology, 14,* 144–185.

Received April 30, 1982
Revision received May 16, 1983 ■